

COLLOQUIAL LEXEMES IN JOURNALISTIC TEXTS

LUCIA JASINSKÁ

Department of Slovak Studies, Slavonic Philologies, and Communication
Faculty of Arts, P. J. Šafárik University in Košice, Slovakia

JASINSKÁ, Lucia: Colloquial lexemes in journalistic texts. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 139 – 147.

Abstract: In our paper we mainly focus on the research of colloquial lexical units in journalistic texts. The aim of the research is colloquiality as a marked attribute of journalistic texts. At first we define the terms *hovorovost'* (*colloquiality*) (also in relation to the term *hovorenost'* (*spokenness*)) and *hovorový* (*colloquial*). Since the point was the research of “living language” – represented by field of journalism – our source material were journalistic texts from the database of the Slovak National Corpus. The number of occurrence of colloquial lexical units was recorded according to their absolute frequency and the results were categorized and interpreted. The most frequented means of expression were verified in current lexicographic processing and the changes of the indicator of colloquiality was studied. With style parameters in background, we evaluated the markedness of the vocabulary of analyzed journalistic texts.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 INTRODUCTION

Journalistic texts have currently an important place in the sphere of social communication. Their primary mission is to inform the addressees of various social layers, age, and education as quickly and accurately as possible. The authors of journalistic texts should represent social events in a semantically unambiguous and understandable way as to their expression aspect. The field of journalism may be considered the most flexible, most dynamic, and most formative phenomenon of the present, which is being developed and modified even in connection with the use of intralingual, especially lexical, means.

2 THEORETICAL BACKGROUND

In the present paper, we will focus on **colloquial lexemes** in journalistic texts. In the introduction to this paper, we consider it necessary – even with regard to the definition of variance in linguistics – to define the term *hovorovost'* (*colloquiality*) (also in relation to the term *hovorenost'* (*spokenness*)), as these two partially overlap).

The language notion of **hovorový (colloquial)** refers to words that are “characteristic of ordinary unofficial verbal manifestations” [13, p. 161] or by using its attribute form in the combination of *colloquial style* of the standard language we mean the so-called “ordinary communication” (ibid.).

Hovorený (spoken) is said to be the one “which is expressed in its audial form, oral: utterance; spoken form of standard Slovak [13, p. 160]. This is then such a materialization of language that may be perceived as verbal or oral. We consider verbal manifestations made in writing its counterpart.

Colloquiality as now a linguistic term is “the property of the means of language (audial, morphological, lexical, syntactic ones) resulting from their attachment to spoken, unofficial, spontaneous, and usually oral utterances” [13, p. 160]. Several linguists have addressed the meaning of the above key words.¹

In connection with the concept of colloquiality, J. Bosák contemplates a kind of reflection of the social situation in language consisting in “increasing the utterances in standard communication” and also in the “ever wider and deeper effect of the means of mass communication” [2, p. 65]. As stated by the author, the term colloquiality may be understood as a functional language style, the colloquial lexicon, or the colloquial layer within the vocabulary, but also the means of colloquiality. From the point of view of language stratification, colloquial units, including colloquial language, are included herein.

In her reflections, M. Ivanová-Šalingová [7] identifies the notions of *hovorenosť (spokenness)* and *hovorovosť (colloquiality)* when she characterizes colloquial Slovak as a spoken form of standard language. Contrary to that, Š. Peciar considers this statement to be unsatisfactory, as it concerns the identification of the colloquial style of standard Slovak with the spoken form of the standard language. As stated by the author, colloquial style is one of the basic functional language styles that may be characterized as a “purposeful selection and arrangement of those standard means (lexical, word-forming, grammatical, and phonetic ones) that are employed by active users of standard language in utterances in everyday language communication, in the milieu of friends and family” [10, pp. 47–48].

Systematic approach to this issue is applied by J. Findra who classifies the **colloquial style** developing on the basis of the national language as a model structure of colloquial texts and identifies it by such properties as being of private, unofficial, verbal, and dialogical nature which imply the addressee’s presence and communicative function. Of course, the domain of colloquial acts of communication is to be found in the private milieu and in the oral form of their language materialization [5]. In this paper, we identify ourselves with the narrower definition

¹ Some linguists’ views on defining the notions of spokenness, colloquiality, and colloquial style are only categorized with a view to the limited scope of the paper.

of the colloquial style (as suggested by Bosák [2]) as one of the forms by which manifestations are materialized in the common communication sphere.

Confrontation of the notions of common spoken language and colloquial style also appears to be a constructive element in defining colloquiality and colloquial means in communication (including the media one). Colloquiality with a differentiated degree of representation in individual acts of communication may then be perceived as a manifestation of the dynamism of language and its individual spheres. The framework of colloquiality is also enhanced by colloquial lexical means of expression.

Colloquial lexemes make up the basic component in the vocabulary of colloquial style. This is an inventory of language units used primarily in “private-communication oral language manifestations” [5, p. 55]. J. Findra defines here properties such as spontaneity, privacy, dialogical and unofficial nature, and situational anchoring of utterance. While these are typical means of expression in oral private acts of communication, they are not used in written texts, especially not in those of educational or administrative type. More specifically, they should not be used there or their use would have to be well founded and functional.

In this paper, we analyse lexical means in the texts of mostly journalistic style, which finds its space in the media product environment, whereby these are “not just newspapers, but nowadays also radio, television, and journalistic film” [9, p. 459] and we add that in the 21st century even the Internet environment should be included.

According to J. Mistrik, the **journalistic style** – oscillating on the borderline between objective and subjective styles – is “a way of purposeful selection and thematic arrangement of those standard language means that are used for making the most up-to-date, accurate, and convincing information [intended] to the public” [9, p. 460]. Essential features of this style include written form, monological nature, publicity, and conceptuality, its specific features including information, variability, consistency, and topicality. J. Findra defines the following dominant features of the journalistic functional style as a model of the surface organization of the media text: “the public, official nature, written, monological form, absence of the addressee, communicative function” [5, p. 262]. Although in connection with the development of the electronic media the presence of the addressee may be justifiably considered.

By comparing the characteristics of the journalistic and colloquial styles, we state that they form a kind of opposition to the surface organization of the structure of the texts, since it is only the communicative function that is their shared attribute. While this is a distinctive feature in the colloquial-style texts, in the journalistic acts of communication the communicative function should be linked to the cognitive one, particularly in terms of accuracy, information load, and objectivity.

In the journalistic style, lexical means prevail, by which information are provided in an accurate and comprehensible way. Thus, the authors of the texts should avoid “special expressions, professional terms with a narrow or limited

extent of the meaning of the word (texts of the journalistic style must be comprehensible to a wide range of recipients), but also words typical of colloquial style, i.e. ordinary, unofficial, spontaneous speeches (these, however, penetrate into emotionally tinted journalism, partly even into the rational-type journalism, including non-literary words)” [6, p. 20]. The relevant question is therefore contained in the rate of penetration of colloquiality into official media acts of communicating a message.

Focusing on the expressive means of colloquiality, we perceive **colloquiality** as “the property of the linguistic means occurring not just in the colloquial style, but also in other styles, specifically in the manifestations of other styles that are materialized in their oral form” [1, p. 182]. We add that the means having the attribute of colloquiality may occur not just in spoken acts of communication, but also in written acts of communication as forms of any functional language style, although the aim here is not their prototype-like application. This fact is the stimulus for our research, therefore we will focus in the analysis on the written acts of communication of journalistic style in which the presence of colloquial means of expression is marked.

3 RESEARCH METHODOLOGY

With regard to the extent and striking dispersion of the means of expression in the semantic subsystem of language (lexical, syntactic, and morphological levels), we focus our research intention at the lexical level. The sub-corpus of journalistic texts² in the current version of the Slovak National Corpus was the source text material for the analysis.

At first in the *Krátky slovník slovenského jazyka* (Short Dictionary of the Slovak Language) [8] we have been manually searching for all of the auto-semantic lexical units with the qualifier “hovor.” (coll.), i.e. the words that are classified as colloquial. In particular, we focused on all the auto-semantic lexemes, even though we have also been searching for syn-semantics and examining the qualifier of colloquiality in them at various levels:

a) primarily at the level of lexemes (*áčko*, *hit*, *krach*, *macher*, *presilovka*, *rámus*, *surfovať*, *zubačka* ‘A-class, hit, smashup, coxcomp, power play, hurly-burly, surf, cog-rail’);

b) secondarily at the level of lexia or lexias (*bugina* = 1. a folding stroller; *citovať* ‘quote’ – 2. summon; *diplomátka* – 2. a lighter briefcase; *doktor* – 4. physician; *glazúra* – 2. glaze; *mangľovať* – 2. beat, batter; *nakladačka* – 2. cucumbers

² Slovak National Corpus – prim-8.0-public-inf. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2018. Available at: <http://korpus.juls.savba.sk>. There are 1 009 613 215 tokens (791 376 893 words) located in the sub-corpus of informative acts of communication.

for pickling; *odprášiť* – 2. run away, 3. expel, dust; *plačka* – 1. a weepy woman; *vizita* – 2. visit); or

c) at the level of contextual meaning in relation to the corresponding lexia (*číslo* ‘number’ – to 3. “part of the performance; one performance” fig. coll. *to je číslo!*; *dirigent* ‘director’ – to the meaning of “who directs”, coll. *dirigent mužstva* ‘director of the team’; *lietací* – to the meaning of “designed, adapted to fly” coll. *lietacie dvere* = swinging door; *vizita (visit)* – meaning “regular visit by a group of physicians in patient rooms”, coll. “a group of such physicians”).

We recorded separately homonymous colloquial lexemes as long as they were marked with the appropriate qualifier. There were mono-semantic words found in this group, such as:

*fuk*¹ – „adv. **coll. expr.** in the word group *to, on (mi) je f.* jedno, ľahostajné (-ý) (I don’t care)“;

*fuk*² – „usually plural. **coll. expr.** peniaz, peniaze (money): *(ne)mať (have-not to have) f-y*“ (KSSJ, 2003, s.169);

but also polysemantic lexical units, such as:

*baba*¹ -y *báb f.*

1. **coll.** stará žena ‘an old woman’: *stará b.*,

2. **pejor.** neprijemná, zlá, protivná žena ‘unpleasant, evil, annoying woman’: *klebetná, zlostná b.*,

3. **slang.** mladá žena, dievča ‘a young woman, girl’: *b-y z internátu,*

4. **coll.** pôrodná asistentka ‘childbirth assistant’: *pôrodná b.*,

5. **pejor.** zbabelec, bojazlivec, slaboch ‘coward, warlord, weak’: *nebud’ b.!*,

6. *slepá b.* children game: *hrať sa na slepú b-u and phras. sham,*

*baba*² -y *báb f. coll.*

1. empty pie,

2. haruľa (potato cake): *zemiaková b.*,

3. white grub,

4. colon sausage (KSSJ, 2003, p. 56).

Identifying the exact occurrence of polysemantic and homonymous lexemes in the corpus was problematic in part. Therefore, we have not included in the database of the lexemes searched – with regard to further word processing – the verbs with a number of lexias higher than 4 (e.g. *rezať, prísť, sedieť* ‘cut, come, sit’ – the qualifier “hovor.” (coll.) is only given in some lexias); sporadically, such a phenomenon also occurred with adjectives (*silný, ťažký* ‘strong, heavy’) or nouns (*chlapec, život* ‘boy, life’).

In this way we searched and subsequently recorded 1,770 lexical units in order to verify their occurrence in the texts of the journalistic style. Since this is an extensive material that require time-consuming processing, we are currently checking the occurrence of the first 550 lexemes (1. *abonentka* ‘female subscriber’ → 550. *krstňa* ‘god-child’) in the SNK.

In the relevant sub-corpus, we were searching for its absolute frequency of each lexical unit. However, for several multi-semantic words, it was not possible to detect the exact incidence of colloquial lexias due to polysemia or the occurrence of lexemes in case of homonymy (*akord*² – úkolová práca ‘chord’ – task work; *baba*; *baba*¹, *baba*² – non-colloquial + colloquial semes, altogether 11,249 occurrences; *babka*² – 1. part of a hasp and staple, 2. ankle forging of scythe; *cifra*² ‘cipher’ – ‘ornament’; *citovať*² – ‘summon’; *cvik*² – ‘vee’; *dóza*¹ (*box*); *huba*² – ‘mouth’, ‘yap’; *kopačka*¹; *koštovať*¹, *koštovať*²).

A high level of homomorphism was also problematic in research, for example: *áčka* (*As*); *advokát* ‘advocate’ (= advocate in general to the basic meaning of “legal representative, attorney”); *béčka* ‘Bs’; *bežat’* ‘run’ (3. ísť, ponáhľať sa ‘go, rush’; *bláznit’ sa* ‘be crazy’ (2. show striking interest in someone, something); *blázon* ‘fool, crazy, mad’ (3. unreasonable person; crank); *bledý* ‘pale’ (3. evil, critical); *bronz* ‘bronze’ (2. bronze-coloured); *céčka* ‘Cs’; *ceremónia* ‘ceremony’ (2. okolky ‘shilly-shally’; corgoň (= a grown up, big boy, young man); *cukrovka* (1. sugar beet); *čara* (= exchange); *čarať*, *čarovať*¹ (= meniť, vymieňať, zamieňať) (change, replace, swap); *číslo* (fig. coll. *to je číslo!* ‘what a feat!’ to the 3rd meaning “part of the show, performance”); *darovať* ‘give, donate’ (2. usually in the negative of forgive – *to mu nedarujem!* ‘I can’t wink at that’); *déčka* ‘Ds’; *detská* ‘playroom’; *dial’kar* (2. who is studying part-time, 3. driver on the long distance tracks); *dodávka* (3. van); *doktor* (4. physician); *éčka* (*Es*); *expresný* (= coll. this was an express job to the basic meaning of “quick, fast, rapid”); *fantastický* (2. huge, unheard of); *filmovat’* (3. play a role in the film, 4. pretend); *chlapec* (4. = male); *chlieb* (2. livelihood); *kontra* (2. opposition in card games). These lexemes, or their colloquial semes, were not identifiable by frequency.

For many lexemes, it was necessary to work with the filtering tool or manually filter out the irregular forms that were part of the searched expression paradigm. For example, in the corpus paradigm of the lemma *káro*, there are also grammatical forms of other lexemes, namely the forms of the non-verbal lexemes of *kára* (*jazdil na káre* ‘man rode the car’), or, alternatively, as well as the form-identical proprium (*Sergej Kára*) (*Sergei Kára*), or the searched expression occurs as part of another word (... *niekoľko SBS-károv na riadenie plynulosti parkovania* ‘... a few security men to control the smooth flow of parking’).

4 RESEARCH RESULTS

After searching for excerpted colloquial lexemes, we have created 8 groups of words with differentiated incidence to clarify the results of their absolute frequency in the sub-corpus of journalistic texts. The data is presented in the following table:

Group	ABSOLUTE FREQUENCY OF OCCURRENCE	NUMBER OF LU
1	over 10 000	18
2	9 999 – 5 000	18
3	4 999 – 1 000	61
4	999 – 500	34
5	499 – 100	132
6	99 and less	205
7	zero incidence	39
8	unidentified incidence	43

Tab. 1. Groups of words in accordance to their absolute frequency

As can be seen from the above overview, the most sought-after colloquial means (205) are in the range from 1 to 99 (*fotokrúžok, abonentka, filc, kancel', hotel, dvojkár, čertovina, háčko, buksa, elpéčko, kamrlík, gyps, fáč*) 'photo circle, subscriber, felt, office, armchair, a student usually being awarded Bs, devilish trick, an H, box, LP, cubbyhole, gyps, bandage', in general, it is thus a relatively low incidence of colloquial words that do not disturb the course of communication, on the contrary, they functionally express the intention of the author. We add that the use of colloquial expression is determined by the theme of the act of communication and also reflected by the zero incidence of some lexemes (*babra, figľovať, golfky, došpatiť, javeru*). The absence of these words in the texts of the journalistic style reflects their position in the language vocabulary. These are in fact lexical units that are located on the periphery within the lexical stratification, and thus are little used in ordinary spoken communication. In addition, one may assume their higher level of colloquiality. Compared to them, there are words with a lower level of colloquiality that are the most frequent in the journalistic texts. Such colloquial expressions include the lexemes: *automobilka* 'car-making factory' (46 207), *inkasovať* 'cash' (28 743), *chalan* 'boy' (27 807), *hit* 'hit' (25 643), *akurát* 'just' (24 927), *foto* 'photo' (22 639), *kanonier* 'canonier' (20 361), *fabrika* 'factory' (18 659), *bytovka* 'apartment building' (17 737), *krčma* 'pub' (15 698), *fotoaparát* 'camera' (14 409), *kontaktovať* 'contact' (14 363), *garantovať* 'to guarantee' (13 909), *garancia* 'guarantee' (13 029), *krach* 'crash' (12 965), *fajn* 'fine' (12 912), *kilo* 'kilo' (11 559), *kemp* 'camp' (10 365).

The words with the highest incidence were examined in the SSSJ (2006, 2011) to determine whether they were reclassified to neutral words or lexemes/lexias with another qualifier³. A significant change was found in 7 words, namely *automobilka* 'car factory', *hit*, *bytovka* (= residential house; in the case of a new, differentiated

³ Since the processing of all the volumes of the *Slovník súčasného slovenského jazyka* (Dictionary of Contemporary Slovak Language) (2006, 2011, 2015) is based on corpus findings, we assumed that one, though not the only one, determinant of lexicographic entries – with an impact on the classification symptoms of lexeme and their lexicon – is frequency the occurrence of the word in the SNK.

meaning “bytová krádež” (“housing theft”) the colloquiality qualifier is given), *krčma*, *fotopaparát*, *kontaktovať*, *garancia*, ‘pub, camera, to contact, guarantee’ that are in the current dictionary processed as non-colloquial, that is, stylistically unmarked. Conversely, as colloquial are classified the words as *inkasovať* (= get a goal, blast ...), *chalan*, *akurát*, *fotka*, *fabrika* a *kilo* ‘boy, just, photo, factory and kilo’. In some words, the qualifier changes are partial. Either the word is only colloquial in one of its meanings (e.g. *garantovať* ‘guarantee’ – 3. coll. give a verbal guarantee; *krach* ‘crach’ – 2. loss of favourable state), or the original colloquial mark has changed to another one, signalling the use of the word or of its lexia as a professional expression (e.g. *kanonier* – 2. sport. a player famous for shooting at the goal; *kemp* – 2. sport. professional training center).

Our research suggests that the words that are often used in journalistic acts of communication get into the core of the vocabulary, and at the same time their stylistic markedness is attenuated (see also Bosák [2], [5]). A similar tendency in the language was indicated by I. Bónová, who confirmed by research that various forms of words from the so-called common language, when used for a longer period of time, will infiltrate the standard variety and will gradually acquire a neutral stylistic validity within the dynamizing processes [3].

Thus, we may state that the frequency of incidence of the marked means of expression is a determining factor in relation to its stylistic markedness. As the frequency is decreasing, the presence of the mark of colloquiality is becoming higher. Moreover, it is not a state but a process in which markedness, e.g. colloquiality, attenuates gradually.

5 CONCLUSION

The subject matter of the submitted paper has been the research into lexical means with the attribute of colloquiality in journalistic texts. It turns out that the incidence of the lexemes (and lexias) examined is significant in some cases, suggesting the penetration of these elements into the field of journalistic style. However, we do not assess this fact as counterproductive, i.e. the incidence of colloquial means of expression – primarily used in interpersonal unofficial, familial communication – does not decrease the official nature of the journalistic communication, on the contrary, the authors are also “approaching” the percipient or audience in this way.

On the basis of our partial research, we may confirm the tendency to dynamize the language system, since a significant rate of use of some of the originally colloquial words leads to their reclassification into non-colloquial lexemes or lexias or to re-classifying them into terms from different areas (sport, culture, etc.). The changes that have occurred in the process of compiling a new glossary (in comparison to the KSSJ) confirm at the same time the research methodology set and mean an

indirect feedback in relation to the relevance of examining the incidence of colloquial means of expression in the texts representing a “live language”. This constantly confirms the “live speech” pressure on the status quo in the language system.

ACKNOWLEDGMENTS

This paper has been elaborated as part of the grant project No. 008UPJŠ-4/2017 Science Without Barriers (Interdisciplinary Inspirations of Contemporary Literary Scholarship and Linguistics in Educational Practice at University, Project Leader: Prof. PhDr. Ján Gbúr, CSc.).

References

- [1] Bodnárová, M. (2013). Terminologická reflexia komunikácie v súkromnej sfére vo vzťahu k hovorovému štýlu a jemu blízkym javom – niekoľko úvah a poznámok. *Slovenská reč*, 78(3–4), pages 174–186.
- [2] Bosák, J. (1984). Hovorovosť ako dynamický faktor. *Slovenská reč*, 49(2), pages 65–73.
- [3] Bónová, I. (2017). K problematike hovorenej podoby jazyka na východnom Slovensku. *Slovenčinár*, 4(1), pages 4–8.
- [4] Bosák, J. (1995). Sociolingvistická stratégia výskumu slovenčiny. In *Sociolingvistické aspekty výskumu súčasnej slovenčiny*. Sociolinguistica Slovaca. Ed. S. Ondrejovič – M. Šimková. Bratislava, Veda, pages 17–42.
- [5] Findra, J. (2013). *Štylistika súčasnej slovenčiny*. Martin, Osveta, 320 p.
- [6] Gladiš, M. (2015). *Žáner v prostredí masových médií*. Košice, UPJŠ, 128 p.
- [7] Ivanová-Šalingová, M. (1963). Hovorový štýl súčasnej spisovnej slovenčiny. *Slovenská reč*, (28), pages 17–32.
- [8] *Krátky slovník slovenského jazyka*. (2003). Eds. J. Kačala – M. Pisárčiková – M. Považaj. 4th amended and modified issue. Bratislava, Veda, 985 p.
- [9] Mistrík, J. (1989). *Štylistika*. 2nd edition. Bratislava, SPN, 584 p.
- [10] Peciar, Š. (1965). Jazykové prostriedky hovorového štýlu spisovnej slovenčiny. In *Jazykovedné štúdie*. 8. Bratislava, Vydavateľstvo SAV, pages 42–70.
- [11] Reifová et al. (2004). *Slovník mediální komunikace*. Praha, Portál, 328. p.
- [12] *Slovník súčasného slovenského jazyka. A – G* (2006). Schol. ed. K. Buzássyová – A. Jarošová. Bratislava, Veda, 1,134 p.
- [13] *Slovník súčasného slovenského jazyka. H – L* (2011). Schol. ed. A. Jarošová – K. Buzássyová. Bratislava, Veda, 1,087 p.
- [14] *Slovník súčasného slovenského jazyka. M – N* (2015). Schol. ed. A. Jarošová. Bratislava, Veda, 1,104 p.

FREQUENCY IN CORPORA AS A SIGNAL OF LEXICALIZATION (ON THE ABSOLUTE USAGE OF COMPARATIVE AND SUPERLATIVE ADJECTIVES)

PAVLA KOCHOVÁ

Czech Language Institute of Czech Academy of Sciences, Prague, Czech Republic

KOCHOVÁ, Pavla: Frequency in corpora as a signal of lexicalization (on the absolute usage of comparative and superlative adjectives). *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 148 – 157.

Abstract: The study deals with the category of comparison of Czech adjectives from the semantic point of view; it concentrates especially on the so-called absolute (or relative) usage of comparatives and the absolute usage of superlatives and their lexicographic treatment (or absence of the lexicographic treatment) in Czech monolingual dictionaries. The question is whether their frequency in corpora can prove lexicalization of this usage.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 STARTING POINTS

1.1 The status of the adjective gradation in the language description

It is a well-known fact that in the language description the gradation of adjectives stands between inflection (morphology) and derivation (word-formation). This has been clearly expressed in [11]. J. Panevová [18], [19] presents a brief overview of arguments both for morphology and for word-formation and notes that descriptions in various grammar handbooks of the Czech language differ: the *Grammar of Czech (Mluvnice češtiny)* from 1986 treats gradation differently in different places – in the first volume as part of word-formation [5, pp. 378–380, 448–49] and in the second volume as part of morphology [14, pp. 79–80]; the *Handbook of Czech Grammar (Příruční mluvnice češtiny)* [12, pp. 176–180, 222] and the handbook *Czech – Speech and Language (Čeština – řeč a jazyk)* [4, pp. 132–133, 141] consider the comparative and the superlative to be part of word-formation, in contrast to the *Grammar of Contemporary Czech (Mluvnice současné češtiny)* [3, pp. 205–209] and the *Academic Grammar of Contemporary Czech (Akademická gramatika současné češtiny)* [21, pp. 380–385, 503–507], which consider them to be part of morphology. Concerning the lexicographic treatment, in Czech monolingual academic dictionaries, i.e. the *Reference Dictionary of the Czech Language (Příruční slovník jazyka českého)* [8], the *Dictionary of the Standard Czech Language (Slovník spisovného jazyka českého)* [7], hereinafter SsJČ), the *Dictionary of Standard Czech*

for Schools and the General Public (*Slovník spisovné češtiny pro školu a veřejnost* [6], hereinafter SSČ), the Academic Dictionary of Contemporary Czech (*Akademický slovník současné češtiny*, hereinafter ASSČ), as well as commercial dictionaries, i.e. the Internet Dictionary of Contemporary Czech (*Internetový slovník současné češtiny* [10], hereinafter ISSČ), comparatives are listed as part of grammatical information, e.g. the comparative *větší* ‘bigger’ is listed in the entry for *vel(i)ký* ‘big’ – mainly because it is an irregular comparative form.¹ In a similar way the comparative *lepší* ‘better’ is listed in the entry *dobrý* ‘good’, at the same time it is treated as a separate entry – due to the lexicalization of the absolute (elative) usage of this comparative form.² The “grammatical” treatment of comparatives is probably related to the space limitations of the dictionary. The formation of comparative (and superlative) forms and their semantic structure are very regular; therefore, there are usually no lexicographic reasons for their special description.

It is evident that the absolute usage of comparative (or superlative) forms should be taken into consideration by lexicographers. As [1], [18], [19], [13]³, and grammar handbooks (see [14, pp. 79–80]; [21, pp. 383–384]) indicate, there exists an asymmetry between the form (comparative) and its function (non-comparative, absolute usage),⁴ i.e. comparative and superlative forms are used without explicit comparison. Some of these occurrences could be considered as their “regular” comparative usage – although comparison is not explicitly expressed – because the comparative aspect is more or less obvious (implicit comparison with the average or usual value is present) (see [18], [19]).

1.2 Treatment of lexicalized comparatives and superlatives in monolingual dictionaries

This fact influences the number of registered comparative (or superlative) forms with a lexicalized meaning, i.e. those that are listed in monolingual dictionaries. Surprisingly, the number of registered comparatives which are treated as a lexical unit (i.e. not only as adjective morphological forms) is very low. In the SSČ there are lexemes *lepší* ‘better’, *vyšší* ‘higher’, *menší* ‘smaller’, *starší* ‘older’, *horší* ‘worse’, *nížší* ‘lower’, *mladší* ‘younger’, *blížejší* ‘closer’, *dřívější*⁵ ‘earlier’, *pozdější* ‘later’,

¹ Superlative forms are not usually listed because of their regularity.

² On the treatment of gradation of adjectives (and adverbs) in the ASSČ (and in other existing Czech monolingual dictionaries), see [20].

³ See the entries *Stupňování* (P. Karlík), *Elativ* (Z. Hladká) and *Komparativ and Superlativ* (both entries by K. Osolsobě).

⁴ The absolute usage is usually analysed for comparative forms ([18], [19]; [21]); [1], [14], [13] (in the entry *Superlativ*) mention it for superlative forms as well.

⁵ The comparative status of the form *dřívější* (allegedly belonging to the positive form *brzký* ‘early’) was questioned in [13, *Komparativ*]. Nevertheless in the Internet language handbook (*Internetová jazyková příručka* [9]), in [3], and in the ASSČ it is registered as a comparative for the lemma *brzký*.

*širší*⁶ ‘wider’, there are the same lexemes (except *horší* and *starší*) in the ISSČ. In a larger, but older dictionary (SSJČ) we find only lexemes *lepší*, *menší*, *starší*, *mladší*, *pozdější*, *dřívější*, *blíže*⁷. There are no lexicalized superlatives in any Czech monolingual dictionary,⁸ but in the Dictionary of the Contemporary Slovak Language (*Slovník súčasného slovenského jazyka*, hereinafter SSSJ) we can find the superlative adjective forms *najbližší* ‘the closest, the nearest’, *najmenší* ‘the smallest’, which are listed as separate lexical units; the treatment of some adjectives in the positive form indicates the lexicalization of some other superlative forms, e.g. *drahý* ‘dear’. Entries for superlative forms *najhorší* ‘the worst’, *najlepší* ‘the best’, *najsvätejší* ‘the most holy’, and *najvyšší* ‘the highest’ are listed because they form phraseological and multi-word units.

2 ADJECTIVE GRADATION IN THE LATEST LEXICOGRAPHIC DESCRIPTION OF CZECH

2.1 Adjective gradation from the frequency point of view

It is a well-known fact that gradation applies to a small part of adjectives for semantic and formal reasons. In [3, p. 205] it is presented that only 6% of adjectives have graded forms, in [17, p. 21] the proportion is indicated to be 8% (i.e. 5440 lemmas of 65 400 adjective lexemes in the most complete corpus then, in 2010). In the corpus SYN2015 adjectives with comparative or superlative forms make up 5% of all adjectives.⁹ The above-mentioned distinctions lie in different corpus sources that were used. As it was pointed out, the possibility of the adjective gradation is related to semantics.

Commonly, qualitative adjectives create comparative and superlative forms, relational (classifying) adjectives do not. From a different point of view, the occurrence of these forms within the category of (originally) classifying adjectives indicates semantic changes. Therefore, in the ASSČ (started in 2012) comparative forms are registered in all adjective entries where comparatives are commonly used (the same principle is used in the SSSJ); this information has two functions – to indicate the correct form and to signal the semantic type of the adjective.

At the same time, gradable adjectives occur in comparative and superlative forms quite rarely. Experts from the Institute of the Czech National Corpus were

⁶ It is not listed as a separate entry but it is registered within the entry for *široký* ‘wide’; it is placed in the sense ‘relating to as many people as possible from a certain field’ with the comment “often *širší* (without comparison)” and with the examples *pořady pro široký, širší okruh posluchačů; výstava pro širokou veřejnost; zboží široké spotřeby* ‘programs for a wide, wider audience; an exhibition for the general public; goods of wide consumption’.

⁷ It is registered within the entry *blízký* in the sense “detailed, elaborate” with the comment “only the comparative”.

⁸ Only in the ISSČ lexicalized nouns from superlative adjective forms are registered – *nejmenší* ‘very young child’, *nejhorší* ‘the worst alternative’.

⁹ The methodology of determining this number was similar to that of [17, pp. 20–21].

asked to make a list of adjectives that are commonly used in comparative and superlative forms.¹⁰ The list contains units that have an absolute frequency of at least 10 in the corpus SYN2015, and at the same time appear in their graded forms, i.e. comparatives and superlatives together, at least in 20% of their occurrences. The list is quite short, including the adjectives with suppletive comparatives and superlatives (i.e. *dobrý* ‘good’, *lepší, nejlepší*; *velký* ‘big’, *větší, největší*; *malý* ‘small’, *menší, nejmenší*; *špatný* ‘bad’, *horší, nejhorší*; *dlouhý* ‘long’, *delší, nejdelší*; *brzký* ‘early’, *dřívější, nejdřívější*).

	Lemma	Comparative + superlative forms		Lemma	Comparative + superlative forms		Lemma	Comparative + superlative forms
1	velký ‘big, great’	72 053	31	přísný ‘strict’	1 023	61	vytížený ‘busy’	104
2	dobrý ‘good’	55 740	32	podrobný ‘detailed’	1 019	62	prozaický ‘prosaic’	93
3	vysoký ‘high’	38 906	33	pomalý ‘slow’	998	63	povolaný ‘qualified’	91
4	malý ‘small’	25 034	34	mocný ‘powerful’	866	64	markantní ‘striking’	77
5	starý ‘old’	23 701	35	závažný ‘serious’	860	65	odrostlý ‘grown-up’	76
6	špatný ‘bad’	14 382	36	rozšířený ‘widespread’	855	66	sdílný ‘communicative’	55
7	nizký ‘low’	13 764	37	tmavý ‘dark’	784	67	smířlivý ‘conciliatory’	55
8	mladý ‘young’	12 300	38	početný ‘numerous’	512	68	hořejší ‘top, upper’	53
9	dlouhý ‘long’	9 540	39	prodáváný ‘best-selling’	509	69	roztodivný ‘odd’	45
10	důležitý ‘important’	9 229	40	zazší ‘later’	409	70	vroucný ‘fervent, dear’	41
11	blízký ‘near, close’	8 290	41	vlivný ‘influential’	407	71	sebenepatrný ‘slight’	41
12	silný ‘strong’	6 439	42	produktivní ‘productive’	371	72	žádoucný ‘desirable’	36
13	široký ‘wide’	5 290	43	žádaný ‘desired’	347	73	obsažný ‘comprehensive’	33
14	pozdní ‘late’	4 300	44	navštěvovaný ‘visited’	274	74	obsazovaný ‘cast’	32
15	jednoduchý ‘simple’	4 168	45	šetrný ‘friendly [environmentally], thrifty’	258	75	fajnový ‘fine’	32
16	rychlý ‘fast’	4 008	46	hodnotný ‘valuable’	247	76	benevolentní ‘benevolent’	28

¹⁰ We express our thanks to Dominika Kovářiková and Václav Cvrček.

	Lemma	Comparative + superlative forms		Lemma	Comparative + superlative forms		Lemma	Comparative + superlative forms
17	levný 'cheap'	3 955	47	sofistikovaný 'sophisticated'	230	77	vyznamenávaný 'honour'd'	26
18	krátký 'short'	3 798	48	chutný 'tasty'	218	78	stahovaný 'downloaded'	24
19	drahý 'dear, expensive'	3 694	49	náchylný 'susceptible'	202	79	otužilý 'hardy'	24
20	častý 'frequent'	3 433	50	frekventovaný 'busy'	175	80	vzrušivý 'exciting'	20
21	hluboký 'deep'	3 387	51	příhodný 'appropriate'	172	81	zavilý 'fierce, ferocious'	20
22	složitý 'complicated, complex'	2 963	52	vyhledávaný 'sought'	164	82	onaký 'another, better'	18
23	dřívější 'earlier, previous'	2 900	53	lidnatý 'populous'	162	83	skladný 'compact'	18
24	slabý 'weak'	2 815	54	čtený 'read'	135	84	obletovaný 'adored'	18
25	snadný 'easy'	2 495	55	palčivý 'burning'	131	85	hovorný 'talkative'	14
26	náročný 'difficult'	2259	56	výstižný 'concise'	125	86	pregnantní 'succint'	14
27	výhodný 'favourable, advantageous'	1537	57	niterný 'inner'	122	87	poslouchaný 'listened to'	11
28	efektivní 'effective'	1146	58	výnosný 'profitable'	122	88	křepký 'sprightly'	10
29	účinný 'effective'	1129	59	důrazný 'strong'	119	89	první 'first'	10
30	cenný 'valuable'	1087	60	schůdný 'viable'	110	90		

Tab. 1. List of adjectives that are commonly used in comparative or superlative forms (in the corpus SYN2015).

Thirty of these adjectives were analyzed in more detail. Those were the adjectives with the biggest proportion of comparative forms among their total occurrences. The proportion of comparatives was chosen as the basic criterion because comparative forms are more often lexicalized (see above, 1.2). The same adjectives are sorted both by the proportion of comparatives (in the left part of the table) and by the proportion of superlatives (in the right part of the table). The list of them mostly contains the most frequent adjectives. It is obvious that a high proportion of graded forms among the occurrences of certain adjectives is conspicuous. Units with the biggest difference between the proportion of comparatives and superlatives are in bold type.

	Lemma	Comparatives	% of comparatives	Total frequency		Lemma	Superlatives	% of superlatives	Total frequency
1	brzký 'early'	2 900	77,81	3 727	1	blízky 'near, close'	6 175	34,66	17 814
2	pozdní 'late'	4 276	60,33	7 088	2	častý 'frequent'	2 379	30,27	7 858
3	nízký 'low'	11 081	42,70	25 951	3	dobrý 'good'	29 806	25,59	116 490
4	levný 'cheap'	2 621	35,19	7 448	4	špatný 'bad'	6 298	20,69	30 447
5	podrobný 'detailed'	967	30,30	3 363	5	vysoký 'high'	15 781	20,09	78 558
6	vysoký 'high'	23 125	29,44	78 558	6	levný 'cheap'	1 334	17,91	7 448
7	špatný 'bad'	8 084	26,55	30 447	7	velký 'big, great'	33 233	17,38	191 203
8	starý 'old'	18 439	25,81	71 442	8	důležitý 'important'	6 490	15,86	40 918
9	složitý 'complicated, complex'	2 644	25,02	10 566	9	drahý 'dear, expensive'	1 544	13,55	11 392
10	široký 'wide'	4 529	24,14	18 760	10	nízký 'low'	2 683	10,34	25 951
11	slabý 'weak'	2 289	24,13	9 488	11	výhodný 'favourable, advantageous'	456	9,03	5 049
12	dobrý 'good'	25 934	22,26	116 490	12	efektivní 'effective'	342	7,75	4 413
13	snadný 'easy'	2 214	22,04	10 045	13	silný 'strong'	2 126	7,61	27 923
14	výhodný 'favourable, advantageous'	1 081	21,41	5 049	14	rychlý 'fast'	1 265	7,49	16 900
15	pomalý 'slow'	910	21,31	4 270	15	jednoduchý 'simple'	1 527	7,41	20 617
16	velký 'big, great'	38 820	20,30	191 203	16	starý 'old'	5 262	7,37	71 442
17	hluboký 'deep'	2 499	20,05	12 461	17	hluboký 'deep'	888	7,13	12 461
18	malý 'small'	18 864	19,72	95 652	18	malý 'small'	6 170	6,45	95 652
19	mladý 'young'	9 643	19,68	48 998	19	slabý 'weak'	526	5,54	9 488
20	drahý 'dear, expensive'	2 150	18,87	11 392	20	mladý 'young'	2 657	5,42	48 998
21	efektivní 'effective'	804	18,22	4 413	21	náročný 'difficult'	577	5,16	11 182
22	dlouhý 'long'	8 130	16,98	47 872	22	krátký 'short'	837	4,35	19 236
23	rychlý 'fast'	2 743	16,23	16 900	23	široký 'wide'	761	4,06	18 760
24	silný 'strong'	4 313	15,45	27 923	24	dlouhý 'long'	1 410	2,95	47 872
25	krátký 'short'	2 961	15,39	19 236	25	složitý 'complicated, complex'	299	2,83	10 566
26	náročný 'difficult'	1 682	15,04	11 182	26	snadný 'easy'	281	2,80	10 045
27	častý 'frequent'	1 054	13,41	7 858	27	pomalý 'slow'	88	2,06	4 270
28	jednoduchý 'simple'	2 641	12,81	20 617	28	podrobný 'detailed'	52	1,55	3 363
29	blízky 'near, close'	2 115	11,87	17 814	29	pozdní 'late'	24	0,34	7 088
30	důležitý 'important'	2 739	6,69	40 918	30	brzký 'early'	0	0,00	3 727

Tab. 2. List of thirty selected adjectives sorted by the proportion of comparatives (in the left part) and by proportion of superlatives (in the right part) in the total frequency.

2.2 Adjective gradation from the semantic point of view

Concerning the lexicographic treatment, we are interested in comparative and superlative forms being used “absolutely”, i.e. in examples without a general aspect of comparison (as pointed out above, comparison is not always explicitly expressed because of the occurrence of the average or usual value of the compared quality).

A high proportion of comparatives among the occurrences of a certain adjective suggests that the collocability of its comparative form compared with the collocability of both the positive form and the superlative form might need to be explored. The same should be done for superlative adjective forms, which need to be compared with both positive adjective forms and comparative forms. The collocability of positive, comparative and superlative forms with nouns is for some adjectives different (some examples see in C.2), for some adjectives identical (for some examples, see C.5).

The high proportion of comparative or superlative forms in Table 2 can be accounted for by:

(A) phraseological units:

– *hlad je nejlepší kuchař* ‘hunger is the best sauce’; *cesta nejmenšího odporu* ‘the line of least resistance’; *v nejhorším případě* ‘at worst’; *bližší košile než kabát* ‘blood is thicker than water’; *je snadnější uhlídat pytel blech, než...* ‘it’s impossible to keep tabs on it’; *je nejvyšší čas* ‘it’s high time’¹¹;

(B) multi-word names (mainly terminological units):

– *vyšší odborná škola* ‘higher vocational school’, *schůzka na nejvyšší úrovni* ‘top-level meeting’.

These structures with comparatives and superlatives (in both (A) and (B)) must enter the lexicon as a unit. The comparatives or superlatives are not substitutable by the positive, or the substituted version loses its idiomaticity (see [18]). In a dictionary these lexical units need to be registered with definitions.

(C) There is other group of comparatives and superlatives used with a meaning which cannot express comparison. Some of them are already listed in Czech monolingual dictionaries as separate entries (see 1.2), but some are not. The difference between the meaning of the graded forms and the meaning of the positive typically results from different collocability. Sometimes the adjective-noun combination with a comparative or superlative has a meaning different from the common comparison, or the comparative or superlative meaning is neutralized. We find:

(C.1) terminological units in which a comparative (or superlative) form is repeated in a systematic way, i.e. comparative (or superlative) form is used in multi-word

¹¹ Phraseological units and their corpus detection and their exploration were not our concern.

units of the same type. Some of the multi-word names can be registered within an “abstract” sense which associates (terminological) lexical units with similar meaning, e.g.:

– *jitrocel větší, bedrník větší, vlašovičník větší; citroník největší, liska největší* (binomial botanical names);

– *vyšší* ‘higher-level’ with nouns *rostliny* ‘plants’, *živočichové* ‘animals’, *organismy* ‘organisms’, *formy života* ‘forms of life’; *jednodušší* ‘less complex’ with nouns *sloučeniny* ‘compounds’, *formy života*; *složitější* ‘more complex’ with nouns *látky* ‘substances’, *organismy*;

– *mladší/starší žáci, mladší/starší dorostenci; nejmladší žáci* (adjectives *mladší* ‘younger’, *starší* ‘older’, *nejmladší* ‘the youngest’ in connection with nouns *žáci* ‘pupils’, *dorostenci* ‘adolescents’ express a well-defined age category);

(C.2)

– superlatives with different collocability compared with an adjective positive and comparative, e.g.: *nejbližší* ‘the nearest / the earliest’ collocates with the nouns *příležitost* ‘opportunity’, *jednání* ‘negotiation’, *(možný) termín* ‘possible date’ etc., the positive *blízký* and the comparative *blíže* are rare; similarly *nejvyšší štěstí/blaho* ‘absolute happiness/welfare’;

– the same with comparatives, e.g.: *hlubší* ‘deeper’ collocates with the nouns *význam* ‘meaning’, *smysl* ‘sense’, *analýza* ‘analysis’ etc., the comparative form is the most frequent (the positive *hluboký* and the superlative *nejhlubší* are rare);

– superlative and positive forms can collocate with the same nouns, different from comparative forms, e.g. *nejdražší* ‘the dearest’ collocates with *přítel* ‘friend’ and with family member nouns (*rodiče* ‘parents’, *maminka* ‘mum’, *tatínek* ‘dad’), and is thus similar in collocability to the positive *drahý* but different from the comparative *dražší*; *nejhlubší* ‘the deepest’ collocates with *úcta* ‘reverence’, *soustrast* ‘condolence’; *noc* ‘night’, *tajemství* ‘secret’, *nitro* ‘heart’, and is thus similar in collocability to the positive form, but different from the comparative form;

(C.3) comparatives which are used in the euphemistic way:

– *slabší povaha* ‘a weaker man’, *slabší zdraví* ‘rather poor health’; *má silnější postavu* ‘he/she is rather fat’; *nejsem už nejmladší* ‘I’m not that young’;

(C.4) comparatives which are very frequently used absolutely and their meaning is lexicalized in a “terminological” way, e.g. *kratší* ‘not long’, *delší* ‘not short’ – *kratší/delší vlasy, na kratší/delší vzdálenosti*.¹²

(C.5) A big number of graded forms are not indicative of semantic changes. The collocability of positive, comparative and superlative forms is also similar, e.g. collocability of *podrobný* ‘detailed’, *podrobnější*, and *nejpodrobnější* is basically the

¹² In C.1 – C.4 we list mainly examples of lexicalized comparatives or superlatives which are not yet registered in Czech monolingual dictionaries.

same – among the ten most frequent noun collocates are *informace* ‘information’, *popis* ‘description’, *zpráva* ‘message’, and *vyšetření* ‘examination’.

We propose that the examples listed in C.1 – C.4 need to be registered in monolingual dictionaries. This can be done in several ways (in a separate entry, in the exemplification section by means of an additional definition or by a comment relating to limited collocability), according to conceptual principles of the particular situation.

3 CONCLUSION

This paper deals with the treatment of lexicalized comparative and superlative forms in Czech monolingual dictionaries, which appears insufficient. We look for corpus signals providing clear evidence of the lexicalization of absolute comparatives or superlatives. High frequency of these forms can indicate idiomaticity and semantic changes. However, the most prominent signal for lexicalization assessment of a comparative or superlative adjective is its collocability.

References

- [1] Buzássyová, K. (1979). Príspevok k vymedzeniu neutralizácie v kategórii stupňovania. *Jazykovedný časopis*, 30(1), pages 6–17.
- [2] Buzássyová, K. (1st and 2nd volumes) and Jarošová, A. (1st, 2nd and 3rd volumes) (eds.) (2006, 2011, 2015). *Slovník súčasného slovenského jazyka*. A – G. [1st volume]. H – L. [2nd volume]. M – N. [3rd volume]. Bratislava, Veda.
- [3] Cvrček, V. et al. (2010). *Mluvnice současné češtiny 1. Jak se píše a mluví*. Praha, Karolinum.
- [4] Čechová, M., Dokulil, M., Hlavsa, Z., Hrbáček, J., and Hrušková, Z. (2011). *Čeština – řeč a jazyk*. 3rd revised edition. Praha, SPN – pedagogické nakladatelství, a. s.
- [5] Dokulil, M., Horálek, K., Hůrková, J., and Knappová, M. (eds.) (1986). *Mluvnice češtiny 1. Fonetika. Fonologie. Morfonologie a morfemika. Tvoření slov*. Praha, Academia.
- [6] Filipec, J., Daneš, F., Machač, J. (1st edition), and Mejstřík, V. (2nd and 3rd edition) (eds.) (1st edition, 1978; 2nd revised edition, 1994; 3rd revised edition, 2003): *Slovník spisovné češtiny pro školu a veřejnost*. Academia, Praha. [CD-ROM] (1997, 2004, 2005). Voznice, LEDA.
- [7] Havránek, B., Bělič, J., Helcl, M., Jedlička, A., Křístek, V., and Trávníček, F. (eds.) (1960 – 1971). *Slovník spisovného jazyka českého*. Praha, Nakladatelství.
- [8] Hujer, O., Smetánka, E., Weingart, M., Havránek, B., Šmilauer, V., and Ziskal, A. (eds.) (1935–1957). *Příruční slovník jazyka českého*. Praha, Státní nakladatelství, Školní nakladatelství, Státní pedagogické nakladatelství.
- [9] Internetová jazyková příručka. Ústav pro jazyk český, v. v. i., Praha. Accessible at: <<https://prirucka.ujc.cas.cz>>.
- [10] Internetový slovník současné češtiny. Brno, Lingea. Accessible at: <<https://www.nechybujte.cz/slovník-soucasne-cestiny/>>.
- [11] Karlík, P., and Hladká, Z. (2004). Kam s ním (problém stupňování adjektiv). In *Život s morfémou*. Sborník studií na počest Zdenky Rusinové, pages 73–93. Eds. P. Karlík and J. Pleskalová. Brno, Masarykova univerzita.

- [12] Karlík, P., Nekula, M., and Rusinová, Z. (eds.) (2012). Příruční mluvnice češtiny. 2nd revised edition. Praha, Nakladatelství Lidové noviny.
- [13] Karlík, P., Nekula, M., and Pleskalová, J. (2017). CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: <<https://www.czechency.org/>>.
- [14] Komárek, M., Kořenský, J., Petr, J., and Veselková, J. (eds.) (1986). Mluvnice češtiny 2. Praha, Academia.
- [15] Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářiková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondříčka, P., and Zasina, A. Korpus SYN, verze 7 z 29. 11. 2018. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <<http://www.korpus.cz>>.
- [16] Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářiková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondříčka, P., and Zasina, A. (2015). SYN2015: reprezentativní korpus psané češtiny. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <<http://www.korpus.cz>>.
- [17] Křivan, J. (2012). Komparativ v korpusu: explanace morfemtické struktury českého stupňování na základě frekvence tvarů. Slovo a slovesnost, 73(1), pages 13–45.
- [18] Panevová, J. (2007). Gradation of adjectives and valency. In Gramatika a korpus. Grammar & Corpora 2005, pages 197–204. Eds. F. Štícha and J. Šimandl. Praha, Ústav pro jazyk český AV ČR.
- [19] Panevová, J. (2008). Povaha stupňování adjektiv (K „nesrovnávacímu“ užití stupňovaných forem). In Iugi Observatione: Zborník z konferencie Jazyk – kultúra – spoločnosť, venovanej 80. narodeninám prof. PhDr. L. Ďuroviča, pages 149–156. Ed. S. Ondrejovič. Bratislava, Veda.
- [20] Světlá, J. (2016). Stupňování přídavných jmen a příslovčí. In Kapitoly z koncepce Akademického slovníku současné češtiny, pages 62–63. Eds. P. Kochová and Z. Opavská. Praha, Ústav pro jazyk český AV ČR, v. v. i. Accessible at: <<https://lur1.cz/SMMCN>>.
- [21] Štícha, F. (ed.) (2013). Akademická gramatika současné češtiny. Praha, Academia.

ON THE VALENCY OF VARIOUS TYPES OF ADVERBS AND ITS LEXICOGRAPHIC DESCRIPTION

JAKUB SLÁMA – BARBORA ŠTĚPÁNKOVÁ

Czech Language Institute of the Czech Academy of Sciences, Prague, Czech Republic
Charles University, Prague, Czech Republic

SLÁMA, Jakub – ŠTĚPÁNKOVÁ, Barbora: On the valency of various types of adverbs and its lexicographic description. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 158 – 169.

Abstract: This paper deals with the neglected issue of the valency of adverbs. After providing a brief theoretical background, a procedure is presented of extracting the list of potentially valent adverbs from two syntactically parsed corpora of Czech, SYN2015 and PDT. Taking note of the methodological and theoretical problems surrounding this task, especially those relating to the fuzzy boundaries of word classes, we outline the types of adverbs identified as having valency properties. Where appropriate, we comment on – and occasionally suggest improvements in – the lexicographic treatment of valent adverbs.

Keywords: adverbs, valency, dictionary, syntactically parsed corpora, Czech

1 INTRODUCTION

Valency is undoubtedly a central topic in syntax and “a primary concern of all approaches to the grammar of human languages” [1, p. 39]. As Spevak [2, p. ix] puts it, however, the valency of nouns has remained “in the shadow of the valency of verbs,” and the same arguably applies to the valency of adjectives, and to an even more serious extent to the valency of adverbs, which is only occasionally paid lip service to but remains conspicuously underresearched. We therefore provide an overview of how the notion of valency has broadened to cover not only verbs, and of what has been written about the valency of adverbs in Czech. Then we propose a way of extracting a list of potentially valent adverbs from syntactically parsed corpora of Czech, and we outline the types of adverbs which appear to have valency properties, touching upon some problematic points concerning the fuzzy boundaries of linguistic categories (cf. for instance [3, pp. 568–570]), and, where appropriate, commenting on the lexicographic description of the adverbs under discussion.

2 THEORETICAL BACKGROUND

2.1 The scope of the notion of valency¹

Valency can be delineated as “the capacity a verb (or noun, etc.) has for combining with particular patterns of other sentence constituents” [5, p. 301]. Tesnière [6, p. 670] famously defined valency as “nombre d’actants qu’un verbe est susceptible de régir”, i.e. as a property of verbs. The very next year saw the publication of an early transformational generativist account of nominalizations, which includes some interesting observations on the valency of nominals [7, pp. 66–73]. However, it was only later that it began to be widely recognized that valency is not restricted to verbs; for instance, Matthews [8, p. 115] – still somewhat cautiously – states that adjectives can also have “semantic properties akin to valency.” Over a decade later, Matthews [9, p. 394] defines valency as a property of “a verb or other lexical unit.” This is in accord with a more general trend towards admitting lexical units other than verbs into the description of valency, which resulted in valency being redefined generally as “the number and type of bonds which syntactic elements may form with each other” [10], which, according to Matthews [11, p. 4], suggests that valency should not be the foundation “just of the syntax of verbs, or of verbs and other lexical units, but of syntax generally”. Nevertheless, the valency of adverbs is still a rather neglected area: for instance, couched within the Pattern Grammar, there is a two-volume grammar of the patterns for verbs [12] and nouns and adjectives [13], with adverbs not attended to. On the other hand, the database FrameNet does include some adverbs, whereby, however, these are only said to evoke frames (e.g. the adverb in *Bill wisely sold the piano* is said to evoke the Mental property frame), which hardly amounts to postulating a valency frame for such adverbs.

2.2 The valency of adverbs in Czech linguistics

Adverbs in Czech have been noted to have valency properties; Karlík & Biskup [14] give the example of *nezávisle* (*Jednal nezávisle na rodičích*. ‘He acted independently of his parents’), while noting that the Czech grammatical description has largely neglected the valency of adverbs. Panevová [15, p. 7] briefly mentions the valency of adverbs when noting that we have to acknowledge not only that deadjectival adverbs might have valency properties, but also that primary, non-derived adverbs might require obligatory complements, giving the example of *daleko* ‘far’, to which she applies the question test: *Je to daleko. Odkud (od čeho)? *Nevím*. lit. ‘It is far. From where? *I don’t know.’.

In older Czech dictionaries, derived adverbs are not provided with entries of their own; they are nested with their adjectival bases. For instance, in *Slovník spisovného jazyka českého*, *nezávisle* is nested in the entry for *nezávislý* ‘independent’, whose

¹ Parts of this section are loosely based on [4, esp. 2.2.1].

colligation (*na kom, čem* ‘of sb, sth’) is nevertheless listed in the entry. It is assumed that the adverb can take complements with the preposition *na* followed by a locative form as well, as shown by the example given (*vznikat nezávisle na něčem* ‘arise independently of’). In *Akademický slovník současné češtiny* (hereafter ASSČ)² deadjectival adverbs have their own entries [16, p. 123], and so they might have – as the only type of adverbs [16, p. 68] – their own valency specification (nevertheless, the dictionary does not list only obligatory complements but also typical modifiers, without the two being explicitly distinguished). Thus, for instance, the adjective *bezohledný* ‘reckless’ and the adverb *bezohledně* ‘recklessly’ have (*ke komu, k čemu; vůči komu, vůči čemu* ‘with sb, sth; to sb, sth’) listed as their complements.

3 DATA & METHODOLOGY

In an attempt to arrive at as complete a list of valent adverbs in Czech as possible in a way as objective as possible, we decided to rely on the two syntactically annotated corpora of Czech available, i.e. the automatically parsed SYN2015 of the Czech National Corpus project and the manually annotated Prague Dependency Treebank (PDT). In order to find valent adverbs we searched for nodes whose parent nodes are occupied by adverbs.

3.1 SYN2015

In the corpus SYN2015, we used the following query:

```
1:[tag="D.*"] []{0,5} 2:[p_tag="D.*"] & 1.lemma=2.p_lemma within <s />
```

The results were then checked manually. Tagging and parsing errors were excluded, including, for instance, the following:

- (1) *Efekt chyb DNA na případný vývoj nemocí v mozku je také **jednou** z oblastí, která bude v hledáčku.*³ [*jednou* mistagged as an adverb]
‘The effect of DNA errors on the possible development of diseases in the brain is also **one of** the areas which are going to be monitored.’
- (2) *Učenci jsou vlídní a přívětiví, hledají, **jak být užiteční bližním i** jak přispět k obecnému blahu.* [*jak* parsed as the parent node of *i*]
‘Scholars are kind and friendly; they are looking for ways **how to be helpful to their fellows and** how to contribute to the public good.’

Similarly, we excluded concordances in which there is a terminal punctuation mark between the node and its alleged daughter. Next, we excluded concordances in which the adverb seems to have a complement, but this dependent is part of a larger syntactic construction, most notably the comparative construction:

² See <<http://www.slovníkcestiny.cz>>.

³ Unless explicitly stated otherwise, the examples come from SYN2015.

- (3) *nikdo neumí tabuli umýt **tak dokonale jako já***
'no one can clean the blackboard as **perfectly as me**'

Similarly, we excluded other comparative constructions and similes, as in (4). Furthermore, we excluded expressions that can be classified as secondary prepositions, e.g. *včetně* (5), linking adverbials, e.g. *jmenovitě* (6) and *resp.* (7), quantity expressions, e.g. *hodně* (8) and *drahně* (9), and particles, e.g. *samozřejmě* (10) and *už* (11):

- (4) *Plížíš se po domě **tíše jako** kočka a já si toho hned všimnu.*
'You are creeping about the house as **soundlessly as** a cat and I immediately notice that.'
- (5) *Z celého Izraele bylo přijato pouze dvacet čtyři lidí **včetně Suzany.***
'Of the whole of Israel only 24 people were accepted **including Suzane.**'
- (6) *Velké šelmy, **jmenovitě tygři**, na ostrovy nepronikly.*
'Big cats, **namely tigers**, did not reach the islands.'
- (7) *je nežádoucí, **resp. nepotřebný***
'he is undesirable, **or rather useless**'
- (8) *Jezte **hodně vlákniny.***
'Eat **a lot of fibre.**'
- (9) *Tři dny mě vídal, je to **drahně let.***
'He was seeing me for three days, it's been **many years.**'
- (10) ***Samozřejmě peněz** by mohlo a mělo být více.*
'**Naturally** there could, and should, be more **money.**'
- (11) ***Už měsíce** ji uháněl, aby s ním šla na skleničku...*
'He's been trying to get her to go out for a drink with him **for months.**'

Finally, we naturally excluded words that can be considered adverbs, but that cannot be deemed to have valency (such as *francouzsky*, *krátce*, and *maloplošně* in (12)–(14)):

- (12) *musíme udělat nádivku (**francouzsky farce**)*
'prepare the stuffing (**farce in French**)'
- (13) *Za svou snahu však byl odměněn **krátce před** pauzou.*
'was however rewarded for his efforts **shortly before** the break.'
- (14) *Celek je chráněn **maloplošně jako** PR Křížová cesta a je součástí CHKO Broumovsko.*
'The whole is protected **on a small-scale as** the Nature Reserve Křížová cesta and is part of the Broumovsko Protected Landscape Area.'

Overall, we discarded hundreds of adverb candidates: the initial list of parent adverbs in SYN2015 included 1 810 words tagged simultaneously as adverbs and parent nodes at least in three instances.

3.2 PDT

In the PDT, we used the following query:

```
a-node $adv := [  
  m/tag ~ '^D',  
  echild a-node $ch := [  
    ! afun in { 'AuxZ', 'ExD', 'AuxG', 'Adv', 'AuxX', 'AuxY', 'AuxK' }  
  ]  
]  
>> $adv.m/form  
>> $1, count(1 over $1)  
>> distinct $1, $24
```

The results include adverbs that function as parent nodes for auxiliary nodes, e.g. *apod*, which is always the parent of a node corresponding to a full stop. Moreover, the query results included adverbs in coordination with another adverb yet without any complement; these results were naturally excluded as well (including e.g. *citlivěji* 'sensitively' in *obezřetněji a citlivěji* 'more carefully and sensitively'). Another group of cases that were discarded includes foreign language expressions, such as *memoriam* in *in memoriam* (with *in* parsed as a dependent of *memoriam*), and words mistagged as adverbs, such as *ostrožně* in *Jeruzalém byl postaven na ostrožně Chrámové hory* 'Jerusalem was built on the promontory of the Temple Mount'. We also excluded instances in which there is a dependent of the adverb under scrutiny, but this dependent functions as a modifier rather than a complement (e.g. *zcela běžně* 'quite routinely' with *běžně* as the parent node). Finally, we also excluded cases such as the following, in which it is clear that the word in bold cannot be deemed a valent adverb:

(15) *A právě ono **kdy** vám dokážeme říci.*

'And we are able to tell you just this **when**.'

(16) ***Jinak** – a řekněme hned, že mnohem hůře – je na tom čtenář denního tisku.*

'The reader of the daily press is doing **differently** – and let us say straight away that much worse.'

(17) *Piloty jsou zaberaněny shora **dolů** do bažiny...*

'Stilts are stuck from above **down** to the swamp.'

This produced a list of only nine potentially relevant adverbs, all of which were also arrived at when working with SYN2015, viz. the predicative adverbs *líto* 'sorry' and *oblačno* 'cloudy', and the adverbs *přiměřeně* 'adequately', *nezávisle* 'independently', *stejně* 'just (as)', *západně* 'west', *úměrně* 'proportionately', *blízko* 'near', and *daleko* 'far'.

⁴ The query can be executed at <<http://lindat.mff.cuni.cz/services/pmltq/#!/treebank/pdt30/query/>>.

4 RESULTS

The results included various types of adverbs, which we discuss one by one in the following sections. In each section we give examples which should be clearly indicative of why we consider the respective adverbs to be valent. For instance, *kolmo* ‘perpendicularly’ is included in the second section with the following example:

- (18) *Holeně zůstávají kolmo k zemi, chodidla umístěte na šířku kyčelních kloubů.*
‘Keep your shins **perpendicular** to the ground, and place your feet straight below your hips.’

Since we consider (18a) and (18b) not to be simultaneously acceptable and equivalent in meaning to (18), *kolmo* is a valent adverb in our view:

- (18a) **Holeně zůstávají kolmo, chodidla umístěte na šířku kyčelních kloubů.*
‘Keep your shins perpendicular, and place your feet straight below your hips.’
(18b) **Holeně zůstávají k zemi, chodidla umístěte na šířku kyčelních kloubů.*
‘Keep your shins to the ground, and place your feet straight below your hips.’

4.1 Deadjectival adverbs typically used as adverbials

This group includes several adverbs that can be grouped semantically as follows:

a) adverbs expressing spatial relations (with possible metaphorical semantic extensions): *kolmo* ‘perpendicularly’, *paralelně* ‘parallel’, *rovnoběžně* ‘parallel’, *nalevo* ‘left’, *napravo* ‘right’, *vlevo* ‘left’, *vpravo* ‘right’, *vodorovně* ‘horizontally’, *symetricky* ‘symetrically’, *příčně* ‘diagonally’

- (19) *Vleče mě to stranou, paralelně s pobřežím, přímo do cesty jachty hnané větrem.*
‘It’s dragging me aside, **parallel** to the coastline, straight in the way of a yacht driven by the wind.’

- (20) *Umístil jej v blízkosti hlavní brány, vlevo od příjezdové cesty.*
‘He placed it nearby the main gate, to the **left** of the driveway.’

- (21) *Symetricky k očekávání budoucnosti existuje podle našeho soudu rovněž “očekávání minulosti”.*
‘**Symetrically** to the future expectations there exist, in our view, “past expectations”.’

Other potentially valent adverbs (esp. *daleko* ‘far’, *blízko* ‘near’, and *nablízku* ‘close’) were discarded as prepositions, cf. [17, pp. 39–40].

b) directional adverbs requiring a complement with the preposition *od* followed by a genitive form: *jižně* ‘south’, *severně* ‘north’, *západně* ‘west’, *východně* ‘east’,

jihovýchodně ‘southeast’, *jihozápadně* ‘southwest’, *severovýchodně* ‘northeast’, *severozápadně* ‘northwest’, *jihojihovýchodně* ‘southsoutheast’, *západoseverozápadně* ‘westnorthwest’

(22) **Západně** od věže máme kostel svatého Martina s okolními stavbami.

‘West of the tower there is the church of Saint Martin with the surrounding buildings.’

c) adverbs of comparison: *analogicky*, *odlišně*, *relativně*, *proporčně*, *protikladně*, *srovnatelně*

(23) *pojímá diskusi radikálně odlišně od všech předřečníků* (SYN V7)

lit. ‘he conceives of the discussion radically **differently** from all the previous speakers’

(24) *ČSSD i ANO hlasovaly protikladně ke svým opakovaným veřejným prohlášením* (SYN V7)

lit. ‘The Czech Social Democratic Party and the party ANO voted **oppositely** to their repeated public proclamations.’

(25) *Figuríny byly použity tím nejpopsnějšíším způsobem, vlastně protikladně Sadeově nespoutané fantazii.* (SYN V7)

lit. ‘The manikins were used in a most prosaic way, or **oppositely** to Sade’s unbridled fantasies.’

(26) *Bulharsko je na tom srovnatelně s námi.*

‘Bulgaria is doing **comparably well** to us.’

Note that some adverbs (such as *protikladně* or *přiměřeně* below) can have complements of more than one form, as witnessed by (24) and (25) or (29) and (30) below.

There are several other adverbs that inherently express comparison and are typically followed by *jako* ‘as’ or *než* ‘than’, namely *stejně* ‘just’, *podobně* ‘similarly’, *přesně* ‘exactly’, *obdobně* ‘similarly’, *opačně* ‘oppositely’, *obráceně* ‘backwards’, *rozdílně* ‘differently’, *identicky* ‘identically’, *jinak* ‘otherwise’, *nejinak* ‘likewise’, *jinde* ‘elsewhere’, *jinam* ‘elsewhere/away’, *jindy* ‘another time’, *odjinud* ‘from elsewhere’, *jinudy* ‘another way’, *jináč* ‘otherwise’, and *od(ni)kud jinud* ‘from (no)where else’. It is somewhat questionable whether or not these adverbs should be viewed as having valency properties. There are also expressions such as *shodně* (s) ‘consistently (with)’, *souhlasně* (s) ‘in agreement (with)’, and *úměrně* s ‘proportionately’, which can be viewed as secondary prepositions [17, pp. 43–44].

d) others: *nezávisle* ‘independently’, *odděleně* ‘separately’, *přiměřeně* ‘adequately’, *loajálně* ‘loyally’

- (27) *Šlechta a města **nezávisle** na státu vydávala správní předpisy.*
 ‘The nobility and the town issued administrative legislation **independently** of the state.’
- (28) *mundumugu vždycky bydlí a jí sám, **odděleně** od svého lidu*
 ‘the mundumugu always lives and dines alone, **separately** from his people’
- (29) *Totéž platí pro kořata, která mají být čilá **přiměřeně** věku.*
 ‘The same holds for kittens, which are supposed to be alert **adequately** to their age.’
- (30) *Určíme, co by měl žák v TV **přiměřeně** k jeho věku a výstupům dle RVP ZV znát.*
 ‘We determine what the pupil should know in PE **adequately** to his age and the curriculum framework.’
- (31) *Když už státní zaměstnanci přece protestovali, tak **loajálně** ke státu v neděli odpoledne, jak tomu bylo 6. listopadu 1910 v Brně.*
 ‘When civil servants finally protested, they did so **loyally** to the state on a Sunday afternoon, which happened on November 6, 1910, in Brno.’

4.2 Non-derived adverbs

The adverbs *pozdě* ‘late’ and *brzy* (*brzo*) ‘early’ belong to this group, cf. the following examples:

- (32) *je příliš **pozdě** na jezevčíky s jejich důchodci, příliš **brzy** na milence*
 ‘it’s too **late** for dachshunds with their pensioners, and too **early** for lovers’
- (33) *zemřela příliš **brzy** na to, aby stihla na toto téma s dcerou promluvit*
 ‘she died too **early** to talk to her daughter about that’
- (34) ***Včas** na to, aby pochopili, oč jde, **pozdě** na to, aby se jich to týkalo.*
 ‘**Just in time** to understand what was happening, too **late** for it to concern them.’

Ex. (34) suggests that *včas* should also be included as a valent adverb, albeit not non-derived.

4.3 Predicative adverbs

The SYN2015 data included the following predicative adverbs: *potřeba* ‘necessary’; *líto* ‘sorry’; *zapotřebí* ‘needed/necessary’; *zataženo* ‘cloudy’; *zima* ‘cold’; *horko* ‘hot’; *netřeba* ‘needless’; *rušno* ‘busy’; *zle* ‘sick’; *mokro* ‘wet’; *vlhko* ‘damp’; *živo* ‘lively’; *oblačno* ‘cloudy’; *polojasno* ‘somewhat cloudy’; *potřebí* ‘need’ (e.g. *nebylo potřebí fantazie* ‘there was no need of imagination’); *smutno* ‘sad’; *sucho* ‘drought’; *veselo* ‘merry’; *parno* ‘hot’; *pusto* ‘desert’; *temno* ‘dark’; *větrno* ‘windy’; *teplo* ‘hot’; *ticho* ‘silent’; *(být) libo* ‘wish’. Considering examples such as (35), in which *s* is misparsed as a daughter node of *zataženo* ‘cloudy’, it seems clear that these adverbs do not take complements.

- (35) *Oba dva dny by mělo být zataženo s občasnými dešťovými přeháňkami.*
'On both of the days it should be **cloudy** with occasional showers of rain.'

Nevertheless, it is somewhat conspicuous that predicative adverbs referring to weather and the like quite systematically co-occur with locative (or, possibly, temporal) adjuncts, which was in fact noted as early as by Komárek [18, p. 23]. Similarly, predicative adverbs expressing physical sensations and internal states systematically co-occur with a dative form referring to the experiencer, as in (36), and modal predicative adverbs systematically co-occur with infinitives (and, in some cases, with nouns), as in (37):

- (36) *Občas se při tom přiotrávím acetonem a je mi potom zas nějakou dobu zle.*
'Sometimes when doing it, I get poisoned with acetone and I feel **sick** for some time.'
- (37) *Je jenom zapotřebí promluvit lidem do srdce.*
'It is **necessary** to touch people's hearts with your words.'

Based on that, the question of whether these adverbs really lack valency might arise. We believe that the predicative adverbs do not have valency properties; their systematic co-occurrence with a copula and a certain type of adjunct is, in our view, the result of the existence of independent constructions in the sense of Construction Grammar [19]. That is, we believe that there is for instance a construction (i.e. a Saussurean sign) with the form [[copula] [locative adjunct] [predicative adverb]], which systematically expresses the meaning of 'there is the specified kind of weather or external state in the specified area' and which licenses constructs such as (38). Postulating such a construction is in accord with the fact that the construction can be used somewhat productively, as witnessed by examples such as the following:

- (38) *když kvetly, bylo tam bílorůžovo* (SYN V7)
'when they were blooming, it was all **white-pinkish** there'
- (39) *v metropoli se začínají rozkládat mrtvá těla a je tam neobyvatelno* (SYN V7)
lit. 'in the capital dead bodies are starting to decompose and it is **inhabitable** there'

In ASSČ, predicative adverbs are marked by the word-class specification "přísl. v přísudku," and their valency specification states the type of adjunct that typically co-occurs with the adverb; e.g., *bílo* '(all) white' has the valency specification (kde).

While we said that we do not consider predicative adverbs to have valency properties in the usual sense, there is one exception, found in SYN2015: *libo*, cf. the following example:

- (40) *Je libo porci za deset krejcarů?*
'Would you like a portion for ten pennies?'

Arguably, while *libo* is used in the same syntactic configuration as other predicative adverbs, it actually differs in that it takes a complement (*porci* in (40)).

5 CONCLUSION AND DISCUSSION

In this paper, we used two syntactically annotated corpora of Czech to extract a list of potentially valent adverbs in Czech. We manually discarded hundreds of candidates and arrived at several formal and semantic groups of valent adverbs, as presented above. There are some notoriously problematic issues, most notably concerning the fuzzy border between adverbs and prepositions. Words like *navzdory* 'notwithstanding' and *kolem* 'around', which are usually treated in terms of adverb-preposition homonymy, were not discussed here either even though they would certainly benefit from further investigation.

Finally, we would like to make three remarks. First, while we see *nezávisle* and other words discussed above as adverbs with valency, some linguistics treat e.g. *nezávisle na* 'independently of/from' as a secondary preposition (cf. [20, p. 511] or [21, p. 47]). To give two further examples, while we included *paralelně* (cf. example 19), Kroupová [17, p. 42] includes *paralelně s* 'parallel to' as a secondary preposition with the lowest degree of conventionalization, and, for instance, Blatná [21, p. 47] includes *úměrně s* 'proportionately to' as a secondary preposition.⁵ We still included these as adverbs with valency based on our intuitions that the adverbs involved were still notably "adverby" (esp. when functioning as clause elements and not being delexicalized) and based on examples such as (41), in which *úměrně s* can hardly be viewed as a secondary preposition:

- (41) *Neschopnost uvěřit ve smrt roste přímo úměrně s tím, jak se blíží.*
'One's disbelief in death grows in **proportion to** its approach.'

Note, however, that this does not in fact exclude the possibility of viewing *úměrně s* as a secondary preposition in other contexts.

Second, it is typically assumed that derived words "inherit" the valency specification from their respective bases, an assumption especially common in noun valency studies; on argument inheritance see e.g. [22, p. 215]. This assumption is,

⁵ Blatná [2, p. 47] lists all the following expressions as secondary prepositions: *nalevo od* 'left of', *napravo od* 'right of', *nezávisle na* 'independently of', *paralelně s* 'parallel to', *shodně s* 'consistently with', *souběžně s* 'parallel to', *současně s* 'simultaneously to', *souhlasně s* 'in agreement with', *společně s* 'together with', *spolu s* 'together with', *úměrně s* 'proportionately to', *úměrně k* 'proportionately to', *zároveň s* 'along with'.

however, in fact problematic, and it is questionable whether we should specify the valency of derived words with or without recourse to derivation, as suggested by remarks made, among others, by Herbst [23, p. 267], Williams [24, p. 584], Allerton [5, p. 311], and Goldberg [25, p. 24], as summarized in Sláma [4, pp. 27–28]. While this issue is definitely beyond the scope of the present paper, it is still worth pointing out that the directional adverbs mentioned above suggest that the issue of inheritance is not as straightforward as it is usually assumed to be. While e.g. *severně (od)* ‘north (of)’ is usually described as a derivative of *severní*, the adverb does not seem to inherit the valency frame of the adjective but perhaps of the corresponding noun (*na sever od* ‘north of’).

Third, related to the previous point is the question of the lexicographic treatment of adverbial valency: while the current practice in the ASSČ allows only deadjectival adverbs to have their valency specification [16, p. 68] but also deals with predicative adverbs as if they had valency, it seems reasonable to conclude that some adverbs whose corresponding adjectives presumably lack valency properties (e.g. *severně*) should be described as valent as well.

ACKNOWLEDGEMENTS

The work has been in part supported by the LINDAT/CLARIN and LINDAT/CLARIAH-CZ projects of Ministry of Education, Youth and Sports of the Czech Republic (LM2015071 and LM2018101).

References

- [1] Thompson, S., and J. Hopper (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In J. Bybee and P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*, pages 27–60. Amsterdam, John Benjamins.
- [2] Spevak, O. (2014). Editor’s foreword. In O. Spevak (ed.), *Noun Valency*, ix–xiii. Amsterdam & Philadelphia, John Benjamins Publishing Company.
- [3] Taylor, J. R. (2015). Prototype effects in grammar. In E. Dabrowska & D. Divjak (eds.), *Handbook of Cognitive Linguistics*, pages 562–579. Berlin; Boston, De Gruyter Mouton.
- [4] Sláma, J. (2018). The prepositional phrase with the preposition at as a valency complement of nouns. MA thesis. Prague, Department of English Language and ELT Methodology.
- [5] Allerton, D. J. (2006). Valency Grammar. In K. Brown (ed.), *Encyclopedia of Language & Linguistics*, pages 13, 301–314. 2nd ed. Amsterdam, Elsevier.
- [6] Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris, Klincksieck.
- [7] Lees, R. B. (1960). *The Grammar of English Nominalization*. The Hague, Mouton.
- [8] Matthews, P. H. (1981). *Syntax*. Cambridge, Cambridge University Press.
- [9] Matthews, P. H. (1997). *The Concise Oxford Dictionary of Linguistics*. Oxford, Oxford University Press.

- [10] Crystal, D. (2003). *A Dictionary of Phonetics and Linguistics*. 5th ed. Oxford, Blackwell.
- [11] Matthews, P. H. (2007). The scope of valency in grammar. In T. Herbst & K. Götz-Votteler (eds.), *Valency: Theoretical, Descriptive and Cognitive Issues*, pages 3–14. Berlin; New York, Mouton de Gruyter.
- [12] Francis, G., S. Hunston, and E. Manning (1996). *Collins Cobuild Grammar Patterns 1: Verbs*. London, HarperCollins.
- [13] Francis, G., S. Hunston, and E. Manning (1998). *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London, HarperCollins.
- [14] Karlík, P., and P. Biskup (2017). Adverbium. In P. Karlík, M. Nekula, and J. Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. Brno, Masarykova univerzita. Accessible at: <<https://www.czechency.org/slovník/ADVERBIUM>>.
- [15] Panevová, J. (1998). Ještě k teorii valence. *Slovo a slovesnost* 59(1), pages 1–14.
- [16] Kočová, P., and Z. Opavská (eds.) (2016). *Kapitoly z koncepce Akademického slovníku současné češtiny*. Praha, Ústav pro jazyk český AV ČR.
- [17] Kroupová, L. (1985). *Sekundární předložky v současné spisovné češtině*. Praha, ÚJČ ČSAV.
- [18] Komárek, M. (1954). K otázce predikativa (kategorie stavu) v češtině. In *Sborník VŠ pedagogické v Olomouci (Jazyk a literatura)*, pages 7–25. Praha, Státní pedagogické nakladatelství.
- [19] Hoffmann, T., and G. Trousdale (2013). *The Oxford Handbook of Construction Grammar*. New York, Oxford University Press.
- [20] Štícha, F. (ed.). *Akademická gramatika spisovné češtiny*. Praha, Academia.
- [21] Blatná, R. (2006). *Víceslovné předložky v současné češtině*. Praha, Nakladatelství Lidové noviny.
- [22] Booij, G. (2007). *The Grammar of Words*. Oxford, Oxford University Press.
- [23] Herbst, T. (1988). A valency model for nouns in English. *Journal of Linguistics*, 24(2), pages 265–301.
- [24] Williams, E. (1991). Meaning Categories of NPs and Ss. *Linguistic Inquiry* 22(3), pages 584–587.
- [25] Goldberg, A. E. (2006). *Constructions at Work*. Oxford, Oxford University Press.

THE SYNCHRONIC DYNAMICS OF WORDS ENDING IN *-ita/-ost'*

MÁRIA ŠIMKOVÁ

Slovak National Corpus, L. Štúr Institute of Linguistics, Slovak Academy of Sciences,
Slovakia

ŠIMKOVÁ, Mária: The synchronic dynamics of words ending in *-ita/-ost'*.
Journal of Linguistics, 2019, Vol. 70, No 2, pp. 170 – 179.

Abstract: This paper focuses on the potential of using corpora to study manifestations of the synchronic dynamics of language and on the analysis of how words with the suffixes *-ita/-ost'* function in contemporary texts. The analysis is based on data from the Slovak National Corpus: the corpus of older texts (texts from 1955 to 1989), the primary corpus (texts from 1955 to 2017, especially since 2000), and the corpus of online texts (until 2017). A comparison of the frequency and collocations of the analyzed words shows the dynamics of these microsystems in the language of the previous and the current period.

Keywords: synchrony of the language, lexical analysis, dynamics of abstract terms, frequency, Slovak National Corpus

1 SOURCES OF MATERIAL AND POSSIBILITIES OF RESEARCH OF SYNCHRONIC DYNAMICS OF LANGUAGE

After a theoretical distinction between language synchrony and diachrony was established and as soon as the synchronic dynamics of language were accepted as relevant and clarified from a terminological perspective (e.g. [1], [2], [3]), new possibilities opened up in linguistic research, enabling the analysis of dynamic phenomena, processes, and trends in virtually all linguistic areas of contemporary language. Numerous teams and individual authors made use of this situation. A collective monograph titled *The Dynamics of Contemporary Slovak Lexicon* [*Dynamika slovnej zásoby súčasnej slovenčiny*, 4] later became the most significant piece of academic work in this area. In addition to bringing numerous theoretical contributions and research innovations, its authors (Horecký, Buzássyová, Bosák et al.) also built on those methods and outputs of Slovak linguistics that had been focusing on empirical research. During the preparatory phase of the book's production, the authors created a separate card catalog with words and phrases including context. This card catalog consisted of manual excerpts of documents not captured in the dictionaries of Slovak that were available at the time. Although such an extensive preparation of materials was not common in the 1980s, it yielded

valuable results by providing a unique description of the dynamics of Slovak lexicon in the 1970s and the first half of the 1980s.

Another contribution that can be attributed to this book is its partial departure from a purely systemic approach to linguistic phenomena and its adoption of a communication-focused approach. The communication needs that influence the borrowing of new words and the openness of Slovak towards words from other languages, especially internationalisms of Graeco-Latin origin, are two ideas that also appear in other places. Klára Buzássyová, one of the co-authors of [4], mentions them in her own works (e.g. [5]) and a collective monograph titled *The Recent History of Slavic Languages. Slovak* (*Najnowsze dzieje języków słowiańskich*, [6, pp. 40–43]). Just like the former monograph, the latter is also based on a large number of specific excerpts collected from imaginative, professional, and journalistic texts. A large portion of these were manually excerpted by individuals, with written texts being the primary source and various types spoken communication a less frequent one. Research from this period usually quantified the occurrence of linguistic units using the words “frequent”, “less frequent”, “rare”, “productive”, and “unproductive” without precise quantification. Although these examples were collected and analyzed manually, the results of this research still prove to be valid in the present day, when they are verified in large corpora using modern corpus-based methods. Another feature of manual research is that these texts often include a large number of marginal, very infrequent phenomena, which are more likely to be recorded in manual excerption due to their various peculiarities.

In recent decades, there have been significant developments in the potential for using corpus-based textual and linguistic resources, i.e. working with linguistic materials by collecting written and spoken texts and making them available for research in large electronic databases. Corpus-based quantitative analysis of texts and methods of analyzing linguistic changes have been introduced in detail by researchers such as Václav Cvrček [7] and Michal Křen [8]. Cvrček points out that corpus linguistics has, among other things, made linguistic research place emphasis on empiricism and quantitative methods, while focusing more on parole [7, p. 11]. This statement can be specified using the ideas mentioned above: extensive and well-structured corpus resources have enabled the use of corpus-linguistic research methods to significantly expand the possibilities of empirical language research, especially the research of parole, thus meeting the often emphasized needs of many linguists (both Slovak and international) who had previously been aware of the limitations in their research that had caused by a lack of research materials. The current situation in regard to research material is different: “A well-known advantage of corpus linguistics, which might be taken for granted nowadays, is having a sufficient (sometimes even excessive) amount of empirical data and being able to easily quantify all quantifiable phenomena” [8, p. 11]. However, this advantage also has its pitfalls.

When comparing the potential for analyzing research materials before the existence of corpus resources and now, the present approach has numerous indisputable advantages. Corpora, corpus linguistics, and quantitative linguistics now play an irreplaceable role in the analysis of language, especially in the area of synchronic dynamics. Yet despite all the improvements and development it has undergone, corpus-based research still encounters limits and difficulties (cf. e.g. [9]). One problem stems from the difficulty of appropriately managing the amount of available corpus material (also from a time perspective). For instance, if a classical linguist initially tries to do corpus-based research, it often becomes corpus-driven because the linguistic material available in real-world texts starts providing new perspectives and provoking new questions. However, the large number of available documents may cause the analysis only to focus on a few frequently occurring phenomena, which is why less frequent (albeit from a linguistic perspective interesting) linguistic units usually do not receive any attention. It appears that the classical linguists who previously used manual excerption were able to capture these infrequent phenomena well.

This paper aims to combine a corpus-linguistic and quantitative approach with a more detailed manual analysis of the acquired linguistic material. The analysis should reveal dynamics and competing functioning of word-formation types with either an international or domestic formative element, i. e., words ending in *-ita/-ost'*. The paper concludes with a comparison of results from this partial analysis and the findings regarding the dynamics of words ending in *-izmus/-stvo* (cf. [10]). We presume that functioning of these words in contemporary written communication is different, although they represent similar groups of abstract words.

2 MATERIAL RESOURCES FOR THE ANALYSIS OF WORDS ENDING IN *-ita/-ost'*

This research is based on the resources provided by the Slovak National Corpus of the Ľudovít Štúr Institute of Linguistics (SNC) listed in Table 1.

<i>Corpus name</i>	<i>Corpus size</i>	<i>Corpus composition</i>	<i>Texts created in</i>
r1955az1989-5.0	83.6 million tokens; 66.8 million words	5.11% journalistic, 75.73% imaginative, 13.82% professional, 5.34% other texts	1955–1989
prim-8.0-public-sane	1.4 billion tokens; 1.1 billion words	73.75% journalistic, 16.33% imaginative, 8.91% professional, 1.01% other texts	1955–2017; mainly texts written after 2000
web-4.0	3 billion tokens; 2.4 billion words	–	?–2017

Tab. 1. The SNC's corpus resources used for analysis

The SNC's resources were searched using the NoSketch Engine (<https://www.sketchengine.co.uk>) and the [lemma="*.xxx"] CQL command, with the "xxx" string replaced with specific word endings, i.e. ita, ost'. The resulting concordances were then used to create frequency lists of lemmas. In addition to the desired abstract words with the analyzed formative elements, this formal search method also returned all other words with the same ending. The primary focus in each group was to compare the 20 most frequently occurring lemmas and, potentially, to compare the specific characteristics of additional expressions or the whole microsystem of the 1,000 most frequently occurring lemmas.

3 THEORETICAL BASIS FOR THE ANALYSIS OF WORDS ENDING IN *-ita/-ost'*

In the context of Slovak linguistics, the dynamic processes in abstract vocabulary have been systematically studied by Klára Buzássyová. She has focused on them in several partial studies: a detailed analysis of the competing word-formation types with the formative elements *-ita/-ost'* [11], which is also heavily utilized in this paper; an analysis of their negative forms [12]; and a summary of abstract vocabulary, especially the competing formative elements *-ita/-ost'*, which was presented in the collective monograph [4]. As stated by Buzássyová, "words with this structure have the same characteristics as abstract words in general: they are part of intellectual vocabulary, often formal or professional from a stylistic perspective" and they typically "a) form variants (linguistic units that are stylistically and semantically equivalent); b) find use as means of stylistic differentiation, i.e. as stylistic synonyms; c) form units that are partially differentiated, both semantically and from the perspective of communication areas (e.g. terminology, journalistic vocabulary, common vocabulary)" [11, pp. 142–143]. These findings will be verified by analyzing corpus materials with both older and newer written texts of traditional formats and materials from the web corpus, where a significant portion of texts combine features of written and spoken forms and they mainly use common vocabulary (cf. Table 1).

4 WORDS ENDING IN *-ita*

The list of words ending in *-ita*, which was composed using search of words forms, includes abstract nouns as well words denoting various entities (*univerzita* 'university', *lokalita* 'locality', *maturita* 'school-leaving examination'), proper nouns (*Judita*, *Margita*) or an acronym (*SITA*). Abstract vocabulary is not separately labeled in the corpus annotation, so automatic filters could only be used to remove proper nouns and non-feminine nouns (e.g. *bandita* 'bandit', masculine noun).

	<i>r1955az1989-5.0</i>	<i>prim-8.0-public-sane</i>	<i>web-4.0</i>
1.	kvalita 'quality'	aktivita 'activity'	kvalita 'quality'
2.	univerzita 'university'	kvalita 'quality'	aktivita 'activity'
3.	aktivita 'activity'	univerzita 'university'	univerzita 'university'
4.	realita 'reality'	realita 'reality'	lokalita 'locality'
5.	kapacita 'capacity'	kapacita 'capacity'	realita 'reality'
6.	autorita 'authority'	lokalita 'locality'	kapacita 'capacity'
7.	lokalita 'locality'	SITA	komunita 'community'
8.	Judita	priorita 'priority'	priorita 'priority'
9.	intenzita 'intensity'	komunita 'community'	stabilita 'stability'
10.	Margita	stabilita 'stability'	intenzita 'intensity'
11.	kontinuita 'continuity'	identita 'identity'	identita 'identity'
12.	bandita 'bandit'	autorita 'authority'	imunita 'immunity'
13.	nervozita 'nervosity'	elita 'elite'	autorita 'authority'
14.	Univerzita 'University'	charita 'charity'	špecialita 'speciality'
15.	relativita 'relativity'	solidarita 'solidarity'	elita 'elite'
16.	popularita 'popularity'	popularita 'popularity'	efektivita 'effectivity'
17.	maturita 'school-leaving examination'	intenzita 'intensity'	popularita 'popularity'
18.	individualita 'individuality'	komodita 'commodity'	kreativita 'creativity'
19.	formalita 'formality'	imunita 'immunity'	solidarita 'solidarity'
20.	produktivita 'productivity'	maturita 'school-leaving examination'	produktivita 'productivity'

Tab. 2. The 20 most frequently occurring words ending in *-ita* that appear in the SNC's corpora

A look at the list of the most frequently occurring words ending in *-ita* shows that the three top-ranking words are the same in all three corpora (albeit in different order). A total of 9 words appear in each of the three corpora, which indicates that their high frequency of use has been stable throughout different periods and styles. Two of the most frequently occurring words are abstract nouns that have form ending in *-ost'* (*aktivita*, *kvalita*) whose use is acceptable. Virtually all abstract nouns ending in the international formative element *-ita* can have alternative forms with the domestic formative element *-ost'*, although in many cases this possibility is merely hypothetical and the domestic forms are used less frequently in practice, if they are used at all. Buzássyová [11] points out that these systemically possible variants of derived words almost exclusively used in the international form were recorded by authors of dictionaries (back then, these were the authors of *Dictionary of the Slovak Language* 'Slovník slovenského jazyka', 1959–1968), even though the research materials (the card catalog) did not provide sufficient evidence for their support.

What follows is a list of words that Buzássyová includes in this group along with their actual occurrence in the general *prim-8.0-public-sane* corpus. Since this

comparison lists the occurrence of two words in the same corpus, the quantitative data is presented in the form of absolute frequency:

totalita	9,888	totalitnosť	6	‘totality’
nervozita	17,799	nervóznosť	6	‘nervosity’
brutalita	3,179	brutálnosť	155	‘brutality’
invalidita	2,453	invalidnosť	0	‘invalidity’
sexualita	7,167	sexuálnosť	96	‘sexuality’
periodicita	1,979	periodickosť	18	‘periodicity’

The data indicates that even in contemporary texts these hybridisms formed by derivation (foreign root + domestic formative element) only appear rarely, if they appear at all. Buzássyová confirms this idea by saying that the “forms with the formative element *-ita* are virtually the only versions that appear, they are neutral from a stylistic perspective, and their formal/professional character is not perceived [by users]” [11, p. 144]. However, a similar situation arises when studying pairs of words with international bases and international/domestic formative elements. Buzássyová [Ibid., p. 145] labels these “expressive variants, i.e. linguistic units that are neutral and equivalent from a semantic and stylistic perspective”. In this group of words in the prim-8.0-public-sane corpus, those with an international formative element are strongly predominant. However, there are two pairs (*absurdita* – *absurdnosť* ‘absurdity’, *objektivita* – *objektívnosť* ‘objectivity’) whose occurrences are not so diametrically different:

agresivita	9,926	agresívnosť	329	‘aggressivity’
absurdita	3,477	absurdnosť	2,178	‘absurdity’
aktivita	216,776	aktívnosť	520	‘activity’
genialita	1,925	geniálnosť	152	‘geniality’
kolegialita	619	kolegiálnosť	89	‘collegiality’
objektivita	3,491	objektívnosť	2,493	‘objectivity’
popularita	21,561	populárnosť	170	‘popularity’
stabilita	36,704	stabilnosť	127	‘stability’
nestabilita	5,046	nestabilnosť	117	‘instability’

From the perspective of their occurrence in contemporary texts, next group of words can be labeled as the most heterogeneous one. According to Buzássyová, this group is the “core of correlation between words belonging to the *-ita/-osť* word-formation types” and also the “most dynamic component of this lexico-semantic and word-formation microsystem, which most clearly documents the dynamics in this part of the vocabulary” [Ibid., p. 146]. These are pairs where each word has differentiated itself as a stylistic synonym. Buzássyová states that forms with the formative element *-ita* have a (more) formal character, while forms ending in the domestic suffix *-osť* are neutral and less formal. Although the frequency of these

words is usually low for both words in the pair (perhaps equally formal in situations like these), this group also includes pairs where the versions ending in the domestic suffix *-ost'* prevail, which would indicate that they are more neutral and common, e.g. *slovenskost'*, *plastickost'*, *poetickost'*:

uniformita	941	uniformnost'	43	'uniformity'
autenticita	2,765	autentickost'	2,502	'authenticity'
labilita	514	labilnost'	102	'lability'
atraktivita	4,614	atraktivnost'	2,543	'attractiveness'
akceptabilita	1	akceptabilnost'	0	'acceptability'
		akceptovatel'nost'	314	'acceptableness'
aplikabilita	4	aplikabilnost'	1	'applicability'
		aplikovatel'nost'	199	'applicableness'
direktivita	3	direktivnost'	33	'authoritativeness'
historicita	223	historickost'	321	'historicity'
slovacita	41	slovenskost'	770	'Slovakness'
plasticita	272	plastickost'	328	'plasticity'
poeticita	0	poetickost'	360	'poeticness'
potencialita	117	potencialnost'	24	'potentiality'
sugestivita	64	sugestivnost'	321	'suggestiveness'
simultaneita	5	simultannost'	150	'simultaneousness'

Words ending in the analyzed formative elements *-ita/-ost'* (ale aj *-izmus/-stvo*, cf. [10]) have a strong tendency towards prefixation and compounding (cf. [5, pp. 120–122]). The group of words ending in *-ita* have shown to have 49 different prefixes and prefixoids among the 1,000 most frequent lemmas in the prim-8.0-public-sane corpus, including compound words with the same format as *seropozitivita* 'seropositivity'. These prefixes and prefixoids include: *a-, ab-, agro-, auto-, bi-, bio-, cyto-, de-, dis-, elektro-, ex-, extra-, foto-, gastro-, hepato-, hetero-, homo-, hyper-, hypo-, i-, ichno-, in-, inter-, kardio-, ko-, kon-, kyber-, meteo-, multi-, ne-, nefro-, neuro-, non-, perme-, poly-, post-, pro-, pseudo-, radio-, re-, retro-, sero-, sub-, super-, stvar-, termo-, trans-, tri-, uni-* (e.g. *diskontinuita* 'discontinuity', *hyperaktivita* 'hyperactivity', *iracionalita* 'irrationality', *multikulturalita* 'multiculturality', *radioaktivita* 'radioactivity', *superkvalita* 'superquality').

The most frequently occurring words are hybrid compound words with the domestic prefix *ne-* (16 derived forms; e.g. *nestabilita* 'instability', *nekvalita* 'poor quality'), which are interpreted as negative forms. Most of the other prefixes are international: *inter-* (9 derived forms), *hyper-* (6), *multi-* (4), *a-, bi-, dis-, i-, in-* (3 derived forms each). The most used word bases are *-aktivita* (9 derived forms), *-sexualita* (7), *-toxicita* (4).

5 WORDS ENDING IN *-ost'*

The list of words ending in *-ost'* also includes forms that are not abstract names of qualities, which could be automatically filtered out by removing non-feminines and non-nouns (*dost'* 'quite' – adverb, *host'* 'guest' – masculine). Not a single word presented in Table 3 allows the creation of a form ending in *-ita* and the entire analyzed group of 1,000 most frequent lemmas ending in *-ost'* includes very few words that do (e.g. *efektívnosť* – *efektívita* 'effectivity', *aktuálnosť* – *aktualita* 'actuality', *atraktívnosť* – *atraktívita* 'attractivity').

	<i>r1955az1989-5.0</i>	<i>prim-8.0-public-sane</i>	<i>web-4.0</i>
1.	<i>dost'</i> 'enough'	<i>spoločnosť</i> 'company'	<i>spoločnosť</i> 'company'
2.	<i>spoločnosť</i> 'company'	<i>možnosť</i> 'possibility'	<i>možnosť</i> 'possibility'
3.	<i>činnosť</i> 'activity'	<i>host'</i> 'guest'	<i>činnosť</i> 'activity'
4.	<i>skutočnosť</i> 'reality'	<i>činnosť</i> 'activity'	<i>skúsenosť</i> 'experience'
5.	<i>možnosť</i> 'possibility'	<i>dost'</i> 'enough'	<i>dost'</i> 'enough'
6.	<i>radosť</i> 'joy'	<i>súčasnosť</i> 'the present'	<i>súčasnosť</i> 'the present'
7.	<i>pozornosť</i> 'attention'	<i>príležitosť</i> 'opportunity'	<i>skutočnosť</i> 'reality'
8.	<i>udalosť</i> 'event'	<i>skutočnosť</i> 'reality'	<i>príležitosť</i> 'opportunity'
9.	<i>miestnosť</i> 'room'	<i>skúsenosť</i> 'experience'	<i>schopnosť</i> 'ability'
10.	<i>host'</i> 'guest'	<i>minulosť</i> 'the past'	<i>povinnosť</i> 'duty'
11.	<i>príležitosť</i> 'opportunity'	<i>budúcnosť</i> 'the future'	<i>veľkosť</i> 'size'
12.	<i>skúsenosť</i> 'experience'	<i>radosť</i> 'joy'	<i>host'</i> 'guest'
13.	<i>minulosť</i> 'the past'	<i>súvislosť</i> 'connection'	<i>starostlivosť</i> 'care'
14.	<i>budúcnosť</i> 'the future'	<i>verejnosť</i> 'the public'	<i>budúcnosť</i> 'the future'
15.	<i>súvislosť</i> 'connection'	<i>povinnosť</i> 'duty'	<i>radosť</i> 'joy'
16.	<i>starosť</i> 'worry'	<i>udalosť</i> 'event'	<i>minulosť</i> 'the past'
17.	<i>povinnosť</i> 'duty'	<i>pozornosť</i> 'attention'	<i>verejnosť</i> 'the public'
18.	<i>vlastnosť</i> 'characteristic'	<i>žiadosť</i> 'request'	<i>udalosť</i> 'event'
19.	<i>schopnosť</i> 'ability'	<i>schopnosť</i> 'ability'	<i>vlastnosť</i> 'characteristic'
20.	<i>prítomnosť</i> 'presence'	<i>osobnosť</i> 'personality'	<i>miestnosť</i> 'room'

Tab. 3. The 20 most frequently occurring words ending in *-ost'* that appear in the SNC's corpora

The cumulative occurrences for the entire group of words ending in *-ost'* in the *prim-8.0-public-sane* corpus are 5,5 times bigger than for the group of words ending in *-ita*, even though the second one relatively frequently includes proper names: **ost'* 10,580,133 – **ita* 1,870,719. This context makes the results from the previous section, which showed that in cases of competing pairs, internationalisms with the formative element *-ita* were predominant, even more remarkable. In comparison with the most frequent words ending in *-ita* (tens of thousands of occurrences), there are generally more words ending in *-ost'* and the frequency of the most frequent individual words is much higher (hundreds of thousands of occurrences).

Compared to the other groups of abstract names that have been analyzed so far (words ending in *-ita*, *-izmus*, *-stvo*), words ending in *-ost'* have a specific tendency towards derivational prefixation and compounding. The prefixes include 10 formative elements (*bez-*, *in-*, *nad-*, *ne-*, *pre-*, *pred-*, *proti-*, *roz-*, *z-*, *za-*), which are almost exclusively domestic (with the exception of *in-*) and with a strong prevalence of the prefix *ne-* (129 derived forms). However, a significant number of words formed using prefixes have been lexicalized and their origin via prefixation is no longer clearly felt (e.g. *zaslepenosť* 'blindness', *prejazdnosť* 'traffic flow', *predvídateľnosť* 'predictability', *rozpracovanosť* 'unfinished character').

Compound words make up a much larger part of this group. They have 53 different initial elements whose semantic content usually refers to measure and quality (e.g. *jedno-* 'one-', *vše-* 'all-', *každo-* 'every-', *dobro-* 'well-', *lahko-* 'easy-'). The most common final elements of compound words are *-schopnosť* (7 compound words), *-hodnosť* (5), and *-myselnosť* (5), e.g.: *cielavedomosť* 'single-mindedness', *lahkomyselnosť* 'carelessness', *mnohotvárnosť* 'multiformity', *samolúbosť* 'self-satisfaction'.

6 CONCLUSION

The synchronic dynamics of language, the topic of internationalization as a general trend, and the relationship between domestic and international words have been studied by several authors. In Slovakia, most of this work was done in the 1980s. This analysis of words ending in *-ita/-ost'* has primarily focused on two aspects of these topics that have not been exhaustively studied: a study of the extensive materials currently offered by corpus databases and especially an analysis of the way abstract vocabulary functions within the whole lexicon and within the microsystems of words ending in *-ita/-ost'*. Some of the previous observations and partial conclusions have been confirmed or specified using frequency data, while others will require further detailed analysis.

The considerable tendency of word formation using prefixes, semiprefixes, and compounding (even by combining domestic and international elements) that has been identified in the group of words ending in *-izmus/-stvo* (cf. [10]) can also be seen in the group of words ending in *-ita*. Here, prefixation typically uses foreign prefixes and semiprefixes. The group of words ending in *-ost'* includes an unusually rich and diverse collection of compound words. Unlike the group of words ending in *-izmus* and partially the group of words ending in *-stvo*, the group of words formed using *-ita/-ost'* does not include any productive type of word formation using first or last names. Furthermore, no active type of word formation using pejoratives and ironic words has been identified in this group. From the perspective of frequency, overall the group of words ending in *-ost'* is much more utilized than the group of words ending in *-ita*, but an analysis of competing forms with international and

domestic formative elements shows that the particular words ending in *-ita* is more utilized than the same word bases with suffix *-ost'*, from the perspective of both function and frequency. This analysis of contemporary texts indicates that the competition between international and domestic formative elements is often very strong and the processes of intellectualization and internationalization in contemporary Slovak do not occur the same way in all lexical microsystems. On the contrary, language users select an appropriate form based on their own communication needs.

References

- [1] Barnet, V. (1981). Synchronní dynamika spisovného jazyka. *Jazykovedný časopis*, 32(2), pages 123–130.
- [2] Jedlička, A. (1981). Vývojové procesy a synchronní dynamika jazyka v konfrontačním osvětlení. *Jazykovedný časopis*, 32(2), pages 107–116.
- [3] Horecký, J. (1988). Dynamickosť a dynamika v jazyku. In *Studia Academica Slovaca*. 17. pages 189–199, Bratislava, Alfa.
- [4] Horecký, J., Buzássyová, K., Bosák, J. et al. (1989). *Dynamika slovnej zásoby súčasnej slovenčiny*. Bratislava, Veda.
- [5] Buzássyová, K. (2010). Vzťah internacionálnych a domácich slov v premenách času. *Jazykovedný časopis*, 61(2), pages 113–130.
- [6] Najnowsze dzieje języków słowiańskich. *Slovenský jazyk*. (1998). Opole, Uniwersytet Opolski – Instytut Filologii Polskiej.
- [7] Cvrček, V. (2013). Kvantitatívni analýza kontextu. Praha, Ústav Českého národního korpusu – Nakladatelství Lidové noviny.
- [8] Křen, M. (2013). Odrasť jazykových zmien v synchronných korpusech. Praha, Ústav Českého národního korpusu – Nakladatelství Lidové noviny.
- [9] Šimková, M., Gajdošová, K., Kmeťová, B., and Debnár, M. (2017). *Slovenský národný korpus. Texty, anotácie, vyhľadávania*. Bratislava, Jazykovedný ústav Ľ. Štúra SAV – Vydavateľstvo Mikula.
- [10] Šimková, M. (2018). Synchronná dynamika slov zakončených na *-izmus/-stvo* v textoch Slovenského národného korpusu. *Jazykovedný časopis*, 69(3), pages 560–571.
- [11] Buzássyová, K. (1986a). Konkurencia slovotvorných typov s formantmi *-ita* a *-ost'*. *Slovenská reč*, 51(3), pages 142–152.
- [12] Buzássyová, K. (1986b). Novšie názvy negatívnych vlastností a stavov. *Kultúra slova*, 20(10), pages 335–340.
- [13] Slovenský národný korpus. Verzia prim-8.0-public-sane. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2018. Accessible at: <https://korpus.juls.savba.sk>.
- [14] Slovenský národný korpus. Korpus r1955az1989-5.0. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2018. Accessible at: <https://korpus.juls.savba.sk>.
- [15] Slovenský národný korpus. Korpus web-4.0. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2018. Accessible at: <https://korpus.juls.savba.sk>.
- [16] <https://www.sketchengine.co.uk>.

SLOVAK COMPARATIVE CORRELATIVES: NEW INSIGHTS

JAKOB HORSCH

Catholic University of Eichstätt-Ingolstadt, Germany

HORSCH, Jakob: Slovak comparative correlatives: New insights. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 180 – 190.

Abstract: Comparative Correlatives (CCs) are structures that have attracted substantial interest. In Slovak, they typically look like the following proverb:

Čím bližšie Rím, tým horší kresťan.

‘The closer (to) Rome, the worse the Christian.’

So far, no extensive research has been conducted on CCs in Slavic languages except Polish [1]. In Slovak, CCs have not received a great deal of attention. Accordingly, this study examines the various forms of CCs in a Slovak National Corpus (SNC) random sample of 500 tokens, showing that there is much more variety than has been acknowledged in the literature. Frequencies will be used to show that there are iconic structures, and it will be argued that there are construction-specific properties that suggest the existence of a specific CC construction in Slovak.

Keywords: Slovak, Comparative Correlative, Slovak National Corpus

1 INTRODUCTION

The comparative correlative (CC), also known as *comparative conditional* [2], *proportional correlative* [3] and *the... the... construction* [4], is a highly interesting structure that has seen increased attention [1], [3], [5–9]. Most of this research has focused on English, neglecting the Slavic languages (an exception is Borsley’s study of Polish CCs [1]), which differ significantly from English with regard to basic parameters such as word order.

In its most simple form, the CC construction consists of two clauses, C1 and C2. It appears that this basic form is comparable across many languages [10], as the following examples illustrate for English (1) [11], Slovak (2), Polish (3) [1] and German (4) [5]:

- (1) [*The more carefully you do your work,*]_{C1} [*the easier it will get.*]_{C2}
- (2) [*Čím menej rečí tu bude,*]_{C1} [*tým skôr zaspím.*]_{C2}
‘The less talking there is here, the sooner I will fall asleep.’
<SNC prim-7.0-public-all JTol2>
- (3) [*Im bardziej zmęczony jesteś,*]_{C1} [*tym gorzej pracujesz.*]_{C2}
‘The more tired you are, the worse you work.’

- (4) [*Je müder Otto ist,*]_{C1} [*desto aggressiver ist er.*]_{C2}
‘The more tired Otto is, the more aggressive he is.’

Semantically, CCs are complex: C2 can be described as the effect (or apodosis/dependent variable) of C1 (the corresponding protasis/independent variable) [12], [13]. More precisely, the semantic properties are both asymmetric as well as symmetric: On the one hand, there is a conditional, or asymmetric, relationship, i.e., in example (1) getting together results in happiness, and on the other hand, there is parallel change over the same time period, i.e., by getting together more and more, happiness simultaneously increases. Sag refers in this context to a “pair of semantic differentials” coupled with a “monotonic relationship” [9].

Concerning the form, each clause is introduced by fixed, i.e., invariable clause-initial elements; in the case of Slovak *čím* (C1) and *tým* (C2). We can also see that these clause-initial elements are followed by comparative elements such as *more carefully* and *easier* in English (1), and *menej reči* and *skôr* in Slovak (2). Finally, there is the option of inserting a clause after these comparative elements, such as *you do your work* in C1 in (1), or *zaspím* in C2 in (2).

In other words, this is what Hoffmann refers to as a “constructional template” [7], [14] that produces CCs that vary in complexity: While the clause-initial elements are fixed, there is a slot for comparative elements that can be freely filled and a further slot for clauses that can but doesn’t have to be filled.

In various languages, idiosyncrasies have been observed in CCs, which has led to increased interest: Borsley, for example refers to the CC as a “notable peripheral construction” that exhibits phenomena that “fall outside the scope of syntax proper” [1]. In fact, Slovak CCs also exhibit a variety of highly interesting idiosyncrasies.

This is why the following study was conducted with evidence from the Slovak National Corpus (SNC). It examines the various forms of Slovak CCs, discussing particularly interesting traits. The aim is to complement the literature, which has so far treated the CC in Slovak marginally: Many of the forms found in the SNC are not mentioned at all in the literature. Authentic examples from a 500 token random sample will showcase the manifold forms CCs can appear in, and by use of frequencies suggest their structures that appear to be clearly preferred over others.

2 ONLINE DICTIONARIES IN GDC

2.1 Slovak CCs in the literature

So far, no extensive research has been carried out on Slovak CCs. However, they have sporadically attracted interest. The earliest mentioning of the co-occurrence of the clause-initial elements *čím* and *tým* can be traced to 1943 [15].

In subsequent years, *čím-tým* as used in CCs was discussed briefly by Betáková and Marsinová. Interestingly, according to Betáková, *čím* and *tým* belong to the category of “correlative conjunctions” (*súvzťažná spojka*) [16]. Marsinová, on the contrary, suggests the pair be excluded “a priori” from the category of conjunctions because she considers the word *čím* not to be related to the conjunction *čo* [17].

In fact, we can observe general uncertainty concerning the classification of *čím* and *tým* as used in CCs: Elsewhere they are classified as “hypotactic conjunctions” (*hypotaktická spojka*) [18], and other sources state that while they look like the instrumental case forms (*siedmy pád*) of relative pronouns (*súvzťažné zámená*) *čo* and *to*, they have become fossilized as a “pair of conjunctions” (*spojková dvojica*) which connects “modal clauses of comparison” (*spôsobové vety porovnávacie*) [19].

In a similar vein, the only longer study on Slovak CCs explicitly excludes *čím* and *tým* as they appear in CCs from the group of pronouns and speaks of “particles” (*častice*) that “modify comparatives” [20], furthermore noting that when *tým* is eliminated from the structure, *čím* loses its validity, thereby implying that both of these words are necessary for the structure to carry the distinct CC meaning.

This view also points to the interpretation that will be argued for later, which is that we are looking at fixed clause-initial elements that may be etymologically related to instrumental-case forms of the pronouns *čo* and *to*, but are in fact construction-specific elements.

Further uncertainty in the literature concerns terminology: It appears that there is no agreement on what to call the CC construction in Slovak. Various terms are used, including “modal or comparative adverbial clauses” (*spôsobové alebo prirovnávacie vety*) [21], “adverbial comparative clause” (*príslovková veta prirovnávacia*) [22], “comparative clauses” (*porovnávacie vety*) [23], “comparative modal adverbial clause” (*príslovková veta spôsobová – prirovnávacia*) [24], [25], and “adverbial subordinate clause of degree” (*príslovková vedľajšia veta miery*) [26].

This disagreement hints at the marginal status of CCs in research, which is also reflected by the little attention they receive in grammars: While some do briefly mention CCs [21–24], other grammars completely ignore their existence [27], [28]. Even the *Morfológia slovenského jazyka* devotes no more than one paragraph to the CC [18].

Of course, this makes Slovak CCs all the more interesting. It is also noteworthy that the examples given in the sources above do not suggest a great variety of possible forms. As will be shown, CCs in Slovak actually appear in many forms, suggesting that CCs are a highly productive structure.

Moreover, Slovak CCs possess unique construction-specific properties such as the invariable clause-initial elements discussed above, and obligatory and optional slots that can accommodate material of varying complexity. Together with the lack of research noted earlier, these features certainly warrant an in-depth corpus study.

2.2 Corpus Study

The following study is based on the *prim-7.0-public-all* version of the SNC, a corpus of written Slovak that consists of 65.1% journalistic, 15.1% fiction, 9.5% professional and 10.3% other texts, with a size of almost 1 billion words [29]. The following CQL query¹ using the SNC web interface [30] was used to find *čím-tým* patterns:

```
[word="čím"] []{,9} [tag="(D*(x|y|z)|A*(x|y|z)|G*(x|y|z)).*" ] []* [word="," ]
[word="tým"] []{,9} [tag="(D*(x|y|z)|A*(x|y|z)|G*(x|y|z)).*" ]
```

In total, this query yielded 10,151 tokens, from which a random sample of 500 tokens was extracted. From these, 17 false positives were determined, leaving 483 relevant CC tokens.

The first notable characteristic of Slovak CCs that the data shows is their variation in complexity, which examples (5) to (8) from the SNC illustrate:

- (5) *Čím ďalej, tým lepšie.*
 ČÍM further TÝM better
 ‘The further/longer, the better.’
 <SNC prim-7.0-public-all InZ2/03>
- (6) *Čím viac k športu, tým viac od nebezpečných ciest (...).*
 ČÍM more to sports TÝM more from dangerous
 ways
 ‘The more inclination towards sports, the less inclination towards dangerous lifestyles.’
 <SNC prim-7.0-public-all KOR2001/06>
- (7) *Čím viac milujeme, tým viac rastieme v slobode.*
 ČÍM more love:we TÝM more grow:we in freedom
 ‘The more we love, the more we grow in freedom.’
 <SNC prim-7.0-public-all MI2010/05>
- (8) *Čím dôkladnejšie popremýšľate o svojich krokoch, tým lepšie to pre vás dopadne.*
 ČÍM thoroughly-more think:you:2:PL about your steps
 TÝM better it for you results
 ‘The more thoroughly you think about your steps, the better it turns out for you.’
 <SNC prim-7.0-public-all MYBB2013/36>

¹ I would like to thank my colleague Dr. Thomas Brunner from the Department of English Linguistics at the Catholic University of Eichstätt-Ingolstadt for helping me compose this regular expression.

As is evident, the complexity of Slovak CCs ranges from very basic constructions with only comparative elements as in (5) (Sabol refers to these as “elliptic” [20], implying the omission of a verb) to complex structures such as (8) that include transitive verbs and prepositional objects. It is interesting that in this context, the *Morfológia* mentions that following the clause-initial elements, there can be a “word or a clause” (*slovo alebo veta*) [18], but does not provide examples of complex structures such as (8).

We can thus say that apart from the invariable clause-initial elements *čím* and *tým*, the C1 and C2 clauses have slots: First, one that contains an obligatory comparative element (e.g. *ďalej* and *lepšie* in (5)) following the clause-initial elements *čím* and *tým* and second, an optional clause slot that follows the comparative elements, as is demonstrated by (6) to (8). These clauses vary considerably in length and complexity.

Generalizing from these observations, we can thus determine a more abstract schema, or “constructional template” [7, 14] for Slovak CCs (9), based on Culicover and Jackendoff’s template for English CCs [10]. Note that the clause-initial elements² are transcribed in IPA to represent their phonological invariability, as they are assumed to be construction-specific and not related to the pronouns *čo* and *to*.

(9) [[tʃi:m] [...]comp. element [...]opt. clause]C1 [[ti:m] [...] [...]comp. element [...]opt. clause]C2

Turning to the comparative element, CCs can contain adjectives such as *horší* in the Slovak proverb *Čím bližšie Rím, tým horší kresťan*, adverbs (e.g. *viac* in (7)), or noun phrases, as in (10):

(10) *čím väčšie ťažkosti treba prekonať, tým prenikavejšie*
 ČÍM bigger problems must:he overcome TÝM brighter
zažiari úspech učiteľa
 shines success teacher:GEN

‘The bigger the problems that must be overcome, the brighter the success of the teacher shines.’

<SNC prim-7.0-public-all BGal1>

In this context, one variable that was coded for the SNC data was FILLER TYPE, which revealed that there is a strong preference for adverb phrases as comparative elements, as Tables and Figures 1 and 2 show:

² The *Morfológia* notes that in informal language, *to* may be used instead of *tým*. [18]

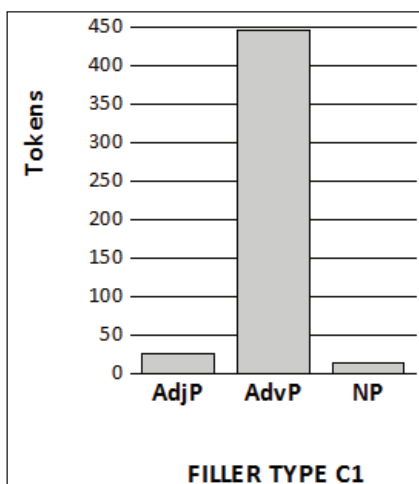


Fig. 1. C1 filler types

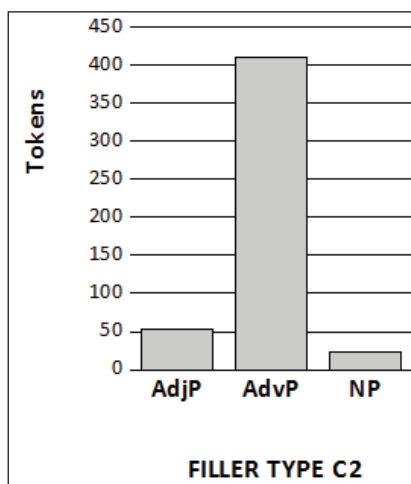


Fig. 2. C2 filler types

FILLER TYPE C1	Tokens	FILLER TYPE	Tokens
ADJP	23	ADJP	51
ADVP	446	ADVP	408
NP	14	NP	24
Total	483	Total	483

Tab. 1. C1 filler types

Tab. 2. C2 filler types

The numbers show that both in C1 and C2, the majority of comparative elements were adverb phrases, 446 and 408 out of 483, respectively. Nevertheless, there was still a significant number of other filler types, which demonstrates the productivity of the pattern.

Moving on, an interesting feature of Slovak CCs that differentiates them from their English counterparts is the relatively free ordering of constituents. Consider (11) and (12), for example:

- (11) *čím som sa viac usiloval, tým*
 ČÍM am:I sa:REFL more tried TÝM
silnejšie sa vo mene táto odporná
 stronger sa:REFL in me this detestable
vlastnosť presadzovala.
 feature asserted
 ‘The more I tried, the stronger this detestable feature asserted itself in me.’ <SNC: prim-7.0-public-all SME2009/10>

- (12) *Čím bol spisovateľ odvážnejší, tým väčší*
 ČÍM was:he author brave:more TÝM bigger
*účinko malo jeho dielo.*³
 effect had:it his work
 ‘The braver an author was, the bigger an effect his work had.’

We see that comparative elements do not have to follow the clause-initial elements but can also be found at the center, as in (11) or at the end, as in (12). This means that the template as suggested in (9) is not entirely satisfactory.

As Tables and Figures 3 and 4 illustrate, however, there is a clear tendency towards placing the comparative element in the front position, right after the clause-initial element (clauses which consist of a comparative element only, as in (5), were omitted from the coding of this variable):

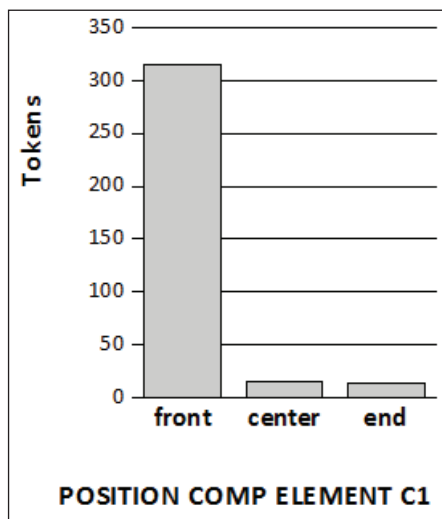


Fig 3. Comparative element position in C1

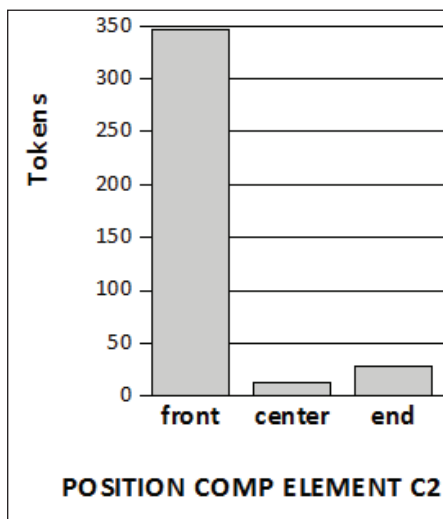


Fig. 4. Comparative element position in C2

COMPARATIVE EL. POSITION C1	Tokens	COMPARATIVE EL. POSITION C2	Tokens
front	315	front	346
center	15	center	12
end	13	end	28
Total	343	Total	386

Tab. 3. Comparative element position in C1

Tab. 4: Comparative element position in C2

³ www.litcentrum.sk/31662

These numbers show that while the generally free word order of Slovak does apply to the CC construction, there is a clear preference for an order with the comparative element immediately following the clause-initial element, as in (5) to (8). It is noteworthy that this is also the only order discussed in the literature; there is no mentioning of any of the alternatives as in (11) and (12).

Another interesting phenomenon is that of so-called “stacked” constructions, which are also known in English [8], where a CC consists of more than two clauses, as in (13) and (14):

- (13) [*A* *čím* *viac* *rastie* *cena* *komodít* *a*
 And ČÍM more grows price commodities:GEN and
energie,]_{C1}
 energy:GEN
 [*tým* *viac* *je* *americká ekonomika* “*spanikárená*”],]_{C2}
 TÝM more is American economy “panicked”
 [tým viac sa prepadá do recesie.]]_{C3}
 TÝM more sa:REFL sinks:it into recession
 ‘And the more the price of commodities and energy grows, the more the American economy panics, the more it sinks into recession.’
 <SNC: prim-7.0-public-all HN2008/04>
- (14) [*čím* *nižšia* *je* *akontácia*],]_{C1}
 ČÍM lower is deposit
 [*a* *čím* *dlhšie* *trvá* *lizing*],]_{C1'}
 and ČÍM longer lasts leasing
 [tým *drahšie* *kúpa* *vyjde*.]]_{C2}
 TÝM expensive:more purchase becomes
 ‘The lower the deposit is and the longer the leasing lasts, the more expensive the purchase becomes.’
 <SNC: prim-7.0-public-all SME04/02>

As these examples show, there are at least two variations of stacked clauses in Slovak: First, there are what we will call C1C2C3 clauses, as in (13), indicated by the *čím-tým-tým* clause-initial elements. The semantics of this CC are as follows: C1 is the cause for the effect in C2, which in turn is the cause for the effect in C3.

Second, there are cases of two causes (C1 and C1’) resulting in the same effect (C2), as suggested by the *čím-čím-tým* clause-initial elements in (14). To paraphrase this CC, the purchase becomes more expensive due to both lower deposits and a longer leasing duration. This is why the designation C1C1’C2 is suggested.

Note that neither of these two stacked arrangements in CCs is mentioned in the literature. The reason might be that the SNC data suggests these are not iconic: Out of the 483 CC tokens, only 40, or just over 8% were such structures with a clear majority of 443 iconic C1C2s.

A further noteworthy feature is reverse, i.e., C2C1, order. Two such examples were found in the SNC sample as false positives. This phenomenon is known from Polish as well, as Borsley’s variation (15) of example (3) [1] demonstrates. In Slovak, C2C1s appear as in example (16).

- (15) [*Tým gorzej pracuješ,*]_{C2} [*im bardziej jesteś zmęczony.*]_{C1}
 ‘The more tired you are, the worse you work.’
- (16) [*filozofi boli tým lepší,*]_{C2} [*čím boli*
 philosophers were TÝM better ČÍM were:they
starší.]_{C1}
 older
 ‘Philosophers were the better the older they were.’
 <SNC: prim-7.0-public-all AGI1>

It is notable that such a C2C1 order is generally not discussed in the literature. While some sources do provide examples [25], these are never discussed with regard to their semantics. Only one source comments on the possible non- iconicity of C2C1s, noting the C1C2 order is more or less “consistent” (*ustálené*) as opposed to the C2C1 arrangement, which is called an “extraordinary occurrence” (*výnimočný jav*) [20], thereby implying an iconic C1C2 structure. Whether such an iconic structure exists in Slovak is a question that must be answered in a future corpus study with a dataset obtained from a regular search expression that includes C2C1s.

3 CONCLUSION

The phenomena discussed in this paper reveal the great variety of forms in which Slovak CCs can appear, far more than the examples provided in the literature to date suggest. Generalizing from the many forms, we can derive a template as suggested in (9), consisting of three slots that follow the words *čím* and *tým*.

These words are referred to as “clause-initial elements” here because, as suggested by their phonetic transcription, *čím* and *tým* are neither conjunctions of any sort, nor instrumental case pronouns, but rather construction-specific elements that are unique to the Slovak CC: If either is removed, the construction as a whole loses its CC meaning (cf. also [20]).

The template (9) suggested for Slovak CCs is highly productive, leading to the creation of structures that range from very simple, such as (5), to highly complex, such as (8), or even stacked C1C2C3/C1C1’C2 structures, as in (13) and (14). Furthermore, CCs can even appear in inverse C2C1 order as in (16).

The frequencies determined in the SNC random sample suggest that despite the possibility of variation in Slovak CCs, there are arrangements that are clearly preferred over others: There appears to be a clear preference for adverb phrases in

the comparative element slot, for placing this element at the front of the clause, and for “iconic” C1C2 structures as opposed to more complex, stacked ones. These findings confirm the existence of a template as in (9).

The present study has thus managed to shed light on a construction in Slovak that has so far received little attention, despite its many interesting traits. We have been able to show that the Slovak CC is highly productive pattern which produces structures that by far exceed the possibilities that have so far been discussed in the literature, and that it can be regarded as a construction in its own right, with construction-specific properties such as invariable clause-initial elements.

The corpus data provided many more interesting examples, such as interrogative CCs and split fillers that had to be excluded from this study in the interest of brevity. Together with C2C1s, they could form the base for further research on a highly interesting construction in Slovak.

References

- [1] Borsley, R. D. (2004). On the Periphery: Comparative Correlatives in Polish and English. *Proceedings of Formal Approaches to Slavic Linguistics*, 12, pages 59–90.
- [2] McCawley, J. D. (1988). The Comparative Conditional Construction in English, German, and Chinese. In *General Session and Parasession on Grammaticalization*. Berkeley Linguistics Society, pages 176–187.
- [3] Den Dikken, M. (2005). Comparative Correlatives Comparatively. *Linguistic Inquiry*, 36(4), pages 497–532.
- [4] Cappelle, B. (2011). The the... the... construction: Meaning and readings. *Journal of Pragmatics*, 43 (1), pages 99–117.
- [5] Beck, S. (1997). On the Semantics of Comparative Conditionals. *Linguistics and Philosophy*, 20 (3), pages 229–271.
- [6] Borsley, R. D. (2004). An Approach to English Comparative Correlatives. In *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar*, Center for Computational Linguistics, Katholieke Universiteit Leuven. Ed. S. Müller, pages 70–92. Stanford, CA: CSLI Publications.
- [7] Hoffmann, Th. (2014). Comparing English Comparative Correlatives. Post-doc thesis.
- [8] Hoffmann, Th. et al. (2019). The More Data, The Better: A Usage-based Account of the English Comparative Correlative Construction. *Cognitive Linguistics*, 30(1).
- [9] Sag, I.A. (2010). English Filler-Gap Constructions. *Language: Journal of the Linguistic Society of America*, 86(3), pages 486–545.
- [10] Culicover, P. W., and Jackendoff, R. (1999). The View from the Periphery: The English Comparative Correlative. *Linguistic Inquiry*, 30(4), pages 543–571.
- [11] Fillmore, C. J. et al. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pages 501–538.
- [12] Goldberg, A. (2005) *Constructions at Work*. Oxford University Press.
- [13] Hoffmann, Th. (2017). Construction Grammar as Cognitive Structuralism: The interaction of constructional networks and processing in the diachronic evolution of English comparative correlatives. *English Language and Linguistics*, 21(2), pages 349–373.

- [14] Hoffmann, Th. (2019). *English Comparative Correlatives: Diachronic and Synchronic Variation at the Lexicon-Syntax Interface*. Cambridge, Cambridge University Press.
- [15] Vavro, J. (1943). Zo Syntaxe Záměna Čo. *Slovenská reč*, 10(7–8), pages 222–224.
- [16] Betáková, V. (1955). Poznámky k Učebnici jazyka slovenského pre štvorročné odborné školy. *Slovenská reč*, 20(5), pages 313–323.
- [17] Marsinová, M. (1955). Spracovanie gramatických kategórií v normatívnom slovníku. *Slovenská reč*, 20 (1), pages 29–39.
- [18] Dvonč, L. et al. (1966). *Morfológia slovenského jazyka*. Bratislava, Slovenská akadémia vied.
- [19] Oravec, J. (1954). Používanie slova ‘čo’ v spisovnej slovenčine. *Jazykovedný časopis*, 8, pages 216–233.
- [20] Sabol, F. (1982). Slovo čím v platnosti záměna, Častice a spojky. *Slovenská reč*, 47(1), pages 51–54.
- [21] Pauliny, E. et al. (1963). *Slovenská gramatika*. Bratislava, Slovenské pedagogické nakladateľstvo.
- [22] Pauliny, E. (1981). *Slovenská gramatika (Opis jazykového systému)*. Bratislava, Slovenské pedagogické nakladateľstvo.
- [23] Stanislav, J. (1977). *Slowakische Grammatik*. Bratislava, Slowakischer Pädagogischer Verlag.
- [24] Mistrik, J. (2003). *Gramatika slovenčiny*. Bratislava, Slovenské Pedagogické nakladateľstvo.
- [25] Orlovský, J. (1971). *Slovenská syntax*. Bratislava, Obzor.
- [26] Pavlovič, J. (2012). *Syntax slovenského jazyka I*. Accessible at: <http://pdf.truni.sk/e-ucebnice/pavlovic/syntax-1>.
- [27] Mistrik, J. (1988). *A Grammar of Contemporary Slovak*. Bratislava, Slovenské pedagogické nakladateľstvo.
- [28] Pauliny, E. (1997). *Krátka gramatika slovenská*. Bratislava, Národné literárne centrum – Dom slovenskej literatúry.
- [29] Šimková, M. et al. (2017). *Slovenský národný korpus: Texty, anotácie, vyhľadávania*. Bratislava, Mikula.
- [30] *Slovenský národný korpus – prim-7.0-public-all*. Bratislava, Jazykovedný ústav Ľ. Štúra SAV 2015. Accessible at: <http://korpus.juls.savba.sk>.

ANALYSIS OF VERBAL PREPOSITIONAL “OF” STRUCTURES

MARIANNA HUDCOVIČOVÁ

University of Ss. Cyril and Methodius in Trnava, Slovakia

HUDCOVIČOVÁ, Marianna: Analysis of verbal prepositional “of” structures. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 191 – 199.

Abstract: The article presents empirical research of verbal prepositional “of” structures, grammatical collocations of the verb and the preposition OF. The preposition OF belongs among the most frequent prepositions in the English language. The study is based on comparisons of English and Czech sentences containing verbs and prepositions that are followed by the object. Material was taken from the electronic data bank Prague Czech-English Dependency Treebank 2.0. The structures were examined and analyzed from morphological, syntactical and semantic points of view. The aim of the study is to create English-Czech verbal prepositional counterparts; to create verbal prepositional groups on the grounds of the similar semantic, syntactic features; to identify the features that are the same for each verb group and generalize them; to identify trends and tendencies for verbs when they collocate with a certain preposition. The findings are presented in several charts and tables.

Keywords: verbal prepositional structure, grammatical collocations, verbal semantic group, preposition “of”

1 AIMS OF RESEARCH

There are 83 simple prepositions in English. For the analysis, one of the most frequent ones was chosen, the preposition OF. Verbal prepositional groups (verb complementation) are to be established due to similar semantic and syntactic features. Their generalization, the trends in verbs collocating with a particular preposition will be sought. This sort of analysis requires study of the context in which the prepositions occur.

The analysis will be based on the following hypothesis: “Regular“ verbal collocations prevail over the “irregular“, i.e. coincidental contexts. Verbs with a similar meaning belong to the same semantic group and collocate with the same prepositions.

2 SYNTACTIC AND SEMANTIC ASPECT OF VERBAL PREPOSITIONAL STRUCTURES

In the analysis of prepositional verbal phrases from a syntactical point of view the group of verbs is defined that is usually complemented by and object and

define transitive verbs. Kudrnáčová [5] contributed to the semantico-syntactical analysis of the selected groups of verbs, i.e. motion verbs. The author postulated that “the number and types of complements (and their possible combinations) are not associated with individual verbs but with verbal classes. A certain set of semantic features is shared by all members of the given verb class. These features then represent those components of the verbal lexico-semantic content that are syntactically relevant, i.e. that determine the verbal syntactic behaviour” (p. 8).

Panevová [7], Levin [6], Anderson [1], Wierzbicka [9] and Jackendoff [3] claim that the syntactic and semantic levels are very closely related. In order to detect the specific features of the verbs and prepositions it is necessary to study both levels of the linguistic system. Katz [4] stated – regarding the analysis of verbal-prepositional structures – that the decomposition of the verb can serve as a key. Quirk et al. [8] and Dušková [2] offer classification of verbs according to semantic criteria, i.e. certain semantic features that are shared by a particular group of verbs. Levin (1993) claims that syntactic properties are semantically determined and sought to what extent the meaning of a verb determines its syntactic behaviour. What is really important is to find an effective method for identification of the relevant components of verbal meanings.

3 RESEARCH METHODOLOGY

The source of complete sentences with verbal prepositional structures (VPS) was the Prague-English Dependency Treebank 2.0. The Prague Czech-English Dependency Treebank 2.0 is a sentence-parallel manually annotated treebank. It is a manually parsed Czech-English parallel corpus sized over 1.2 million running words in almost 50,000 sentences for each part. The annotation includes also links to two valency lexicons, PDT-VALLEX for Czech and Engvallex, which contains 6 213 valency frames for 3 823 verbs for English. PDT-VALLEX holds 10 593 valency frames for 6667 verbs. The English part contains the entire Penn Treebank-Wall Street Journal section. The Czech part consists of Czech translations of all of the Penn-Treebank- WSJ texts. The corpus is 1:1 sentence aligned. I used PML-TQ open source search tool for parallel English-Czech treebanks.

The analysis comprised VPS containing preposition OF in English and their equivalents in Czech (also registered in the full context). The VPS were classified into groups of the verb phrases devised for this study. The VPS belonging to an identical semantic group are examined according to two criteria: syntactic relations and lexico-semantic relations. Next, the tendencies for each semantic group were sought. In the survey, qualitative and quantitative approaches were used and the method of contrastive analysis was applied.

4 SEMANTIC GROUPS

Based on the data taken from the English Dependency Treebank, it was possible to include almost all verbs into semantic groups. The following semantic groups were formed: communication; consist; take away sth. from sb.; be guilty of a crime; purify and ask. Only 4 verbs were not included into semantic groups because they did not share similar lexico-semantic features. (i.e. *remain*, *partake*, *relieve* and *dream*).

4.1 Semantic group with the meaning “communication”

The first group consists of verbs with a similar meaning denoting communication and cognitive processes. The verbs *tell*, *say*, *talk* and *speak* represent oral communication and transmission of messages. These verbs collocate with other prepositions, e.g. the verb *talk* with prepositions *to*, *of*, *about*, *with* and *at*, each with a difference in meaning. The verbal prepositional structure *hear of* denoting getting information through an audio channel belongs to this group as well.

The verbs *notify* and *inform* can be considered synonymous. In Czech they collocate with the preposition *o*, e.g. *uvědomiti o plánu*, *informovat o*; *they already put warning labels in their catalogs informing customers of the one-party law* ‘již do katalogů zařazují varování informující zákazníky o zákoně o jednostranném souhlasu’. Together with the preposition *of* and the postponed nominal phrase, e.g. *notify of invitation*, *plan*, *plot*, *responsibility*, *transaction*, they express the subject of the communication. The verb *warn* can be considered as a synonym of *notify* or *inform*, only with a stronger meaning *to inform someone of a possible danger or problem*.

The verbs *think*, *know* and *learn* can be classified as verbs denoting cognitive processes. The verbal prepositional structure *think of* was the most frequent in the group that is collocated with the preposition *of* in this research. It was translated into Czech as *myslet o*, *uvažovat o*, *vzpomenout si na* or *přijít na*. Translation of the verbal prepositional structure *think of* depends on the context of the whole sentence. As the frequency of the verb is high, there are more translation options. The preposition *of* together with the nominal phrase that follows expresses the topic of the cognitive process – thinking, e.g. *think of commuting*, *cooperation*, *future*, *money*, etc.

The verbal prepositional structure *know of* belongs to the group of cognitive processes as well. It was translated by the Czech structure *vědět o*, e.g. *know of ambition*, *a plan*, *the risk*, *technology*, *the use*, etc. The preposition *of/o*, together with the nominal phrase, expresses the result of the cognitive process of thinking. The verb *know* collocates with other prepositions as well, e.g. *about*, as does the verb *think*. Unlike the verb *think*, however, the preposition *about* is not interchangeable with *of*. They have, like the verb *know*, different meanings, e.g.

know of is defined in the Oxford Advanced Learner's Dictionary as: "to have information about or experience of somebody, something; *know about* is defined as: to have knowledge of something, to be aware of something."

The last verbal prepositional structure that denotes cognitive processes is *learn of*, translated into Czech as *dozvědět se o*, e.g. *learn of development, infection, practice, etc.* The complete structure refers to the cognitive process of getting information.

The meaning of the verbal prepositional structure *assure of* is defined in the Oxford Advanced Learner's Dictionary as follows: "to tell somebody something positively or confidently, esp. because they may have doubts about it," e.g. *assure of a paycheck*. Its Czech translation was 'ujistiti se o platu'.

The verbal prepositional structure *convince of* expresses an oral communication used to persuade somebody of something. The research material contained such collocations as *convince of need, support or worthiness; to convince anti-abortion activists of his stalwart support* 'aby přesvědčil aktivisty vystupující proti potratům o své věrné podpoře'. In all cases the structure was translated as *přesvědčit o*.

The reaction to persuasion can be an agreement. This is expressed in the verbal prepositional structure *approve of*, e.g. *approve of abortion* 'souhlasit s potratmi'. The counterpart of *of* is *s*.

This structure is defined in the dictionary as: "to say that one is annoyed, unhappy or not satisfied." With this definition comply the phrases *complain of loss and policy*. The second meaning of the verb *complain* was found in the context *complain of moonlighting*. People complain of moonlighting, a person having a second job, when it interferes with the quality of their work or is a conflict of interest.

The most frequent Czech preposition was *o*, e.g. *think of* – 'myslet o', *tell of* – 'říct o', *talk of* – 'mluvit o', *say of* – 'říct o' and *inform of* – 'informovat o'. The second most frequent preposition was *na*, e.g. *think of* – 'vzpomenout si na, přijít na', *warn of* – 'upozornit na'. The preposition *před* occurred once, e.g. *warn of* – 'varovat před'.

4.2 Semantic group with the meaning "consist"

This is the second most frequent group, containing the verbs *consist*, *compose*, *make*, and *come*. The verbs share a similar meaning: *to be made of* or *form from*. In the following sentence the verb *compose* was translated as 'patří'. The translator translated the sentence to make it sound natural in Czech and the verb *patřit* is a better choice than the original *sestávat z čeho*, e.g. *and a third category is composed of disorders whose treatment is difficult or impossible if a person lacks adequate shelter*. 'a do třetí kategorie patří potíže, jejichž léčba je obtížná nebo nemožná, pokud osoba postrádá vhodné přístřeší'.

The verb *consist* was translated as *sestávat z*, e.g. *sestávající ze 100 milionů dolarů* 'consisting of \$100 million'. (114) *Záruka sestává ze zaručených půjček*, 'The collateral consists of collateralized whole loans'.

The verbal prepositional structure *make of* occurred only in the passive voice and expresses creation of a product from a raw material, e.g. (115) *Cheerios and Honey Nut Cheerios are made of oats* ‘Řady Cheerios a Honey Nut Cheerios se vyrábějí z ovsa’.

The expression *come of* has a similar original meaning as the previous verbal prepositional structures. It expresses origin. (116) *Ringers, she added, are “filled with the solemn intoxication that comes of intricate ritual faultlessly performed”*. ‘Zvoníci, dodala, jsou “prodchnuti slavnostním opojením, které vychází z rafinovaného, dokonale provedeného obřadu”’.

The preposition *z*, the Czech equivalent to the English *of*, occurs in all verbal prepositional structures that belong to this group. The preposition makes it clear that the product is made of a certain material.

4.3 Semantic group with the meaning “take away something from somebody”

The verbal prepositional structures *strip of*, *defraud of* and *deprive of* denote the definition of this group: to break principle. The verbs share the same meaning: “take away something from somebody.” In the case of the verb *strip*, property or honours is taken away. In the second case, the verb *defraud* expresses taking something illegally from a person. The last verb *deprive* denotes taking something necessary or pleasant from someone. Translators used the preposition *o* in all verbal prepositional structures to specify what is being taken away, e.g. *...attempts to strip the president of his powers* ‘zkušely připravit prezidenta o jeho moc’; *to defraud the Army of \$21 million* ‘připravit armádu o 21 milionů dolarů’; *deprive of right* ‘připravit o právo’.

4.4 Semantic group with the meaning “be guilty of a crime”

There are two prepositional structures with the preposition *of* that express the meaning to be guilty of a crime, e.g. *accuse of* “to say that somebody done something wrong, is guilty of something or has broken the law” and *convict of*: “to decide in a law court that somebody is guilty of a crime”. Both of them are translated by the preposition *z*, *accuse of* – ‘vinit’ *z* and *convict of* – ‘uznat vinným z’. The preposition *of/z* expresses origin, e.g. (117) *I’m not accusing insurers of dereliction of duty*. ‘Neviním pojišťovny ze zpronevěření se povinnosti’. In the following example, the preposition *of/z* is used as the only preposition that collocates with the verbs *accuse* and *convict*, e.g. *... when someone is convicted of a felony*. ‘... pokud je někdo uznán vinným ze zločinu’. In the survey, expressions were found describing various kinds of crime that collocate with the preposition *of*, e.g. *convict of trespassing, crime, extortion, felony* or *kidnapping*.

4.5 Semantic group with the meaning “purify”

The meaning of the verb *clear somebody of something* is: “to show a person’s innocence”. Translation of this verb was identical with the verbal prepositional structure *cleansed of* ‘očistit od’. The verbs, however, have different meanings. By the verb *clear* the original meaning “to purify” is shifted into a metaphoric or figurative meaning.

The verbal prepositional structure *cleanse of* is defined in the Oxford Advanced Learner's Dictionary as follows: "to make somebody or something thoroughly clean." The survey brought two occurrences of this verbal prepositional structure. In the first case, the meaning is original, to wash away dirtiness, e.g. *cleanse of muck*. In the second example the meaning is figurative, e.g. *cleanse of sin*. The preposition *of/od* with the verbs *clear* and *cleanse* expresses "getting rid of something".

4.6 Semantic group with the meaning "ask"

There are two verbal prepositional expressions with the same fixed structure, i.e. *ask something of somebody* and *require something of somebody*. It can be said that the verb *ask* represents a mild form of request. Czech translators used the same Czech verb in both cases: *vyžadovat od*, e.g. (118) *He says the big questions aren't asked of companies coming to market. 'Říká, že odpovědi na hlavní otázky se od společností, které přicházejí na burzu, nevyžadují'. (119) It is required of me that I give evidence. 'Vyžaduje se ode mně, že podám důkazy'.*

The Czech counterpart of the preposition *of* is again *od*, which expresses an administrator of the request.

4.7 Other verbal prepositional structures taking the preposition "of"

There are verbs that cannot be put into any group because of their different lexico-semantic features. The following verbal prepositional structures taking the preposition OF were found in the survey material and classified with this last miscellaneous group, e.g. *remain of*, *partake of*, *relieve of* and *dream of*.

Prepositional structures are translated into Czech by using different prepositions that are connected with verbs. The verbal prepositional structure *remain of* was translated into Czech as *zůstat z*, e.g. *what remains of the oil tycoon's once-vast estate* 'co pak zůstane z kdysi obrovského majetku olejového magnáta'.

The verbal prepositional structure *partake of*, synonymous with the verb *take part* means "to become involved or take part in something". It is interesting that this verb with the previous meaning collocates in English with the preposition *of*, which expresses mainly the partitive meaning, not sharing or involvement. In Czech or Slovak it collocates with *na*, which denotes involvement, e.g. *domestic franchisees apparently didn't partake of the improvement*. 'domáci provozovatelé licence se zjevně na tomto zlepšení nepodíleli'.

The verbal prepositional structure *relieve of* expresses the meaning "to release somebody from a duty or task by taking their place or finding somebody else to do so." It was translated as e.g. *relieve of duty* 'uvolnit z funkce'. The Czech counterpart *z* denotes, together with the nominal phrase *duty*, a partitive object.

The third verbal prepositional structure *dream of* was translated as *snít o*. The preposition *of/o* expresses together with the object the topic of dreaming, e.g. (120) *It's one more, too, for the fans who dream of a season that never ends*. 'Je to také

další příležitost pro fanoušky, kteří sní o sezóně, která nikdy neskončí’. The verb *dream* may collocate with *about*, and it is interchangeable with *of*.

In all three cases, a different Czech counterpart of the English preposition *of* was used. The verbal prepositional structures from the last group occurred once or twice. It is therefore difficult to analyze semantic features which they might have shared.

5 DISCUSSION ON THE VERBAL PREPOSITIONAL “OF” STRUCTURES

The most frequent prepositional *of* structure was the verbal prepositional structure *think of*, which occurred 34 times. In second place was *make of* with 30 occurrences. *Consist of* with 27 structures came third. Among the 20 most frequent structures, 12 belong to the group with the meaning “communication and cognitive processes”, e.g. *think of*, *say of*, *notify of*, *know of*, *approve of*, *inform of*, *warn of*, *learn of*, *tell of*, *talk of*, *convince of*, and *speak of*. It means that more than the half of the structures collocating with *of* expresses various ways of communication or knowledge. The group of structures representing the meaning “consist”, “take away something from somebody” or “to be guilty” is relatively small with two or three occurrences. The group “communication and cognitive processes” is the largest group and moreover, the most frequently used verbs belong to it. It can be said that communication itself is vitally important in human society. Therefore the ways of communication and their expression vary considerably. This may be the reason why the verbs of this type are the most numerous.

In the most frequent group “communication,” the preposition *of* with the nominal phrase that follows expresses mainly the topic of communication, e.g. *say of resignation* or the result of a cognitive process: getting information, e.g. *know of technology*. The meaning of the preposition *of* in the second group “consist” represents, together with the nominal phrase, the partitive object.

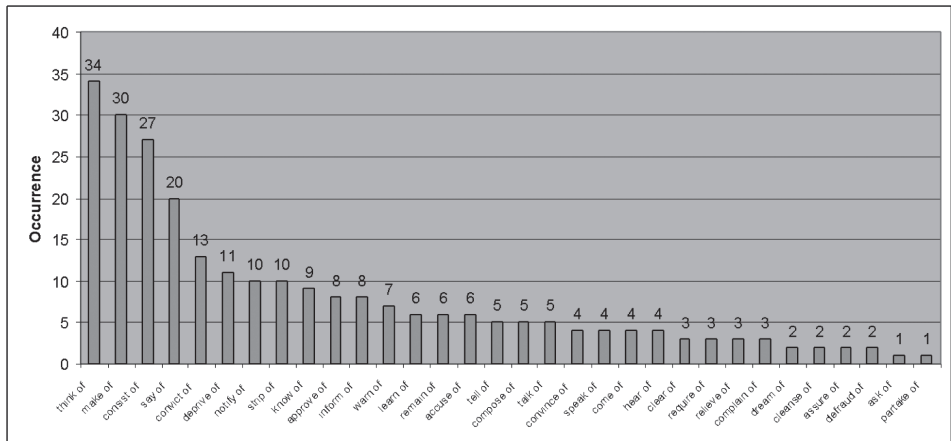


Chart 1. Number of occurrences of verbal prepositional OF structures

think of	myslet o, uvažovat o, vzpomenout si na, přijít na	34
make of	vyrábět z, vytěžit z	30
consist of	sestávat z	27
say of	řít o, uvádět o, prohlásit o	20
convict of	uznat vinným z	13
deprive of	připravit o	11
notify of	uvědomiti o	10
strip of	připravit o	10
know of	vědět o	9
approve of	souhlasit s	8
inform of	informovat o	8
warn of	varovat před, upozornit na	7
learn of	dozvědět se o	6
remain of	zůstat z, zbýt z	6
accuse of	vinit z	6
tell of	řít o	5
compose of	sestávat z	5
talk of	mluvit o	5
convince of	přesvědčit o	4
speak of	svědčit o	4
come of	vycházet z	4
hear of	slyšet o	4
clear of	očistit od	3
require of	vyžadovat od	3
relieve of	uvolnit z	3
complain of	stěžovat si na	3
dream of	snít o	2
cleanse of	očistit od	2
assure of	ujistiti se o	2
defraud of	připravit o	2
ask of	vyžadovat od	1
partake of	podílet se na	1
Total		258

Tab. 1. English verbal prepositional OF structures and their Czech equivalences with number of occurrences

Group “communication”	O/106; NA/15
Group “consist”	Z/ 72
Group “take away sth. from sb.”	O/23
Group “be guilty of a crime”	Z/19
Group “purify”	OD/5
Group “ask”	OD/4

Tab. 2. Occurrence of Czech equivalent prepositions in semantic groups

The data reveal that each semantic group is represented by one preposition, with one exception. The most frequent group, “communication”, was represented by two prepositions of which O represents 88% and NA 12%. This is due to the shifted meaning of *think of*, which was translated according to the context as: *přijít na*.

The most frequent equivalency was OF/O. It can be found in connection with verbs expressing the topic or subject of communication. The second most frequent equivalency is OF/Z, found in the semantic group “consist,” detecting the origin or material from which the product was made.

Occurrence of one preposition confirms hypothesis that “Occurrence of the “regular“ verbal prepositional phrases prevail over “irregular coincidental” ones.” The survey produced 28 “regular” verbs that collocate with the preposition *of*, ‘regular’ referencing verbs that can be classified and put into semantic group together with verbs of similar semantic features. There were only four “irregular” verbs which cannot be put into any semantic group. That also confirms hypothesis that “verbs with a similar meaning belong to the same semantic group and are bound with the same preposition.” The most frequent group contains many verbs expressing communication, e.g. oral communication is served by the verbs *say*, *tell*, *talk* and *speak*.

References

- [1] Anderson, J.M. (1971). *The Grammar of Case: Towards a Localistic Theory*. Cambridge, Cambridge University Press.
- [2] Dušková, L. et al. (1988). *Mluvnice současné angličtiny na pozadí češtiny*. Praha, Academia.
- [3] Jackendoff, R. (1992). *Semantic Structures*. Cambridge, Massachusetts: The M.I.T. Press.
- [4] Katz, L. (2003). *Semantics*. 2nd ed. Oxford, Blackwell Publishing.
- [5] Kudrnáčová, N. (2008). *Directed Motion at the Syntax-Semantics interface*. Brno, Masarykova Univerzita.
- [6] Levin, B. 2009. *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago, The University of Chicago Press.
- [7] Panevová, J. (1974). On Verbal Frames in Functional Generative Description I. *The Prague Bulletin of Mathematical Linguistics*, 22.
- [8] Quirk, R., Greenbaum, S. et al.(1985). *A Comprehensive Grammar of the English Language*. London, Longman.
- [9] Wierzbicka, A. (1988). *The Semantics of Grammar*. Amsterdam, John Benjamins.

TEMPORAL ‘SINCE’ IN SLOVAK: CONJUNCTION(S) AND ASPECT CHOICE – A CORPUS STUDY

PAULA KYSELICA – RENÉ M. GENIS
University of Amsterdam, The Netherlands

KYSELICA, Paula – GENIS, M. René: Temporal ‘since’ in Slovak: conjunction(s) and aspect choice – a corpus study. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 200 – 215.

Abstract: It has recently been shown by especially [1] through [4] and [12] for Russian and by [8] and [9] for Polish that conjunctions corresponding to Dutch *sinds* (cf. also [1], [2], [3]) and English *since* (cf. also [7], [10]) have temporal functions, which are subject to restrictions on the choice of tense and aspect. Ultimately these restrictions can be related to the semantic input of tense and aspect into complex sentences with these connective items. For Polish extensive data provided by corpus research enabled us to shed light on the usage and restrictions in this area and also to establish which constellations with particular conjunctions are more or less likely or not possible (cf. [8], [9]). In the present contribution we present freshly sourced quantitative Slovak SNK-corpus data. We consider the sixteen logically possible tense-aspect constellations, and the Slovak connective items: *odkedy; odvtedy, čo / ako; od chvíle, keď / čo / ako; od tých čias, čo / ako; od tej doby, čo / ako*. This quantitative data study is intended to pinpoint the areas of future research; for this purpose at certain instances comparisons are made with Polish, the only other language we have such data for to date.

Keywords: conjunction, tense, aspect, anteriority, simultaneity, taxis, Slovak, Polish

1 INTRODUCTION

This contribution presents the finding of our investigation into the Slovak correspondences of Dutch and English temporal conjunctions, respectively *sinds/sedert* and *since* and the tense-aspect (hereafter TA) constellations of the complex sentences they appear in. The underlying research is part of the ongoing taxis project of the research group “Comparative Slavic Verbal Aspect” at the university of Amsterdam.¹ Our research group usually work within a cognitive-structuralist framework. Earlier research on this particular conjunction included Czech, Polish and Russian.

¹ Cf. <https://aclc.uva.nl/content/research-groups/comparative-slavic-verbal-aspect-and-related-issues/comparative-slavic-verbal-aspect-and-related-issues.html?origin=8ZtCo6MjS%2B6atiMzaszh6A>.

1.1 Setting the scene

The following example is in a few ways very typical of the kind of complex sentence we are dealing with.

- (1) *Odkedy ťa stretol, všetko sa zmenilo.* [LŠti3]²
Since you.ACC meet.M.3SG.PST.PFV, everything REFL change.N.3SG.PST.PFV³
'Since he (had/has) met you, everything (had/has) changed.'⁴

We see here the three noteworthy distinctive elements.

- the Secondary Clause (hereafter SC) *Odkedy ťa stretol* with conjugated verb *stretol* and
- 'since'-connective item *odkedy*;
- the Main Clause (hereafter MC) with conjugated verb *zmenilo*.

Each of the two conjugated verbs introduce TA-meaning, which interacts between SC and MC, but also probably the selected connective item.

A basic semantic analysis of this type of construction based on the earlier research into Dutch, English, Russian and especially Polish (all earlier references) is provided here for a better understanding as it probably will largely coincide in its generalities, although such an analysis is not the main focus of the current contribution, and a lot more can probably be said about that once the Slovak samples have been thoroughly scrutinized for that purpose. The invariant that has been established, then, consists of the following elements:

- The *sinds/since*-connective item introduces an SC-event which starts in the past;
- The connective item carries a sense of anteriority, which has to do with the beginning in the past;
- The SC-event sets, one might say, "opens" a temporal frame (the "SC-frame"), which stretches from that beginning in the past up to and including the deictic center, which may be at the moment of utterance or before, but which need not be "filled" entirely/throughout with the SC-event itself;
- The MC-event takes place in the temporal frame set by the SC (although it need not "cover" it exactly);
- There is also a sense of simultaneity in this construction, which comes about as the MC-event takes place against the SC-temporal frame (= at some time in that temporal frame).

² The source references are verbatim as they are provided by the Slovak National Corpus – prim-8.0-public-sane (hereafter SNK).

³ Our interlinear glosses follow the Leipzig Glossing Rules. (<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>). See also our list of abbreviations below.

⁴ Please note that in the glosses and translation of examples we have not wanted to pinpoint the exact English aspect-forms. That would be quite impossible and is very dependent on context and a few other factors, which have no bearing on the Slovak originals.

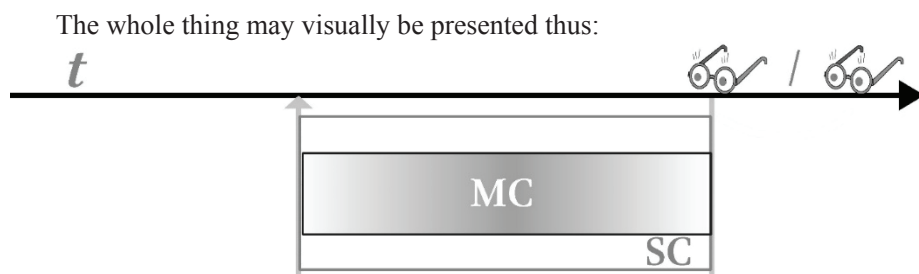


Fig. 1. Invariant meaning of *sinds/since* constructions (after [9] for Polish)

The spectacles represent the location from which an event is observed and it may or may not coincide with the moment of utterance or come after, depending on the particular TA-constellation used.

1.2 Sample examples

The following is a list of examples with more or less randomly chosen *sinds/since*-connective items – it was not thought necessary in this paper with its main focus on the quantitative data to provide an example of each TA-constellation for each connective item, although the exchangeability of the items has not yet been studied. You will notice that passive forms have been deselected here, and choices are intended to be as “riskless” as possible, again, because the deeper semantic analysis is beyond the scope of this paper and the following provides a sufficient impression for our present purposes.

Please note, that in Dutch and English (for which latter, cf. [10, p. 91] the temporal location of the SC-event may not be placed in its entirety in the future. SC-future tenses have not been encountered for Polish and for Slovak they have been omitted here; moreover, they are not likely to fall within the scope of *sinds/since* and one of our purposes is a comparison with Polish. Please note in this respect that PRS. PFV is not qualified here as future, although it deictically often functions as such.

	SC		MC		example (ex SNK)
1	PST	PFV	PST	PFV	(1) <i>Odkedy ťa stretol, všetko sa zmenilo.</i> [LŠti3] 'Since he (had/has) met you, everything (had/has) changed.' ⁵
2	PST	PFV	PST	IPFV	(2) <i>Od chvíle, keď odišla sanitka, pri Holmbergových nohách sa chúlil Wurst.</i> [KBlo1] 'From the moment the ambulance left, Wurst snuggled to Holmberg's feet.'

⁵ Please note that the English translations in this table are for working purposes only: their TA-constellations will depend on matters at play in English that are beyond the scope of our research here. Only in example (1) have we given a few alternatives to give an impression. Please, note also that example (1) is the same as provided above and so we kept the same number.

	SC		MC		example (ex SNK)
3	PST	PFV	PRS	IPFV	(3) <i>Sledujeme vás od chvíle, čo ste vkročili do močiara</i> [JČerv3] 'We've been following you since you stepped into the swamp.'
4	PRS	IPFV	PRS	IPFV	(4) <i>Od tej doby, ako hrávam futbal, viem čo mám robiť.</i> [MYTR2009/01] 'Since [*I play] I have been playing football, I know what to do.'
5	PST	IPFV	PST	PFV	(5) <i>S nostalgiou si uvedomil, že odvtedy, čo bol bezstarostný chalan, zmenilo sa mnoho vecí.</i> [EMcb11] 'With nostalgia, he realized that many things had changed since he was a carefree boy.'
6	PST	IPFV	PST	IPFV	(6) <i>Prvý raz odvtedy, čo žil na veľchánovom dvore, nesprevádzal cisára na letné sídlo.</i> [WMei1] 'For the first time since he had been living at court, he did not escort the emperor to the summer residence.'
7	PRS	IPFV	PST	PFV	(7) <i>Od tej doby, čo sa zúčastňujem tohto projektu, som sa naučila vážiť si samu seba.</i> [ASP2002/01] 'Since I attended this project, I have learned to appreciate myself.'
8	PST	PFV	PRS	PFV	(8) <i>A odvtedy, čo ušli trpaslíci, nikto sa neodváži prehľadávať šachty a poklady v hĺbinách.</i> [JTol2] 'And since the dwarves have gone, no one dares to search the shafts and treasures in the depths.'
9	PRS	IPFV	PST	IPFV	(9) <i>Nielen bohovia, ale aj ľudia vždy potrebovali smiech a zadovažovali si ho od tých čias, čo tu existujú, lenže veľmi primitívne...</i> [ABed8] '[...] but also the people always needed laughter and they obtained it since they existed here [...] ...'
10	PST	IPFV	PRS	IPFV	(10) <i>Ved' odvtedy, ako účinkovali v našom programe, nepretržite pracujú až dodnes.</i> [MYNO2016/09] 'After all, ever since they participated in our program, they have been working [in work] uninterruptedly until today.'
11	PRS	IPFV	PRS	PFV	(11) <i>Odvtedy, čo navštevujem Grécko, vždy si priveziem 5 litrov olivového oleja, ...</i> [MYŽN2009/31] 'Since I have been visiting Greece, I every time bring 5 liters of olive oil, ...'
12	PST	IPFV	PRS	PFV	(12) <i>... ešte vlani ho piekli každý druhý deň, ale odvtedy, čo šiel starý Nérer do penzie, zoženieš ho už len v piatok)...</i> [JJoh2] '... Last year they bake it every other day, but since old Nerer went to the pension, you will only get it on Friday) ...'

Tab. 1. Sample examples of all researched TA-constellations – random connective items

These 12 examples cover all the basic TA-constellation types and the following further examples are provided to cover the special types that are mentioned for

Polish [8] and have now been recognized in Slovak. To complete this issue, we will briefly discuss these.

In both the Polish and now also the Slovak dataset, many examples of various TA-constellations occurred with the meaning ‘remember’.

- (13) *Ja od chvíle, čo si pamätám, som vždy cítil potrebu bojovať.* [HN2011/05]
‘Me, ever since I can remember / for as long as I can remember, I have always felt the need to fight.’

These are translatable by a *sinds/since* construction but there the alternative ‘for as long as’ is usually lacking for other verbs. In terms of taxis this is a noteworthy difference.

Another frequent subtype, for which its distribution across the TA-constellations in Slovak still needs to be studied, might be described as ‘the passing of time’, cf.(14).

- (14)a *Zistuje, že ubehlo asi dvadsať minút od chvíle, keď zdriemol.* [LT1998/07]
‘He discovers that about twenty minutes have passed since he took a nap.’

A perhaps somewhat unexpected type has future relevance, but is nevertheless translatable with *since*.

- (14)b *Dvadsiateho siedmeho decembra uplynie mesiac odvtedy, čo má sadru.* [KOR2001/12]
‘On the twenty-seventh of December, a month will have passed since he has a cast.’

Not at all surprising is the fact that phase verb ‘begin’ crops up often in this dataset; the connective’s meaning ‘since’ marks a point at which an event commences. Example (15) is rather special as it has such a phase verb both in SC as in MC.

- (15) *Narodí sa nám dieťa, no nie dieťa, ktoré jednoducho začne umierať od chvíle, keď začína žiť.* [KRah1]
‘A child was born unto us, but not a child who simply begins to die from the moment he begins to live.’

The last of our special mentions is an issue touched upon already by [10] and [7] for English and by o.a. [8], [11] and [12] for Slavic languages. It is the issue of the necessity in many examples for an element in the MC that gives some sort of “weight”, some specific “load” or relevance, without which a sentence may be grammatically well-formed, but is nevertheless bad. In (16) the relevant element is underscored.

- (16) *Myšlienka na odchod z ostrova mi neskrsla v hlave ani raz od chvíle, čo som na ňom pristál.* [YMar1]
'The thought of leaving the island didn't occur to me even once, since I came upon it.'

This concludes the samples and we shall leave the remainder of the semantic analysis for future research: the reader at this point should have a sufficient impression of what is involved with the *sinds/since* temporal constructions.

In this paper we proceed by describing the methods used for identifying Slovak correspondences to Dutch *sinds* and English temporal *since* using the *ASPAC* parallel aligned corpus and for compiling our data set using the SNK. Further on we present the quantitative data: on that basis we will discuss the statistical analysis and propose a few lines of research to follow up, as ultimately the (future) goal is to make an extensive semantic analysis and establish the set of usage restrictions for each of the individual connective items.

2 DATA SET, METHODOLOGY

2.1 Identification of the correspondences

As mentioned in 1. the starting point for this research were the Dutch connective item *sinds* and English *since* in temporal usage.⁶ As for our earlier research on Polish, *ASPAC* – Amsterdam Slavic Parallel Aligned Corpus⁷ was used: simple queries via ParaConc yielded the correspondences: *odkedy; odvtedy, čo; odvtedy, ako; od chvíle, ked'; od chvíle, čo; od chvíle, ako; od tých čias, čo; od tých čias, ako; od tej doby, čo; od tej doby, ako.*

Hereafter we will refer to these as 'connective items' rather than as 'conjunctions' on account of their sometimes complex shape, consisting of more than one lexical element, which are sometimes exchangeable. In this respect it is important to note that this method to identify the connective items deselects such items that fall outside the (semantic and other) scope of the connective items in the source languages. This is important as it has been established ([8], [9]) that at least the Polish connective items are not primarily restricted to use with past deixis such as their Dutch and English counterparts. It was expected – and has indeed been established, cf. the following – that this holds for the Slovak counterparts as well.

⁶ The English connective items under scrutiny are *since* and *ever since* in their temporal functions only. As was pointed out in [6], [7], [8], [9] and [10] non-temporal *since* such as we find in e.g. *Since I'm a taxi driver, I know how to get there* allows for TA-constellations that are at variance with those for temporal (*ever*) *since*. Dutch *sinds* is always temporal.

⁷ *ASPAC* is a non-tagged, restricted access corpus compiled by Adrie Barentsen. It consists of original and translated Slavic literary texts. Cf. <http://www.uva.nl/profiel/b/a/a.a.barentsen/a.a.barentsen.html>. We are indebted to Barentsen for the access provided.

2.2 Strategies for compiling the data set with the SNK

Our data, once sourced from the SNK, was set out onto a simple Excel-sheet. It was our purpose to classify all corpus search results per connective item and TA-constellation of MC and SC taken together. Of the sixteen logically possible combinations, twelve were actually encountered and found to equate *sinds/since* usage and they are plotted onto tables 1 through 3. The four possible constellations with SC:PRS.PFV only yielded a very limited number of samples that (for various reasons) never correspond to *sinds/since*. Obviously the expected future deixis of SC:PRS.PFV, often sets a time frame in the future, which is incompatible with *sinds/since*. There are other, not yet researched instances as well: cf. (17) in which there is a kind of generic/repetitive (exemplary) meaning intended.

(17) *Väčšina ľudí si o vás vytvorí prvý dojem najneskôr do piatich sekúnd od chvíle, ako vás zbadá.* [InZ2000/02]

Most people DAT about you make.3SG.PRS.PFV first impression at_most to five seconds from moment that you.ACC notice.3SG.PRS.PFV.

‘Most people will have made a first impression within five seconds from the moment they will (have) notice(d) you.’

As such these types have been deselected from this research.

As we were searching the SNK for our connective items, it soon became clear that some search strings yielded many thousands of hits (e.g. 16067 for *odkedy* alone), whilst others amounted to just a tiny handful of examples. It was decided early on that it would be more practical to have a different treatment for the highly frequent connective items as opposed to the rather lower scoring items; the material for the lower scoring items – *od chvíle, ako; od tých čias, čo; od tých čias, ako; od tej doby, čo; od tej doby, ako* – counted “manually”, which means that the TA constellations for both MC and SC were classified by the researchers for all hits without searches for tagged grammatical categories. We considered this to be the most accurate way to deal with material with low scores.

The higher scoring connective items – *odkedy; odvtedy, čo; odvtedy, ako; od chvíle, ked; od chvíle, čo* – needed to be sourced, classified and counted via SNK-search strings with specified TA for both MC as well as SC. For this purpose we used the fact that the SNK corpus is tagged for tense as well as for aspect: verbs in PST and PRS are tagged respectively VL.* and VK.*, and PFV and IPFV are tagged V.d.* and V.e.* respectively. The search was done in two tiers to cover the two possible sequence-types of the clauses: SC-MC and MC-SC. The queries for *odvtedy, čo*, then, are as follows:

Type 1: clause sequence SC:PST.PFV-MC:PST.PFV:

Odvtedy, čo odišli zo Slovenska, už ubehlo desať rokov. [SME2014/07]

‘Since they (had/have) left Slovakia, ten years (had/have) passed.’

- Search CQL: [lemma="odvtedy"] [lemma=","] [lemma="čo"] within <s/>

- Filter: positive, range (0,10) incl KWIC, CQL: [tag="VLd.*"] [] {0,5} [word="",] [] {0,5} [tag="VLd.*"]⁸

Type 2: clause sequence MC:PST.PFV-SC:PST.PFV:

Prešiel už rok odvtedy, čo ocko oslávil päťdesiatiny. [EFBS1]

‘A year (has/had) already passed since father (has/had) turned fifty.’

- Search CLQ: [lemma="odvtedy"] [lemma="",] [lemma="čo"] within [tag="VLd.*"] [] {0,10} [tag="VLd.*"] within </>
- Filter: negative, range (-1,-1) excl. KWIC, CQL: lemma = ""

(The filter is applied to ensure the deselection of sentences that begin with the queried string itself when this belongs to the previous sentence in the text.)

For each of the logically possible TA-constellations as well as for each of the identified connective items the above search strings were adjusted.

In the further processing, the “order-types” were not separated out in any of the calculations as the clause order was deemed inconsequential for the present research (although we do not exclude it could be worth researching these variations at a later stage). The totals – and indeed all working figures of the higher scoring items – provided in table 1 are then the sum of the scores for these two order-types. The SNK search options allow for other solutions to our problems, but this turned out to be easy enough to apply successfully, although we realize that an “automatic” count is never as precise as the “manual” one, we applied for the lower scoring connective items. Some existing false positive or negative results do affect accuracy of the results but for totals in the order of thousands we consider the statistical error to be low enough for this early study. Further research is needed to establish the exact inaccuracy of similar corpus queries.

In the following we will present the quantitative data.

nr	SC		MC		sum of all connective items		sum of all connective items – odkedy		odkedy	
	T	A	T	A						
1	PST	PFV	PST	PFV	6456	26.04%	2884	32.90%	3572	22.28%
2	PST	PFV	PST	IPFV	5493	22.15%	2075	23.67%	3418	21.32%
3	PST	PFV	PRS	IPFV	4119	16.61%	1382	15.76%	2737	17.07%
4	PRS	IPFV	PRS	IPFV	1859	7.50%	240	2.74%	1619	10.10%

⁸ The limitations on the word range (= 10) were set after some experimenting with other ranges. Shorter ranges excluded many usable examples whilst wider ranges introduced inaccuracies in the scores on account of interference from other than the targeted verbs in the sentences.

nr	SC		MC		sum of all connective items		sum of all connective items – odkedy		odkedy	
	T	A	T	A						
5	PST	IPFV	PST	PFV	1615	6.51%	751	8.57%	864	5.39%
6	PST	IPFV	PST	IPFV	1431	5.77%	419	4.78%	1012	6.31%
5	PRS	IPFV	PST	PFV	1240	5.00%	87	0.99%	1153	7.19%
7	PST	PFV	PRS	PFV	888	3.58%	634	7.23%	254	1.58%
8	PRS	IPFV	PST	IPFV	834	3.36%	33	0.38%	801	5.00%
9	PST	IPFV	PRS	IPFV	659	2.66%	234	2.67%	425	2.65%
11	PRS	IPFV	PRS	PFV	168	0.68%	10	0.11%	158	0.99%
12	PST	IPFV	PRS	PFV	35	0.14%	18	0.21%	17	0.11%
totals:					24797	100%	8767	100%	16030	100%

nr	SC		MC		odvtedy, čo		odvtedy, ako		od chvíle, keď	
	T	A	T	A						
1	PST	PFV	PST	PFV	1407	32.32%	499	29.72%	455	40.23%
2	PST	PFV	PST	IPFV	716	16.45%	381	22.69%	400	35.37%
3	PST	PFV	PRS	IPFV	709	16.29%	352	20.96%	102	9.02%
4	PRS	IPFV	PRS	IPFV	113	2.60%	67	3.99%	9	0.80%
5	PST	IPFV	PST	PFV	460	10.57%	115	6.85%	103	9.11%
6	PST	IPFV	PST	IPFV	241	5.54%	84	5.00%	21	1.86%
7	PRS	IPFV	PST	PFV	34	0.78%	38	2.26%		
8	PST	PFV	PRS	PFV	507	11.65%	72	4.29%	22	1.95%
9	PRS	IPFV	PST	IPFV	18	0.41%	5	0.30%	1	0.09%
10	PST	IPFV	PRS	IPFV	138	3.17%	61	3.63%	7	0.62%
11	PRS	IPFV	PRS	PFV	6	0.14%		0.00%	3	0.27%
12	PST	IPFV	PRS	PFV	4	0.09%	5	0.30%	8	0.71%
totals:					4353	100%	1679	100%	1131	100%

nr	SC		MC		od chvíle, čo		od chvíle, ako		od tých čias, čo	
	T	A	T	A						
1	PST	PFV	PST	PFV	218	44.86%	159	27.46%	82	24.19%
2	PST	PFV	PST	IPFV	175	36.01%	264	45.60%	88	25.96%
3	PST	PFV	PRS	IPFV	23	4.73%	109	18.83%	44	12.98%
4	PRS	IPFV	PRS	IPFV	6	1.23%	7	1.21%	24	7.08%

nr	SC		MC		<i>od chvíle, čo</i>		<i>od chvíle, ako</i>		<i>od tých čias, čo</i>	
	T	A	T	A						
5	PST	IPFV	PST	PFV	21	4.32%	9	1.55%	32	9.44%
6	PST	IPFV	PST	IPFV	15	3.09%	18	3.11%	37	10.91%
7	PRS	IPFV	PST	PFV			2	0.35%	10	2.95%
8	PST	PFV	PRS	PFV	24	4.94%	8	1.38%		
9	PRS	IPFV	PST	IPFV	1	0.21%		0.00%	5	1.47%
10	PST	IPFV	PRS	IPFV	3	0.62%	3	0.52%	15	4.42%
11	PRS	IPFV	PRS	PFV				0.00%	1	0.29%
12	PST	IPFV	PRS	PFV				0.00%	1	0.29%
totals:					486	100%	579	100%	339	100%

nr	SC		MC		<i>od tých čias, ako</i>		<i>od tej doby, čo</i>		<i>od tej doby, ako</i>	
	T	A	T	A						
1	PST	PFV	PST	PFV	36	30.77%	22	30.99%	6	50.00%
2	PST	PFV	PST	IPFV	32	27.35%	19	26.76%		
3	PST	PFV	PRS	IPFV	24	20.51%	15	21.13%	4	33.33%
4	PRS	IPFV	PRS	IPFV	5	4.27%	7	9.86%	2	16.67%
5	PST	IPFV	PST	PFV	10	8.55%	1	1.41%		
6	PST	IPFV	PST	IPFV	3	2.56%		0.00%		
7	PRS	IPFV	PST	PFV	2	1.71%	1	1.41%		
8	PST	PFV	PRS	PFV	1	0.85%		0.00%		
9	PRS	IPFV	PST	IPFV	3	2.56%		0.00%		
10	PST	IPFV	PRS	IPFV	1	0.85%	6	8.45%		
11	PRS	IPFV	PRS	PFV				0.00%		
12	PST	IPFV	PRS	PFV				0.00%		
totals:					117	100%	71	100%	12	100%

Tab. 2. Quantitative data sourced from the SNK, processed in Excel

3 DISCUSSION OF THE QUANTITATIVE DATA

3.1 Slovak connective item

The numbers of SNK-hits were set out onto an Excel sheet and calculations were made to produce totals and percentages. Table 2 shows the results of this rounded off to two decimal places. The connective items are set out according to

frequency from left to right. The TA-constellations across SC and MC are sorted according to frequency from top to bottom of the total occurrence of each TA-constellation. In this way the preponderance/preference of each connective item per TA-constellation can be easily judged. For better legibility we have refrained from marking zero (and 0.00%) scores.

It will be apparent that of the logically possible TA-constellations four are missing and this was discussed in 2.2 above.

The data shows that *odkedy* is by far the most highly scoring connective item not only in total, but also in all constellation types, and even, albeit with the notable exception of type 7 (and less notably 12), in absolute numbers against the sums of the scores of the other connective items taken together. Its frequency of occurrence, however, does not trace that of the totals for all connective items together – even though those totals are of course heavily influenced by *odkedy* itself. For this reason as well as its unevenly high score we have produced a further column with totals for all connective items except *odkedy*. The differences between *odkedy*-scores against this column are even greater – as expected – and they display some interesting features which will be pointers for future research. For now we want to point out that our research has so far not produced any evidence that the respective connective items should not be exchangeable in any clear way and so we cannot at this point in time pinpoint why, in what circumstances a particular connective item is preferred. Here we can only mention in which TA-constellation there are significant deviations of *odkedy* against the others. The following table 3 shows this.

nr	SC		MC		sum of all connective items – <i>odkedy</i>	<i>odkedy</i>
	T	A	T	A		

others outscored by *odkedy*

4	PRS	IPFV	PRS	IPFV	240	2.74%	1619	10.10%
7	PRS	IPFV	PST	PFV	87	0.99%	1153	7.19%
9	PRS	IPFV	PST	IPFV	33	0.38%	801	5.00%

odkedy outscored by others

8	PST	PFV	PRS	PFV	634	7.23%	254	1.58%
---	-----	-----	-----	-----	-----	-------	-----	-------

Tab. 3. Quantitative data sourced from the SNK, processed in Excel

Apart from the anyway very low scoring *od tej doby, čo* and *od tej doby, ako*, the other items have a preponderance for especially type 5 (SC:PST.IPFV-MC:PST.PFV) and most, but notably not *od tých čias, čo*; *od tých čias, ako*; *od tej doby, čo*; *od tej doby, ako* also for type 8 (SC:PST.PFV-MC:PRS.PFV).

The choice of connective item seems to only be very slightly influenced or determined by the TA-constellation with only a few “preferences”. To date we have, pending our semantic analysis of the data, not conclusively established why these preferences occur and why the distribution of all connective items is not more even, especially that of *odkedy*.

3.2 Sum of all connective items: Slovak & Polish

We took the sum of the occurrences in our data set of each of the researched TA-constellations with SC+MC taken together and compared that data with similar data for Polish, the only language for which such data is at hand. As the purpose of this research was to establish line of further research, it seemed relevant to do so and learn what we could: significant differences would pinpoint areas for further research. That yielded table 4.⁹

nr	SC		MC		Slovak		Polish	
	T	A	T	A				
1	PST	PFV	PST	PFV	6456	26.04%	431	31.23%
2	PST	PFV	PST	IPFV	5493	22.15%	324	23.48%
3	PST	PFV	PRS	IPFV	4119	16.61%	241	17.46%
4	PRS	IPFV	PRS	IPFV	1859	7.50%	115	8.33%
5	PRS	IPFV	PST	PFV	1240	5.00%	52	3.77%
6	PST	IPFV	PST	IPFV	1431	5.77%	61	4.42%
7	PST	IPFV	PST	PFV	1615	6.51%	70	5.07%
8	PRS	IPFV	PST	IPFV	834	3.36%	51	3.70%
9	PST	IPFV	PRS	IPFV	659	2.66%	21	1.52%
10	PST	PFV	PRS	PFV	888	3.58%	12	0.87%
11	PRS	IPFV	PRS	PFV	168	0.68%	2	0.14%
12	PST	IPFV	PRS	PFV	35	0.14%	0	0.00%
totals:					24797	100%	1380	100%

Tab. 4. Sum of all TA-constellations Slovak and Polish [9]

It is rather satisfying – although not entirely unexpected – that Slovak and Polish display virtually the same percentages per TA-constellation. The deviations have been marked in the table by the outline and this concerns types 5, 6 and 7 but they are very slight indeed.

⁹ This data is sourced from [9], who also presented scores for SC:PRS.PFV items. The latter has been omitted in this table, which presents, then, a direct comparison of the TA-constellation types for the complex sentences for these two languages.

Of course, the Polish dataset was considerably smaller, but that should not deter us from suggesting that these two languages have a very similar treatment in terms of TA-constellation of the *sinds/since* construction.

3.3 Comparing SC and MC scores: Slovak & Polish

Next, we considered the TA-data per clause and so with the scores and percentages for all connective items taken together. In the following two tables 5 the scores for all SC have been summed regardless of the MC they are combined with. Also the tables display the scores for all MC regardless of their SC.

SC				MC			
T	A	score	perc.	T	A	score	perc.
PST	PFV	16956	68.38%				
PRS	IPFV	4101	16.54%				
PST	IPFV	3740	15.08%				
PRS	PFV	<i>omitted</i>	-				
				PST	PFV	9311	37.55%
				PST	IPFV	7758	31.29%
				PRS	IPFV	6637	26.77%
				PRS	PFV	1091	4.40%
control total:		24797	100%			24797	100%

Tab. 5a. SC resp. MC: Slovak

This particular Slovak data can at this time be compared to similar quantitative data for Polish, the only other language for which it is available even though the latter is based on considerably smaller quantities of samples. The figures provided by [9] have been adapted to the following table formatted for easy comparison.

SC				MC			
T	A	score	perc.	T	A	score	perc.
PST	PFV	1008	72.41%				
PRS	IPFV	220	15.80%				
PST	IPFV	152	10.92%				
PRS	PFV	12	0.86%				
				PST	PFV	553	39.73%
				PST	IPFV	438	31.47%
				PRS	IPFV	387	27.80%
				PRS	PFV	14	1.01%
control total:		1392	100%			1392	100%

Tab. 5b. SC resp. MC: Polish (cf. [9])

In spite of the difference in sample volumes as well as the fact that for Slovak SC:PRS.PFV has been omitted,¹⁰ we may at this time, however, signal that the scores of the types in percentages are very similar indeed.

4 CONCLUSIONS

As stated above in the introduction, the purpose of this paper is to set out further lines of research concerning Slovak correspondences of Dutch/English temporal *sinds/since* sentences: TA-constellations and connective items. We have utilized the *ASPAC* and the *SNK* corpora to compile a dataset and on the quantitative data thus at hand, we propose the following notes in respect to future research:

- Both in Polish and Slovak the types with PRS.PFV (be it in SC or MC) are rare. Although in these languages the *sinds/since* correspondences were found not to be restricted to past deixis, nevertheless it would seem that connective items such as those under discussion still have a prepondering use for past deixis. The PRS.PFV types need to be further researched to establish the extent of their usage as well as the functioning of the TA-constellations. NB. The particular “exemplary” use of PRS.PFV in (17) certainly needs further research and future deixis does not seem obvious.

This will aid formulating the very definition of the Slovak (and indeed Polish) connective items and their usage.

- The comparison of Slovak vs. Polish data shows that these two languages do not differ significantly in TA-constellations encountered for *sinds/since* complex sentences. There are a few points at which there are small differences and one might surmise that these are due to as yet not researched differences in the respective TA systems of these languages.
- The meaning types of the TA-constellations need to be established more precisely for Slovak (such as has been done e.g. for Polish in [9]).
- The preference(s) for particular TA-constellations by the individual connective items needs further research. An inroad into this will be the outlying scores in our table 2 and also 3. This will further establish the preferred meaning type(s) but also pinpoint the meaning and usage of the connective items.
- In our research to date on Slovak we have not considered register and clause order (SC-MC vs MC-SC) as factors in the selection of particular connective items. There are, however, some indications from closely related Czech set out by [5] that such matters are not without their significance in the choice of connective item. This still needs to be followed up for Slovak.

In future contributions the present authors intend to address some of these matters, especially the semantic analysis.

¹⁰ Although precise counts are not available at this time, it is clear that they are very slight indeed.

ABBREVIATIONS

3SG = third person singular; ACC = accusative; ASPAC see below under corpora; DAT = dative; IPFV = imperfective; M = masculine; MC = main clause; N = neuter; PFV = perfective; PST = past tense; REFL = reflexive pronoun; SC = secondary clause; SNK see below under corpora; TA = tense-aspect.

References

- [1] Barentsen, A. (2009). Taxis v niderlandskom jazyke. In *Typologija taksisnyx konstrukcij*, pages 269–366, V. S. Xrakovskij. Moskva, Znak.
- [2] Barentsen, A. (2016). Taxis in Dutch. In *Typology of Taxis Constructions (= LINCOM Studies in Theoretical Linguistics 58)*, pages 205–276, V. S. Xrakovskij. München, Lincom.
- [3] Barentsen, A. (2017). Nekotorye nabljudenija o niderlandskom i anglijskom sojuzax sinds/ since i ix sootvetstvijax v sovremennyx slavjanskix jazykax. In *Definitely Perfect: Festschrift for Janneke Kalsbeek (= Pegasus Oost-Europese Studies 29)*, pages 21–56. Eds. R. M. Genis, E. de Haard and R. Lučić. Amsterdam, Uitgeverij Pegasus.
- [4] Barentsen, A. (2018). O funkcionirovanii vidovremennyx form v složnyx predloženijax s sojuzom tipa s tex por kak / depuis que: na materiale slavjanskix i neslavjanskix jazykov. In *La relation temps/aspect: approches typologique et contrastive (Collection travaux et recherches UL3 / Éditions du Conseil Scientifique de l'Université de Lille 3)*, pages 153–157, T. Milliaressi. Lille, Université Charles-de-Gaulle – Lille3.
- [5] Duijkeren-Hrabová, M. van (2017). Het probleem van variatie bij Tsjechische voegwoorden met de betekenis ‘sinds’. In *Definitely Perfect: Festschrift for Janneke Kalsbeek (= Pegasus Oost-Europese Studies 29)*, pages 107–128. Eds. R. M. Genis, E. de Haard and R. Lučić. Amsterdam, Uitgeverij Pegasus.
- [6] Fijn van Draat, P. (1903). The loss of the prefix ge- in the modern English verb and some of its consequences – The conjunction since. *Englische Studien*, 32, pages 371–388.
- [7] Fryd, M. (2011). Since when is the Present Tense ruled out with SINCE! *Groninger Arbeiten zur germanistischen Linguistik*, 53(2), pages 89–103.
- [8] <https://ugp.rug.nl/GAGL/article/view/30529/27829>.
- [9] Genis, R. M. (2018). Towards a semantic model for conjunctions of ‘since’ and its interaction with verbal aspect: Polish material. In *La relation temps/aspect: approches typologique et contrastive (Collection travaux et recherches UL3 / Éditions du Conseil Scientifique de l'Université de Lille 3)*, pages 159–163, T. Milliaressi. Lille, Université Charles-de-Gaulle – Lille3.
- [10] Genis, R. M. (2019 to appear). Temporal ‘since’ in Polish: conjunctions & tense-aspect constellations. In *Dutch Contributions to the Sixteenth International Congress of Slavists – Linguistics: Belgrade, Belgrade, August 22–27, 2018 (= Studies in Slavic and General Linguistics 44)*, pages [to appear]. Leiden, Brill – Rodopi.
- [11] Heinämäki, O. (1978). *Semantics of English Temporal Connectives*. Indiana, Indiana University Linguistics Club.
- [12] Popović, Lj. (2012). Kontrastivna gramatika srpskog i ukrajinskog jezika: taksis i evidencijalnost. Beograd, Srpska akad. nauka i umetnosti.

- [13] Zorixina-Nil'sson, N. V. (2011). Semantika perfektnosti i dlitel'nosti i sposoby ee vyraženiya v vyskazyvaniyax s sojuzom s tex por kak (Iz sopostavitel'no-tipologičeskix nabljudenij). In *Sistemnye svjazi v grammatike i tekste. Materialy čtenij pamjati Ju. A. Pupykina*. 30 aprēlja 2010 g. M. D. Voejkova, A. Ju. Pupykina, M. Ju. Pupykina, pages 81–125. Sankt-Peterburg, Nestor – Istorija.
- [14] ASPAC = Amsterdam Slavic Parallel Aligned Corpus. Information on the corpus accessible at: <http://www.uva.nl/profiel/b/a/a.a.barentsen/a.a.barentsen.html>
- [15] SNK = Slovenský národný korpus – prim-8.0-public-sane. Bratislava, Jazykovedný ústav Ľ. Štúra SAV 2018. Accessible at: <http://korpus.juls.savba.sk>

IN WHICH CLAUSE DO SUBORDINATE CONJUNCTIONS PROSODICALLY BELONG?

ZUZANA KOMRSKOVÁ – PETRA POUKAROVÁ

Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague,
Czech Republic

KOMRSKOVÁ, Zuzana – POUKAROVÁ, Petra: In which clause do subordinate conjunctions prosodically belong? *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 216 – 224.

Abstract: This paper deals with the position of three Czech subordinating conjunctions *že* 'that', *když* 'when', and *až* 'when' within the prosodic word, using the phonetic annotation in the ORTOFON corpus. The position of subordinating conjunctions is traditionally described as initial within the subordinate clause, but the situation in spontaneous speech is not so clear. This paper shows the functional differences between the various positions within the prosodic word and presents the words which are most frequently combined with the selected conjunctions.

Keywords: conjunction, spontaneous spoken language, spoken corpus, prosody, prosodic word

1 THEORETICAL BACKGROUND

Grammar and stylistics usually describe spoken language in opposition to written language in a number of characteristics from all language levels. However, they usually omit the features which belong to sound qualities or which are in close relation to the acoustic channel. Therefore, it is highly valued to use in research such spoken data that are tone-aligned. Thanks to the spoken corpora containing authentic recordings with their transcripts, it is possible to combine research methods for both written and audio data and this data source offers wider possibilities and perspectives for the study of language.

The language–sound relation has been studied for a long time (see e.g. the correspondence between language and phonology units). Perhaps most attention has been paid to the connection between syntax and intonation (e.g. [1], [2], in Czech [3], [4])¹, especially to their similar function in communication, which is the ability to delimit the written and spoken texts. The difference between syntax and prosody lies in the perspective of unit description: syntax tends to employ synthesis (the

¹ The description of the role of syntax in spontaneous spoken texts is heavily discussed nowadays (cf. e.g. [5], in Czech [6], [7], but the suprasegmental level is in the spotlight.

composition of words in mutual syntactic relations to the higher structure within the hierarchy of language levels), while prosody uses analysis (the segmentation of speech continuum). The complementarity of both perspectives plays a functional role in communication, e.g. the distinguishing of the yes/no questions and the declarative sentences, or the functional sentence perspective (the topic structuring and the core of the sentence/utterance).

The important question concerning the delimitation of the written texts and the sound continuum is what the relationship between prosody and punctuation is. The studies dealing with this topic have mainly focused on the read texts (cf. [8], [9], for Czech e.g. [10]). It should be emphasized that punctuation marks, especially commas, could be perceived as a signal of pause. The awareness of this might affect structuring of speech, although speakers naturally group words into semantically coherent phrases indicated by timing and pitch cues; these prosodic phrase boundaries often coincide with major syntactic constituent boundaries but have a much flatter structure than syntax. Prosodic phrase boundaries tend to coincide with commas and semi-colons, but they also occur in other syntactically important places and thus they provide smaller (and potentially more useful) units for processing [11, p. 61], similarly [2, pp. 21, 39].

It was mentioned above that prosody is involved in the segmentation of the speech continuum. For this paper, the prosodic word is considered an elementary unit. It is defined as a group of words (or a single word) associated with one accent. This prosodic unit corresponds to the word in terms of written language. A higher prosodic unit is called tone unit, but the relation to the non-sound language unit is rather unclear. In general, it should be added, that one tone unit corresponds to the syntactically and semantically coherent structure.

Our paper presents an application of the spoken corpus in research on the border of two linguistic disciplines: phonetics and grammar. It focuses on the position of three Czech subordinating conjunctions *že* 'that', *když* 'when', and *až* 'when' within prosodic words. The ORTOFON corpus of spontaneous spoken Czech is used (for the description of the corpus see below).

Previous research about this topic gives contradictory results, based, however, on different data. From the syntactic point of view based on written language, the subordinating conjunctions usually occur in initial position within the subordinate clause (e.g. [12]) and they constitute a structural part of the sentence [13]. According to the orthographic rule, it is necessary to write a comma in front of the subordinating conjunctions in Czech. In accordance with the conclusions of the above-mentioned studies, it could be said that subordinating conjunctions are located at the beginning of the prosodic word.

By contrast, [2, p. 21] claims that the prosodic boundary occurs more often rather after than before a conjunction within spontaneous, unprepared speech. It could correspond to the theory of online syntax, introducing the term projectivity [5]

and offering the pragmatic point of view: the prosodic boundary occurring after a conjunction enables the speaker to signal her/his intent to continue the speech with some kind of new or known but rephrased information within the following subordinate clause. At the same time, the speaker indicates that s/he needs more time in order to think her/his next statement through.

This paper tries to find out which of above-mentioned points regarding prosodic words are relevant to spoken spontaneous Czech. It combines several perspectives mentioned above trying to cover the holistic and the most complete view on the prosodic words.

2 DATA AND METHODOLOGY

The selection of the data source is a key factor for further analysis and the expected results. Since we are interested in the occurrence of prosodic words within the most natural and least prepared spoken language, we chose the ORTOFON corpus [14], which covers spoken everyday communication among family members and friends (for more details see [15]). Moreover, it contains balanced data, and, secondly, it has a multi-tier transcription including not only the orthographic, but also the phonetic layer, containing both segmental and suprasegmental annotation and with the annotation of prosodic words within this phonetic transcription.

We searched for the conjunctions *že* ‘that’, *když* ‘when’, and *až* ‘when’ within the recordings with only two speakers. The first two conjunctions belong to the ten most frequent conjunctions in the ORTOFON corpus, whereas the last one is the 12th most frequent conjunction.² These conjunctions were selected according to their indisputable subordinating function.

We separately analyzed the conjunctions at the beginning, at the end, and in the middle of the prosodic word. The middle position means that the conjunction is surrounded on both sides by a different word. Each prosodic word consists of only words, all symbols stand out; therefore the pauses could be found only at the boundaries of the prosodic words. The conjunctions occurred most frequently as the first word within the prosodic word (see Table 1).

Position of conjunctions within the prosodic word	Frequency
beginning	18,346
middle	1,429
end	6,165

Tab. 1. Frequency of the positions of conjunction within the prosodic word

² The frequency of *že* is 21,163, *když* 5,173, and *až* 1,512.

The analysis was primarily focused on the prosodic characteristics of conjunctions. However, we did not analyze solitary words but words in their natural context, therefore it is appropriate and necessary to take into account at least the adjacent words within the prosodic word or within the adjacent prosodic word. As a result, the prosodic analysis, which was conducted manually, was extended to grammatical and semantic characteristics of all items within the prosodic word. These can be transferred into the following questions:

- Which position within the prosodic word does the conjunction occupy?
- Is the conjunction audibly stressed?
- Which parts of speech are included within the prosodic word?
- Is the whole prosodic word used as a collocation³?
- Do the conjunctions connect the main and the subordinate clauses?

3 RESULTS

This section summarizes the results of the whole analysis. It focuses on the position of the conjunction within the prosodic word and is divided into three subsections corresponding to the three possible positions. Some remarks concerning the collocations are added.

As was stated above, the researchers look at the position of the subordinated conjunction within the prosodic word from different points of view (syntactic or phonetic). We tried to describe which viewpoint is dominant and if the position could influence the function of the selected conjunctions.

3.1 Initial position within the prosodic word

As was stated above, the ORTOFON corpus is unambiguously dominated by cases where the subordinate conjunctions are at the beginning of the prosodic word (see Table 1). This means that the syntactic point of view – the conjunction as a structural part of the subordinate sentence – is already in agreement with the prosodic division of the sound continuum. Given that one of the most frequent left-sided collocations is a pause, the conjunction is not only the beginning of the prosodic word, but also the beginning of the higher unit, the tone unit.

The results of an analysis, based on random samples of 100 occurrences of conjunctions at the beginning of the prosodic word, could be gathered into three groups.

In the first group, the subordinate conjunction indicates the beginning of the subordinate sentence most frequently (54%). In this function the conjunction *že* ‘that’ occurs predominantly; this is confirmed (and de facto caused) by the most

³ Since this paper is not focused mainly on collocations, we use this term for such multi-word expressions that have a meaning as a whole unit.

frequent left-sided collocations, which are formed by verbs requiring a syntax argument in form of the content clauses (*verba dicendi*, *myslet* ‘think’, *vědět* ‘know’, *vidět* ‘see’, etc.). It is not unusual that *že* occurs even when the main sentence does not contain a verb that would justify its use (*tak jsem to hnedka volala mamce shodou okolností u toho telefonu byl i táta že to dala jako na hlasitej hovor* ‘so I just called my mom and by coincidence my dad was near the phone **that** she put on speakerphone’), or even in the case of the ellipsis of the main clause (it is typical for the reproduction of speech).

The second group could be marked as pragmatic. We classified here the examples, where the conjunction was used as a discourse marker (the collocation *že jo* ‘right’, 23%), and where the conjunction constitutes the first and the only part of the prosodic word as well (10%). In this case, the conjunction is used, when the speaker needs more time for thinking about the next statement (*oni pro to nemají ještě senzory jo že .. že . já to úplně cítím* ‘they don’t have sensors yet yeah **that** .. **that** . I feel it completely’); according to the theory of projectivity (see [1]), the speaker points out to his/her communication partner that he/she is going to continue speaking. Moreover, the repetition of *že* is related to the nature of spontaneous speech that occurs without previous preparation. The use of *že* can be regarded as the evidence not only of its semantic “emptiness” (we can use it when we need time for formulating our statement, and at the same time, we do not imply any relations by using it); but there is a kind of universality of this expression (we can use it anytime): *holčičky to tam poslouchají ne . že tam taky těch cédéček moc nemají* ‘the girls listen to it there . **that** they don’t have the CDs there much’ (the conjunction *že* does not belong to the verb contained in the main clause – *poslouchat* ‘listen to’ – because its syntactic argument is expressed by the pronoun *it*, on the contrary, it is possible to replace *že* with an expression with more appropriate meaning, here maybe with *i když* ‘although’).

To the third group of the occurrences belong the cases, in which the subordinate sentence is followed by the main clause (13%). In these cases the conjunctions *když* ‘when’ and *až* ‘when’ are definitely in dominance (77%). We suppose it could be caused by the syntactic-semantic characteristics of these words. *Až* and *když* have not only a syntactic function, but a semantic too, and thanks to that their consequence is clear. Compared to that, *že* stays on the borderline between the “introductory segment” and relevant information within the subordinate clause (analogy to functional sentence perspective is evident).

3.2 Middle position

The conjunctions occurring in the middle position are the least frequent (see Table 1). The subordinate conjunction, analyzed in this paper, is mainly placed at the second position within the prosodic word. Looking closer at the first position, we found another conjunction in 47% of all occurrences, then pronouns (17%) and adverbs (14%). Each of these parts of speech is mostly represented by one lemma,

which is quite opposite to the (mainly third) position after the selected conjunctions. The third position within the prosodic word is most frequently occupied by pronouns and adverbs as the first position, however, there are more verbs and higher number of various lexemes in contrast to the first position. It seems that the third position within the prosodic word is not as fixed by any part of speech as the first position is.

The most frequent conjunction at the first position (covering 68% of all conjunctions at this position) is *a* ‘and’, which is the most frequent conjunction in both spoken and written Czech, followed by conjunctions *jako* ‘like’ and *i* ‘also’. Both *jako* and *i* are most frequently connected with one conjunction (i.e. *jako že* ‘as though’, *i když* ‘though’), creating together a collocation. The second most predominant lexemes at the first position *tak* ‘so’ and *to* ‘it’ are used mainly in a collocation as well (see below).

3.2.1 The ten most frequent prosodic words

The analysis of prosodic words intersects with the analysis of collocations. The differentiation of a collocation from a non-collocation is conducted according to the meaning of the whole prosodic word; when there were more than 50% of the prosodic word occurrences used as a unit with one meaning (function), we called it collocation (see the part of the following paragraph in bold).

The ten most frequent prosodic words containing one of the selected conjunctions are the following: ***to víš že jo*** ‘you bet it is’, *a že to* ‘and that it’, *a že se* ‘and that’, ***víš že jo*** ‘you bet it is’, *a že tam* ‘and that there’, *a když to* ‘and when it’, *ne že by* ‘not that’, *a že by* ‘and that’, *jako že to* ‘as though it’, *tak že se* ‘so’. Only two prosodic words from the top ten were marked as a collocation, it is the same collocation actually.⁴ This collocation is used as an expression of agreement. The list contains two more collocations (*jako že* ‘as though’, *tak že* ‘so’), which do not cover the whole prosodic word, however. They both can be used as a collocation and non-collocation during speech and only the audible analysis could help to distinguish between them. We decided to analyze the collocation *tak že* ‘so that’; namely we were interested in finding out whether in this particular collocation the conjunction *že* is stressed or not (clitic). In 25 cases from 50 occurrences,⁵ the first word – *tak* – was stressed; we supposed that it could be caused by the similarity with the word *takže* ‘so’ with the similar meaning and could be written rather as a single word, i.e. non-collocation. In 22 cases⁶ *tak* was not stressed, it fulfilled the function of a particle (usually structuring text).

⁴ The shorter version appeared because the entire collocation was preceded by the word *no*. The word sequence *no to víš že jo* ‘but you bet it is’ is divided into two prosodic words: *no to* ‘but’ and *víš že jo* ‘you bet it is’. The boundary between the words *víš* and *že* usually coincides during their pronunciation.

⁵ This limit was chosen arbitrary.

⁶ There were mistakes in transcription and so we have analyzed only 47 occurrences.

The remaining prosodic words are mainly introduced by the conjunction *a* ‘and’. For comparison: we have analyzed the sample of 50 occurrences of the most frequent non-collocation *a že* ‘and that’. We found out that in 49 instances the first member of this phrase (*a*) is unstressed. It is in accordance with the Czech grammars, which describe this conjunction as a clitic. Finally, the prosodic word *ne že by* ‘not that sb’ is used as an opener for the contrast statement.

3.3 Final position

The final position of the subordinate conjunctions is much less frequent (see Table 1), but the usage of this conjunction is wider. It can be claimed that different conjunctions appear in different contexts and sometimes the label “conjunction” is relative. The very apparent case is *až* ‘when’. Located at the end of the prosodic word, it does not connect any sentences at all, but fulfills the function related to adverbs, or, if appropriate, particles expressing the meaning “extent” (it appears in 6% of all occurrences). The conjunction *když* ‘when’ in the collocations *co když* (‘what if’; e.g. *co když bude pršet?* ‘what if it rains?’) or *i když* ‘although’ evidence the similar behavior; it does not (syntactically) connect the verb of the main clause, but it primarily expresses the opposite meaning (often as a reaction to communication partner’s line). But there were only 6 occurrences in our sample. The phrases going hand in hand with syntactic “relativization” are also evident in the conception of the term *main* and *subordinate clause*: consider the collocations *s tím že* ‘with that’, *právě že* ‘just that’, *jako že* ‘like that’ (they appear in 9% of all occurrences). The conjunction *že* (and not the others) as a part of these phrases does not separate sentences into complex sentences but – rather – it structures the text, summarizes information or expresses the reaction to the communication partner:

*S1: to jsou klávesy jako nějaké speci**
 ‘these are like keys . some spe*’

S2: takové klávesy a a právě že hodně to frčí
 ‘such keys and and **just that** a lot of it’

In 9% the conjunctions have the function of a discourse marker (*všichni odešli že a já si tak sedím* ‘everybody had left **right** and I was sitting and’).

But the standard “conjunctive” use of the analyzed words appears most frequently, and so at 40% in the case of *the main clause – the subordinate clause* order and 15% in the reverse order (in the first case *že* clearly dominates, in the second one *když* occurs only).

The remaining occurrences (13%) are examples of the speaker not having a clear idea how to continue and/or to formulate his/her statement. In the nearest context pauses, unfinished words and hesitation are found (*abych se teda zeptala jak to vypadá že . proč mě* ‘so I asked how it looks that . why me’).

4 SUMMARY

The paper focused on the position of three Czech subordinate conjunctions in prosodic words in spontaneous spoken language. The analysis showed that, in accordance with syntax descriptions, subordinate conjunctions are most often located at the beginning of a prosodic word. The both initial and final position of the prosodic words demonstrate some common features for all three conjunctions as well as differences.

The most frequent use of these conjunctions as an attaching mean on the complex sentence level is common in both the positions. We figured out that while the conjunction *že* ‘that’ appears most often in the sequence of main clause – subordinate clause, the conjunctions *až* ‘when’ and *když* ‘when’ unambiguously prevail in the opposite order of the sequence. This observation should be confronted with the lexical meaning of each conjunction. The conjunction *že* is used in the main clauses, containing mainly the verbs *myslet* ‘think’ and *vědět* ‘know’, which could be seen as the introductory segment to the subordinate clause, including the largest piece of new information. On the contrary, the proportion of the amount of the new information included in both main and subordinate clauses with the conjunctions *když* and *až* is rather the same. The semantic difference between these two groups of conjunctions is also described in the dictionary of Czech [16]: *že* has not any specific lexical meaning, whereas *když* and *až* do. Therefore, *že* is closer to be called the discourse marker; not only the single-word conjunction, but also the collocation *že jo* ‘right’ (cf. [17]) and its expanded version *to víš že jo* ‘you bet it is’.

Considering the differences between the initial and final positions in the prosodic word, they are most pronounced in the case of *až*. The initial position is occupied by the conjunction *až*, however the part-of-speech-categorization changes at the final position to adverb or particle.

The middle position of the conjunction is interesting in terms of particular co-occurring words in the prosodic word. While the position following the conjunction is occupied by arbitrary parts of speech, the preceding position seems to be restricted, mainly to conjunctions, pronouns and adverbs. In this position, the collocations became also most evident. A further analysis shall focus on collocations separately.

ACKNOWLEDGMENTS

This study was written within the program Progres Q08 Czech National Corpus implemented at the Faculty of Arts, Charles University.

References

- [1] Selkirk, E. (1984). *Phonology and syntax*. Cambridge, MIT.
- [2] Wichmann, A. (2000). *Intonation in text and discourse. Beginnings, middles and ends*. London, Pearson Education Limited.
- [3] Daneš, F. (1957). *Intonace a věta ve spisovné češtině*. Praha, ČSAV.
- [4] Palková, Z. (2006). Textové dispozice pro členění na intonační fráze v češtině. In Palková, Z., and Janoušková, J., editors, *Kapitoly z fonetiky a fonologie slovanských jazyků*, pages 227–239.
- [5] Auer, P. (2008). On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences* 31, pages 1–13.
- [6] Müllerová, O. (1994). *Mluvený text a jeho syntaktická výstavba*. Praha, Academia.
- [7] Hoffmannová, J. et al. (2019): *Syntax mluvené češtiny*. Praha, Academia.
- [8] Chafe, W. C. (1982). Integration and involvement in speaking, writing, and oral literature. In Tannen, D. editor, *Spoken and written language: Exploring orality and literacy*, pages 35–53.
- [9] Chafe, W. C. (1988). Punctuation and the prosody of written language. Accessible at: <https://journals.sagepub.com>.
- [10] Janoušková, J. (2008). Shoda percepčního hodnocení hloubky prozodických předělů v závislosti na struktuře čteného textu. In Volín, J. – Janoušková, J., editors, *AUC Philologica 1, Phonetica Pragensia XI*, pages 87–104. Praha, Karolinum.
- [11] Ostendorf, M. et. al. (2008). Speech segmentation and spoken document processing. *Signal Processing Magazine* 25(3), pages 59–69.
- [12] Hrbáček, J. (1967). K poměru mezi spojovacími prostředky členskými a větnými (Podřadicí spojky v jednoduché větě), *Naše řeč* 50(3), pages 138–144.
- [13] Karlík, P. (2017): SPOJKA. In Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. Accessible at: czechency.org/slovník/SPOJKA.
- [14] Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2017): ORTOFON: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>.
- [15] Komrsková, Z., Kopřivová, M., Lukeš, D., Poukarová, P., and Goláňová, H. (2017): New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Jazykovedný časopis* 68(2), pages 219–228.
- [16] Filipec, J. et al. (2006). *Slovník spisovné češtiny pro školu a veřejnost*. Praha, Academia.
- [17] Komrsková, Z. (2017). What does že jo (and že ne) means in spoken dialogue. *Jazykovedný časopis* 68(2), pages 229–237.

RUSSIAN INDEFINITE PRONOUN *kakoj-libo*: NON-STANDARD USAGE AND CHANGES IN THE SEMANTICS¹

YULIA KUVSHINSKAYA

National Research University – High School of Economics, Moscow, Russia

KUVSHINSKAYA, Yulia: Russian indefinite pronoun *kakoj-libo*: non-standard usage and changes in the semantics. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 225 – 233.

Abstract: The paper deals with meaning and use of an indefinite pronoun *kakoj-libo* ‘any/some’ in the modern Russian language. Research based on corpus data revealed non-standard usage of the pronoun *kakoj-libo* ‘any/some’. The paper describes main types of the deviations and evaluates their pragmatic and semantic effect. Finally, tendencies of the change in semantics and use of these pronouns are characterized.

Keywords: Russian language, semantics, indefinite pronouns, nonstandard speech, corpus-based approach

1 INTRODUCTION

The description of the semantics of the indefinite pronouns is widely considered to be an extremely complicated task [1]. Which is why despite numerous researches, both typological [3] and focused specially on the Russian language ([4], [5]), we still lack understanding in this area.

The studies on the indefinite pronouns have described their semantics through distinguishing features [4] and defined formal and functional types [3]. In the linguists’ recent works a list of so called licensed contexts is used to describe these differences ([1], [6]).

This article is dealing with Russian pronouns of the *-libo* series. In special studies this series has hardly received close attention. In the dictionaries, meaning of such pronouns is usually clarified through references to the pronouns of other series [1]. As a result, the researchers know only a little about the peculiarities of their semantics or the rules of their contemporary usage.

Although the pronouns of the *-libo* series used to be defined as “bookish” ([4], [2]), in the contemporary Russian language they are widely used in media as well as in the spoken language. For instance, while during 1900–1990 according to the Russian National Corpus (RNC) the pronoun *kakoj-libo* ‘any/a’ was used far less

¹ The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics.

often than *kakoj-nibud* ‘any/some’ or *kakoj-to* ‘some/a sort of’ (the ratio for the early 50s is 106,6 ipm (*kakoj-libo*) to 230.40 ipm (*kakoj-nibud*) to 824.66 ipm (*kakoj-to*)), in the modern internet speech, according to RuTenTen, *kakoj-libo* prevails (89,4 ipm (*kakoj-libo*) to 59,42 ipm (*kakoj-nibud*) to 0,39 ipm (*kakoj-to*)). Besides, the mentioned changes go alongside with deviations from its standard use.

This article aims at describing the main types of the non-standard use of the pronoun *kakoj-libo* ‘any/a’.

The other pronouns of the *-libo* series (*kto-libo* – ‘anyone’, *chto-libo* – ‘anything’, *gde-libo* – ‘anywhere’ etc.) are not taken into consideration, because there are only a few instances of their both standard and non-standard usages in the evidence of the students’ texts.

While analyzing a non-standard use we cling to the concept that considers speech errors as a source of more profound knowledge of changes in language. Linguistics on the whole is very interested in a “negative” language evidence (see [8]), however only after the A. Frei’s work it turned into a special subject of study. Later the idea of “grammar of errors” resulted in a range of researches dealing with speech deviations, slips of the tongue etc. ([13], [14], [15], [16], [20]) and encouraged study of the SLA, works on the oral language and so on ([9], [10], [11], [12]).

The non-standard cases of use of the pronouns were examined on the evidence of two corpora: a corpus of Russian students’ texts (CORST, 54 instances) and the NRC (123 instances). The students’ texts have revealed deviations in use of the pronoun *kakoj-libo* ‘any/a’. Then we questioned their origins using the RNC’s data, especially for electronic communication, which means the most free speech, to find out if these evidence a global tendency or just mistakes.

Under deviations we understand a wide range of contexts, starting with examples of obvious violation of the rules and ending with examples that are grammatically acceptable, but demonstrate an imperfect communicative choice and illustrate the ongoing processes.

We defined instances with *kakoj-libo* ‘any/a’ as cases of deviation basing on the description of the licensed contexts ([6], [3]) and using an introspective method along with the results of a written interview of 15 adult native speakers of different occupation and age. An interview was carried out in two steps: 1) the interviewees appreciated the instances, which are mentioned in this article, mixed with the sentences, acting as fillers 2) the interviewees compared the instances from this article with the edited sentences, where the pronoun *kakoj-libo* ‘any/a’ had been substituted with another word (for an example see (9) or (10a, 11a)).

The interview revealed that even the native speakers do not fully comprehend the semantics of an indefinite pronoun, find the bookish style and the indefinite meaning unclear and objectionable, but preferable compared to the direct negation. Many (except professional editors and translators) consider non-standard usages of *kakoj-libo* ‘any/a’ agreeing with a noun in singular in an affirmative context

acceptable (examples 3, 7, 9 below). During the second step the native speakers mostly characterized instances with *kakoj-libo*, non-standard from our point of view, as inappropriate or unacceptable at all, preferring the edited examples.

2 NON-STANDARD USE OF THE INDEFINITE PRONOUN *kakoj-libo*

Now we will describe the main types of non-standard use of *k-l*, the semantic and pragmatic effects and probable causes of such usage.

2.1 The use in an unlicensed (affirmative etc.) context

(2) *Yego glavnaya ideya zaklyuchalas' v predlozhenii provodit' eksperimenty i issledovaniya, uznavaya informatsiyu o vozrastnom, professional'nom, territorial'nom i dr. statusakh respondenta* [18]. 'His main idea was an offer to carry out some kind of experiments and researches, finding out information about the age, professional, territorial and other state of the respondent'.

(3) *Takim obrazom, dlya togo chtoby zaimstvovaniya garmonichno sosushchestvovali s iskonnyimi slovami, prisushchimi kakomu-libo yazyku, neobkhodimo razvit' obshcheye prezreniye lyudey k neumestnomu, chrezmernomu upotrebleniyu* [18]. 'Thus, for borrowings to coexist harmoniously with the original words inherent in some language, it is necessary to develop people's general contempt for inappropriate, excessive use of foreign words'.

(4) *Tol'ko ne navyazyvat', a prosto iskrenne delit'sya toy pol'zoy, kotoruyu vy sami izvlekli iz kakikh-libo istochnikov.* [17]. 'Just do not impose, but just sincerely share the benefits that you yourself have gained from some sources'.

In these instances *k-l* is included in affirmative sentences. Besides, in (3, 4) the noun phrase refers to indefinite but real objects, and such use is not typical for *k-l*. There is no doubt that *k-l* makes these sentences more (probably too) formal. In the example (2) the pronoun can be removed without any loss in meaning, i.e. *k-l* is redundant; in the examples (3–4) we cannot remove the pronoun without replacing it with another quantifier with a specific meaning of a nonrandom, definitely real object of the set that can be chosen either by the speaker or by the actor (*opredelenny* 'definite', *konkretnyy* 'certain').

It seems that such usage occurs under the influence of special contexts found in dictionaries, instructions or legislative texts. In such texts *k-l* is used in the meaning of a variable, a gap that should be filled in certain situations:

(5) *Dno – Nizhnaya stenka, osnovaniye sosuda, sudna, kakogo-l. vmestilishcha.* [1]. 'Bottom – A wall below, a base of a vessel or some container'.

(6) *Litsa, uchastvuyushchiye v dele, ikh predstaviteli vprave khodataystvovat' ob oglashenii kakoj-libo chasti protokola...* [Grazhdanskiy protsessual'nyy kodeks Rossiyskoy Federatsii (2002)] [17]. 'Persons, who participate in the deal, and their representatives are allowed to demand announcing any part of the report'.

For such a use of *k-l*, both standard and non-standard, the idea of a potential plurality and variety of the objects and the idea of a choice are very important.

2.2 The use in a licensed context but without taking into account the semantics of the potential set

(7) *Prichinoy [strakha – YU.K.] mozhet byt' kakaya-libo tragediya ili bedstvennoye polozheniye.* [18]. 'The cause [for fear – Yu.K.] can be any tragedy or tribulation'.

(8) *Yesli vy vnosite kakoj-libo vklad v mirovuyu kul'turu, to on sokhranitsya v pamyati lyudey imenno blagodarya memam.* [18]. 'If you make any contribution to the world culture, it will remain in people's memory thanks to memes'.

These sentences are non-standard because the noun on which the pronoun depends refers to a class of objects, its meaning is generic and does not express multiplicity. In (8) *k-l* is used in an idiom, but the meaning of the idiom allows no diversity.

In such contexts *k-l* that defines a noun in singular with a generic meaning actually acts as an indefinite article. Paducheva notes that an indefinite pronoun phrase *tot ili inoj* can sometimes play the role of an indefinite article. This usage is standard for the phrase *tot ili inoj* [6] but not for *k-l*.

The instance (9) clearly demonstrates the deviation caused by misunderstanding both the quantifier nature of the pronoun *kakoj-libo* 'any/a' and the semantics of the set of similar objects:

(9) *Seychas my mozhem nablyudat' tendentsiyu k sozdaniyu osobogo molo-dezhnogo slenga, upotrebyayushchegosya isklyuchitel'no v kakom-libo gorode ili oblasti ...* [18]. 'Now we can observe a tendency to create a special youth slang, used exclusively in some city or a region'.

The phrase *isklyuchitel'no v kakom-libo* means the only possible object, but the appropriate meaning is "one undefined object". This meaning should be expressed by the phrase *v odnom kakom-libo*.

2.3 The use in negative but not expressive contexts instead of negative pronoun *nikakoj* 'no, none, any'

(10) *Yesli Vy chustvuyete chto ona otkryta i u neyo net kakikh-libo problem, to rasslabtes'.* [17], saved source spelling]. 'If you feel that she is open and does not have any problems, then relax'.

(11) *Zametim, chto Fransua Olland ne predstavlyayet kakikh-libo argumentov v pol'zu utverzhdeniya, chto ul'trapravyye ne dolzhny proyti vo vtoroy tur.* [18]. 'Let us note that Francois Hollande does not provide any arguments in favor of the assertion that the ultra-right should not enter the second round'.

(12) *K sozhaleniyu, diakhronicheskiy analiz semantiki vsej vyborki ne pokazal kakoj-libo dinamiki v izmeneniyakh semantiki – za isklyucheniyem pary-troyki*

primerov s nesnyatoy omonimiyey [18]. ‘Unfortunately, the diachronic analysis of the semantics of the entire sample did not show any dynamics in the development of semantics with the exception of a couple of examples with unmodified homonymy’.

The use in a negative context is normal for *k-l* [2]. The standard use of the pronoun we see in the examples (13–15).

(13) *Oblasti, zanyatyye slavyanami v rimskoye vremya (II-IV vv. n. e.), ne imeli kakikh-libo yestestvennykh rubezhey. Tuda neodnokratno vtorgalis' s zapada razlichnyye plemena germantsev* [17]. ‘The areas occupied by the Slavs in the Roman time (II–IV centuries), did not have any natural boundaries. Various tribes of Germans repeatedly invaded this area from the west’.

(14) *Terrorizm nel'zya assotsirovat' s kakoj-libo religiyey, etnicheskoy gruppoy ili geograficheskim rayonom, i u nego net nikakogo opravdaniya* [17]. ‘Terrorism cannot be associated with any religion, ethnic group or geographical area, and it has no excuse’.

We believe that the non-standard use of *k-l* in (10–12) can be explained by the fact that these examples refer to homogeneous objects, which mean no variety (10, 11), or to the same phenomenon (12). The homogeneity of the objects indicated by a noun phrase with *k-l* is important since the contexts imply opposition of these objects to the others. In (10) the psychological problems are opposed to the openness to communication, in (11) the arguments in favor of the statement are opposed to potential objections. As for *k-l*, it introduces the notion of diversity of similar objects that does not correspond with the idea of the sentence.

In contrast, in the standard sentences (13–14) an author speaks about a potentially diverse set of objects: about natural borders of different types (13), different religions, ethnic groups etc. (14). The meaning of diversity should be considered peculiar for *k-l* which is etymologically a pronoun of a free choice [3]. The non-standard sentences can be improved by introducing a restrictive definition to the noun phrase:

(10a) *Yesli Vy chustvuyete chto ona otkryta i u neyo net kakikh-libo cer'yeznykh problem, to rasslabtes'*. ‘If you feel that she is open and does not have any serious problems, then relax’.

(11a) *Zametim, chto Fransua Olland ne predstavlyayet kakikh-libo vesomykh argumentov v pol'zu utverzhdeniya, chto ul'trapravyye ne dolzhny proyti vo vtoroy tur.* ‘Let us note that François Hollande does not represent any strong arguments in favor of the assertion...’.

(12a) *K sozhaleniyu, diakhronicheskiy analiz semantiki vsey vyborki ne pokazal kakoj-libo zametnoy dinamiki v izmeneniyakh semantiki.* ‘...the diachronic analysis of the semantics of the entire sample did not show any notable dynamics in the development of semantics’.

Restrictive definitions designate a part of a set that has a certain feature, which is named by the definition. These definitions thereby oppose the selected subset to

the rest part of the set which do not have this feature. Thus, the noun phrase that includes *k-l* denotes the set of similar, but different objects.

As for the reasons for the use of *k-l* in these contexts, the main one is the speaker's intention to indicate the possibility that other objects exist. In other words, in example (10) by using *k-l* the author introduces an implication that there could be different problems; in (11) he/she implies that arguments were expected; in (12) the student regrets the lack of dynamics that could have been observed (this conclusion comes from the extra-language context of the example). Thus, it appears that the meaning of an object from a potentially existing set, intrinsic to *k-l*, is used here to indicate an unrealized opportunity.

At the same time, this deviation cannot be considered a rude violation of the norm, since the basic conditions for the use of pronouns are met. We consider such cases as examples of not the best communicative choice, as poor understanding of the semantics of the context and the word. Nevertheless, Paducheva believes that in many instances not the indefinite pronoun, but a negative pronoun is preferable [2].

2.4 The expressive use of the indefinite pronoun *kakoj-libo*

As Haspelmat shows, free choice and negative polarity pronouns can be used to express the meaning of the lower level of the pragmatic scale, in Fauconnier's terms [19; 3, p. 116–117]: “at least one X (not) exists ... that corresponds with...”.

Ward also speaks about the “emphatic negative function of *-libo* forms” [5, p. 466].

As the corpus data shows, *k-l* is indeed used in the emphatic function, expressing the meaning of the lower level of the pragmatic scale. But such expressive use of the pronouns is often accompanied by deviations from the standard:

(15) *Blagodarya takomu podkhodu, izuchayushchiye russkiy kak inostrannyi, dazhe pri otsutstvii kakogo-libo lingvisticheskogo obrazovaniya mogut bez osobykh usilyi chitat' nastoyashchiye slovarnyye stat'i* [18]. ‘Thanks to this approach, those, who learn Russian as foreign language, even without any linguistic education, can read real dictionary articles effortlessly’.

(16) *Da lyudey prosto tak sazhayut, bez kakikh-libo krazh.* [17]. ‘But people get imprisoned just so, without any thefts’.

In both examples (15–16) the meaning of the lower level of the pragmatic scale is applied to an inappropriate object. Indeed, in these contexts *lingvisticheskoye obrazovaniye* ‘the linguistic education’ (15), *krazha* ‘theft’ (16) are generic terms and do not allow scaling.

3 THE SEMANTIC FEATURES AND THE USAGE OF THE INDEFINITE PRONOUN *kakoj-libo*

Before defining the types of deviation in use of the pronoun *kakoj-libo* ‘any/a’; further *k-l*) we will try to describe its standard features.

3.1 The semantic features

E. V. Paducheva, summarizing results of her research [2], qualifies this pronoun as non-referent, i.e. the pronoun expresses reference to an object of the set that is not identified in reality (Veyrenc classifies these pronouns as virtual) [4].

(1) ... *YA ne raspolagala kakimi-libo dannymi o svoikh budushchikh sobesednikakh*. [17]. ‘I did not have any information about my future interlocutors’.

Following M. Haspelmath [3], she defines this pronoun as a quantifier that expresses an existential but not universal meaning, i.e. it distinguishes one of the objects of a set [2].

Paducheva argues in terms of the logical approach to semantics that a presumption of nonexistence of the set is necessary when using *k-l*. We understand this statement so that the noun phrase containing *k-l* refers to an indefinite object from a potentially existing set. For instance, the sentence (1) speaks about some objects, possibly in existence, but they are unknown to the speaker, so it is impossible to say if these objects truly exist.

Paducheva as well as Haspelmath considers *k-l* to be a pronoun with negative polarity. Haspelmath notes that “*-libo* indefinites were originally free-choice indefinites” and although they have lost this function, they still have the comparative function [3, p. 89].

The dictionary definition of this pronoun (through the pronoun *kakoj-nibud* ‘any/some’ also shows that *k-l* as well as *kakoj-nibud* means a choice of one element from a range of similar ones: “*I. Tot ili inoy, lyuboy iz ryada podobnykh*” ‘One or another, any of a number of similar’ [1].

3.2 The licensed contexts

Paducheva identifies several typical contexts that make the use of Russian negative polarity pronouns including *kakoj-nibud* ‘any/some’ possible [2].

Negative contexts:

- direct negation,
- contexts with a negative verb (*lishit* ‘deprive’, *otsutstvovat* ‘be absent’), preposition (*bez* ‘without’, *protiv* ‘against’).

Non-affirmative contexts:

- conditional clause
- interrogative clause
- comparative construction
- verbal participle, especially with modality or repeating action

Paducheva points out that Russian pronouns of negative polarity are not licensed in the context with universal quantification as well as under a phrasal stress [2].

4 THE DISTRIBUTION OF THE TYPES OF DEVIATIONS IN SPEECH OF THE STUDENTS AND THE ADULT NATIVE SPEAKERS

The non-standard type of use of the indefinite pronoun *kakoj-libo* 'any/a' in which the pronoun acts as an indefinite article in the context that expresses generic quantification is frequent in students' speech, but rarely occurs in the online speech produced by the adult native speakers.

The expressive use of *k-l* is more common in online speech and is not very frequent in students' written language. Obviously, the latter is due to the bookish, often scientific genres of these texts which are not supposed to be expressive.

5 CONCLUSION

Thus, we have considered the main types of deviations from the standard use of the pronoun *kakoj-libo* 'any/a' based on the corpus data representing the students' written language and the Internet speech of the adult native speakers.

The deviations in the use of *k-l* seem to be caused by:

1) The lack of understanding of the quantifier nature of a pronoun. The authors of the examples confuse the values of a discrete set and the values of a class, existential and universal quantification.

2) An intention to make the speech more official (in some contexts).

3) An intention to make the speech more expressive (in other contexts).

At the same time, it seems that these deviations came from the intention to express meanings that are either untypical for the pronoun in standard speech or limited to special contexts:

- A gap that is potentially filled in when a statement is applied to a specific situation (a consequence of the etymological meaning of free choice).

- The diversity of the elements of the set. It seems that this is also a consequence of the etymological significance of free choice. In such contexts *k-l* is synonymous to the adjective *diferent*. This meaning enables the speaker to implement a variety of pragmatic strategies.

The analysis of the non-standard use of the pronoun *kakoj-libo* 'any/a' reveals, on the one hand, difficulty in implementing the categories and strategies of written speech (quantifier semantics, formal style, expressiveness), and on the other hand, the ongoing search for new ways to express meaning in the modern language, including ideas of diversity, quantification and expressiveness.

References

- [1] Evgenieva A. P. (ed.). Slovar' russkogo yazyka: V 4-kh tomakh. (1999). 4th edition. Moscow, Russkiy yazyk; Poligrafresursy.

- [2] Paduceva, E. V. (2015). Mestoimeniya otritsatel'noy polyarnosti. In *Russkaya korpusnaya grammatika*. Accessible at: <http://rusgram.ru/index>
- [3] Haspelmath, M. (1997). *Indefinite Pronouns*. Oxford, Clarendon Press.
- [4] Veyrenc J. (1964). Kto-nibud'et kto-libo formes concurrentes? *Revue des études Slaves*, 40, pages 224–233.
- [5] Ward, D. (1977). On Indefinite Pronouns in Russian. *The Slavonic and East European Review*, 55(4), pages 444–469.
- [6] Paduceva, E. V. (2017). Oborot tot ili inoy. In *Russkaya korpusnaya grammatika*. Accessible at: <http://rusgram.ru/index>
- [7] Efremova, T. F. (2000). *New dictionary of the Russian language. Explanatory and word-formative. V 2 tomakh*. Moscow, Russkiy yazyk.
- [8] Shcherba L.V. (1931). O troyakom aspekte yazykovykh yavleniy i ob eksperimente v yazyke soznaniya. *Pamyati uchitelya I. A. Boduena de Kurtene. Izvestiya AN SSSR*, 1, pages 113–129. Accessible at: <http://www.ruthenia.ru/apr/textes/sherba/sherba3.htm>
- [9] Vakhtin, N., Mustajoki, A., and Protassova, E. (2010). Russkie jazyki. In *Instrumentarium of linguistics: Sociolinguistic approaches to non-standard Russian*. Helsinki, Yliopistopaino (Slavica Helsingiensia; vol. 40). pages 5–16.
- [10] Zemskaja, E. A. (ed.). (2001). *Jazyk ruskogo zarubežja: obščie process i rečevye portrety*. Vena.
- [11] Bogdanova-Beglaryan N. V. (2015). Agressivnyy uzus – ili evolyutsiya yazykovoy normy? *Verkhnevolzhskiy filologicheskiy vestnik*, pages 25–31.
- [12] Lapteva, O. A. (1990). *Živaja ruskaja reč s teleekrana*. Seged.
- [13] Frei H. (2006). *Grammatika ošibok*. Moscow, Komkniga.
- [14] Rusakova, M. V. (2013). *Elementy antropotsentricheskoy grammatiki ruskogo yazyka*. Moskow, Yazyki slavyanskikh kul'tur.
- [15] Rahilina, E. V. 2014. *Grammatika ošibok: v poiskah konstant*. In *Jazyk. Konstany. Peremennye*, pages 87–95. Cankt-Peterburg. Aleteyya.
- [16] Zevakhina, N. A., and Dzhakupova, S. S. (2015). Corpus of Russian student texts: design and prospects. In *Materialy 21-y Mezhdunarodnoy konferentsii po komp'yuternoy lingvistike "Dialog"*. Moscow.
- [17] *Russian National Corpus*. Accessible at: <http://www.ruscorpora.ru/index.html>.
- [18] *Corpus of Russian student texts*. Accessible at: http://web-corpora.net/learner_corpus.
- [19] Fauconnier, G. (1975). Pragmatic scales and logical structures. *Linguistic Inquiry* 6, pages 353–375.
- [20] Fromkin V. (1973). *Speech errors as linguistic evidence*. The Hague, Mouton.

WAYS OF AUTOMATIC IDENTIFICATION OF WORDS BELONGING TO SEMANTIC FIELD

VICTOR ZAKHAROV
Saint-Petersburg State University, Russia

ZAKHAROV, Victor: Ways of automatic identification of words belonging to semantic field. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 234 – 243.

Abstract: The paper presents results of the ongoing research on creation of the semantic field of the “empire” concept. A semantic field is a collection of content units covering a certain area of human experience and forming a relatively autonomous microsystem with one or several centers. Relations in such microsystems are also called associations. The idea is to extract from data on syntagmatic collocability a set of lexical units connected by systemic paradigmatic relations of various types and strength using distributional analysis techniques. The first goal of the study is to develop methodology to fill a semantic field with lexical units on the basis of morphologically tagged corpora. We were using the Sketch Engine corpus system that implements the method of distributional statistical analysis. Text material is represented by our own corpora in the domain of “empire”. In the course of the work we have acquired lists of items filling the semantic space around the concept of “empire”.

Keywords: semantic field, concept of empire, distributive and statistical analysis, corpus, thesaurus

1 INTRODUCTION

Many automated text processing systems are based on dictionaries, including semantic ones. Semantic fields can also be regarded as semantically simple dictionaries. The notion of “semantic field” is used in linguistics to denote a set of language units with a common semantic feature with some common meaning component. Words (both common and proper nouns) and word groups constitute such lexical units. To cite O. S. Akhmanova, “A field is a set of content units covering a certain field of human experience and forming a relatively autonomous microsystem” [2].

First attempts to identify semantic fields were undertaken when conceptual dictionaries, or thesauri (for example, Roget’s Thesaurus) were created. As construed by V. G. Admoni, a field is characterized by an inventory of elements related through systemic relations. V. G. Admoni discovers a central part in a field, the nucleus, whose elements have a full set of features defining this group, and the periphery whose elements has some of the features characteristic of the field, but can have features peculiar to neighboring fields [1]. A field consists of a continuity of links between objects of a set,

and the links in some areas are particularly dense. In such cases, we speak of lexico-semantic groups or elementary microfield grouping words that are, as a rule, of the same part of speech. In a general case, blurred boundaries between minifields and parts of speech are characteristic of a field. Many papers deal with the semantic field theory ([4], [18], [21], etc.).

The task of modelling a conceptual or terminological system has been a subject of computational linguistics for a long time. It can be divided into two parts, namely: identification of lexical identifiers of concepts and identification of relations between them. In this study, we are interested in the first task. It can be solved “manually” through explication and formalization of professional knowledge accumulated in the course of human activities. However, since our knowledge of the world is to a lesser or greater extent reflected in texts, then a task of extraction of a system of concepts from texts can be set.

In this paper, we deal with the semantic field “empire”. The paper is a part of a greater research dedicated to comparative analysis of the content of this field in Russian, English, Czech and German. The choice of languages can be explained by the fact that the concept of “empire” in these languages is closely related to the historic memory of the people and that it is “alive” in the linguistic consciousness of the native speakers.

2 RESEARCH METHOD

The research method is a corpus-oriented analysis of the paradigmatics and syntagmatics of lexical units using distributional statistical methods. In our study, we used existing tool and corpus linguistic processors and the data from the ad hoc annotated corpora.

One of the earliest and best-known linguistic research methods is distributional statistical analysis where information on the distribution of text elements and their numerical parameters are used. There is an idea that paradigmatic links which belong to language system can be derived from the syntagmatic ones, i.e. neighborhood of words in a linear text chain (see [3], [14], [16], [17]). The principle of transfer from studying textual (syntagmatic) links to systemic (paradigmatic) ones underlies various distributional statistical methods. The assumption is that two elements are linked by paradigmatic relation if both of them are textually systematically linked with some other third element. However, for it to be possible to speak of regularities in any statistical distribution, large bodies of data and, consequently, massive computational power are needed.

3 MECHANISM OF CREATION OF LEXICO-SEMANTIC GROUPS AND FIELDS

As was already mentioned, paradigmatic links can be derived from syntagmatic ones. This concept was suggested by A. Ya. Shaykevich [17] and K. S. Jones [6] as

early as in the 1960s. The mathematical apparatus for computing this similarity was developed later by D. Lin [13]. But this approach has not been implemented until now though it has become possible to create a large database of co-occurrences of lexical units on the basis of text corpora and to “compute” a set of “the closest neighbors” for every word on the basis of such database (see [11], [19]).

However, it is also important to take into account a syntactic relation between contextually close elements [5]. In the Sketch Engine system ([8], [9], [10], [12]), which was used to build corpora and identify syntagmatic and paradigmatic links, the concept of lexico-syntactic patterns is implemented as so-called *word sketches*, an automatically generated summary of the co-occurrences restricted to the set of syntactic formulas. Word sketches are based on sets of rules describing grammatical relations between words in a text (word sketch grammar).

When building a corpus on the basis of morphologically tagged data, a special database consisting of triplets of lexico-grammatical relations is built. Statistical processing of this database computes the data for a distributional thesaurus which, for us, is similar to a lexico-semantic group for a specific term. The algorithm to compute semantic distance between the elements in the group (candidates) is described in [20, sect. 3, 4]. The similarity in distribution of words is calculated statistically on the basis of an association measure logDice [15] taking into account the grammar of lexico-syntactic patterns [11]. In our research, the distributional statistical analysis is based on the grammar of lexico-syntactic patterns for the Russian language developed by M.V. Khokhlova [7].

4 STUDY MATERIAL AND TOOLS

The main study material is an ad hoc corpus (10.25 mln tokens). It has been built from texts relating to the topic of empire in the Russian literature and culture of the end of the 18th and the beginning of the 20th century collected within the scientific project of the Institute of the Russian literature, St. Petersburg. Words’ meanings are subject to continuous change, thus we have to take into account the temporal dimension and the diachronic nature of words. That is why we divided the corpus into 4 subcorpora on a chronological principle: the 18th century (the corpus identifier –XVIII), the 1st half of the 19th century (XIX-1), the 2nd half of the 19th century (XIX-2) and the 20th century (XX). The boundary dates of the subcorpora were chosen as some sort of milestones in the perception of the “empire” concept in the development of the Russian social thinking.

As has been mentioned already, we have used the Sketch Engine system for purposes of our research. Its main feature is availability of special tools implementing the distributional statistical analysis method – *Thesaurus* (building a lexico-semantic group for a specific term; see Fig. 1), *Clustering* (grouping thesaurus units into narrower clusters), and *Collocations* (extraction of steady word combinations, collocations).

империя ^(noun)
 XIX-1 freq = 397 (139.16 per million)

Lemma	Score	Freq
держава	0.143	96
император	0.141	373
государство	0.135	823
церковь	0.129	1,308
европа	0.129	797
христианство	0.127	336
рим	0.125	330
религия	0.121	193
мир	0.120	1,223
просвещение	0.116	740
правительство	0.111	709
монархия	0.109	61
единство	0.108	254
франция	0.107	405
истина	0.103	703
земля	0.102	843
философия	0.102	454
предание	0.101	173
образованность	0.101	335
литература	0.100	494
восток	0.100	240

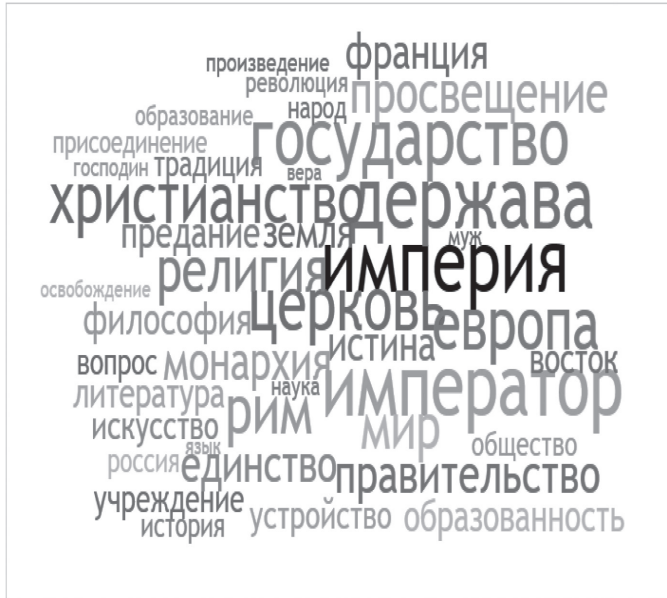


Fig. 1. A fragment of the distributional thesaurus for the word “empire” on the basis of the subcorpus of the 1st half of the 19th century with semantic link strength value (score)

Translation of terms in Fig. 1: держава ‘state, power’, император ‘emperor’, государство ‘state’, церковь ‘church’, Европа ‘Europe’, христианство ‘Christianity’, Рим ‘Rome’, религия ‘religion’, мир ‘world’, просвещение ‘enlightenment’, правительство ‘government’, монархия ‘monarchy’, единство ‘unity’, Франция ‘France’, истина ‘truth’, земля ‘land’, философия ‘philosophy’, предание ‘legend’, образованность ‘education’, литература ‘literature’, восток ‘East’.

A considerable portion of terms in any subject area are, as a rule, represented by word combinations. The tool calculating the strength of syntagmatic links between lexical units is Collocations. It computes the association of the units in a linear sequence based on 7 association measures. It should be added that this tool identifies not only syntagmatic links, but also paradigmatic ones if the “window” for collocates is sufficiently large.

5 CORPUS-BASED ANALYSIS OF PARADIGMATIC AND SYNTAGMATIC RELATIONS

5.1 Research technique

The Russian and Czech corpora were used for the research. A technique for creating a semantic field was developed. The essence of the technique is given below.

Two approaches to identification of lexical units presumably belonging to the semantic field “empire” were used, namely: creation of a distributional thesaurus, creation of a list of collocations. These methods are implemented in the four subcorpora mentioned above. The thesauri terms derived using each subcorpus (minithesauri) are ranked by the association score (see Fig. 1) and are grouped into a summary array. The number of words in each minithesaurus in Thesaurus tool was limited to 40. As a result, it counts 160 term occurrences. But it important how stable an appropriate word is represented throughout the corpus. The “stability factor” k is assigned to each lexical unit (term or word group) We set $k = 1, 2, 3$ or 4 depending on the number of minithesauri that include an appropriate word. Ultimately, the average rank is calculated for all the units of the summary array as well as the normalized rank that represents the semantic association of a relevant lexeme with the head word “empire”, i.e. it is the “appropriateness” factor for this semantic field (Table 1). The normalized rank is derived by multiplication of the average rank and the “rank normalization factor”. The following factors were chosen empirically: 1 for the terms occurring in all the four minithesauri ($k=4$), 1.5 for $k=3$, and 2 for $k=2$. The terms extracted from one subcorpus only are not included in the field. Thus, these factors reduce the ranks of the terms related to the word “empire” in larger number of subcorpora (i.e. in more time periods).

<i>Subcorpus</i>	<i>Rank in the subcorpus thesaurus</i>	<i>Lemma</i>	<i>Score</i>	<i>Freq</i>	<i>Stability factor</i>	<i>Average rank</i>	<i>Norm. rank</i>
XIX-2	36.	англия ‘England’	0.131	1055	2	29	58
XVIII	22.	англия ‘England’	0.095	148	2		
XIX-2	19.	армия ‘army’	0.149	478	1		
.....
XIX-2	24.	государственность ‘statehood’	0.143	201	2	19	38
XX	14.	государственность ‘statehood’	0.141	143	2		
XX	1.	государство ‘state’	0.245	1016	4	2.25	2.25
XIX-2	2.	государство ‘state’	0.200	4240	4		
XVIII	3.	государство ‘state’	0.184	766	4		
XIX-1	3.	государство ‘state’	0.135	823	4		
XX	2.	гуманизм ‘humanism’	0.188	195	1		

<i>Subcorpus</i>	<i>Rank in the subcorpus thesaurus</i>	<i>Lemma</i>	<i>Score</i>	<i>Freq</i>	<i>Stability factor</i>	<i>Average rank</i>	<i>Norm. rank</i>
XVIII	2.	держава ‘state, power’	0.189	424	3	4.3	6.45
XIX-2	10.	держава ‘state, power’	0.165	606	3		
XIX-1	1.	держава ‘state, power’	0.143	96	3		
.....
XIX-1	13.	единство ‘unity’	0.108	254	1		
XIX-2	5.	император ‘emperor’	0.184	1381	3	4	8
XX	5.	император ‘emperor’	0.177	295	3		
XIX-1	2.	император ‘emperor’	0.141	373	3		
XX	8.	империализм ‘imperialism’	0.166	297	1		
.....

Tab. 1. Summary distributional thesaurus for the word “empire” (a fragment)

5.2 Preliminary results

As stated above, summary distributional thesaurus included 160 entries. 79 words occur once, i.e. only in one of the minithesauri, and the distribution of these unique words in the subcorpora is as follows: subcorpus XVIII: 32 words, subcorpus XIX-1: 16, subcorpus XIX-2: 14, subcorpus XX: 17.

The remaining 81 entries which consisted of 33 different words occur in 2, 3 or 4 minithesauri. We call these 33 words the nucleus of the semantic field. Their distribution in subcorpora is as follows: 8 in the 18th century, 24 in the 1st half of the 19th century, 26 in the 2nd half of the 19th century, and 23 in the 20th century.

Here is the list of the nucleus of the semantic field “empire” derived through the implementation of the proposed technique, ranked by different bases (the beginning of the list is given):

a) by the normalized rank:

государство ‘state’, *император* ‘emperor’, *держава* ‘state’, *Европа* ‘Europe’, *царство* ‘kingdom’, *церковь* ‘church’, *Рим* ‘Rome’, *Франция* ‘France’, *христианство* ‘Christianity’, *монархия* ‘monarchy’, *правительство* ‘government’, *страна* ‘country’, *общество* ‘society’, *нация* ‘nation’, *Россия* ‘Russia’, *государственность* ‘statehood’...

b) by the average semantic association factor (score):

держава ‘state, power), *государство* ‘state’, *общество* ‘society’, *союз* ‘union’, *государственность* ‘statehood’, *нация* ‘nation’, *император* ‘emperor’, *политика* ‘politics’, *культура* ‘culture’, *страна* ‘country’, *община* ‘community’, *церковь* ‘church’, *царство* ‘kingdom’, *христианство* ‘Christianity’, *религия* ‘religion’, *мир* ‘world’, *просвещение* ‘enlightment’, *правительство* ‘government’, *монархия* ‘monarchy’...

c) alphabetically:

Англия ‘England’, *государственность* ‘statehood’, *государство* ‘state’, *держава* ‘state, power’, *Европа* ‘Europe’, *император* ‘emperor’, *искусство* ‘art’, *история* ‘history’, *культура* ‘culture’, *литература* ‘literature’, *мир* ‘world’, *монархия* ‘monarchy’, *наука* ‘science’, *нация* ‘nation’, *общество* ‘society’, *община* ‘community’, *политика* ‘politics’, *правительство* ‘government’...

d) by the relative frequency (ipm, instances per million):

Россия ‘Russia’, *общество* ‘society’, *церковь* ‘church’, *мир* ‘world’, *история* ‘history’, *государство* ‘state’, *наука* ‘science’, *просвещение* ‘enlightenment’, *правительство* ‘government’, *держава* ‘state’, *power*, *политика* ‘politics’, *царство* ‘kingdom’, *литература* ‘literature’, *революция* ‘revolution’, *союз* ‘union’, *страна* ‘country’, *Европа* ‘Europe’, *община* ‘community’, *культура* ‘culture’, *император* ‘emperor’...

Let’s give a list of bigram collocations that are candidates to the semantic field “empire” (ranked according to the log-Dice measure).

Total 115 bigrams were extracted, with the majority of them being bigrams including a form of the word *империя* ‘empire’ plus some other word: Adj+*империя*, *империя*+N (genitive), N+*империи* (genitive) where Adj states for adjective, N for noun.

24 bigrams are the nucleus of the syntagmatic collocations according to the normalized rank: *Российская империя* ‘Russian Empire’, *Византийская империя* ‘Byzantine Empire’, *Восточная империя* ‘Eastern Empire’, *Священная империя* ‘Holy Empire’, *падение империи* ‘fall of the empire’, *Австрийская империя* ‘Austrian Empire’, *Великая империя* ‘Great Empire’, *пределы империи* ‘borders of the empire’, *Турецкая империя* ‘Turkish Empire’, *столица империи* ‘capital of the empire’, *etc.*

The experiments with the Czech language were conducted using the synchronous National Corpus of the Czech Language. The collocation search was conducted using the corpus SYN2015 (https://kontext.korpus.cz/first_form?corpname=syn2015). However, the Nosketch Engine system supporting Czech corpora lacks the Thesaurus tool. This is why a corpus on our topic has been created from the Czech Internet (342 mln tokens). To avoid peripheral vocabulary in the resulting field, the output of the distributional thesaurus was limited to 30.

Here are several examples for the Czech language for the word *říše* (empire) (the beginning of the lists is given):

a) by the normalized rank:

království ‘kingdom’, *civilizace* ‘civilization’, *Británie* ‘Britain’, *Rusko* ‘Russia’, *společenství* ‘community’, *vesmír* ‘space’, *Řím* ‘Rome’, *impérium* ‘empire’...

b) by the semantic association factor (score measure in the distributional thesaurus):

civilizace ‘civilization’, *království* ‘kingdom’, *země* ‘country’, *impérium* ‘empire’, *Británie* ‘Britain’, *Amerika* ‘America’, *armáda* ‘army’, *lidstvo* ‘mankind’, *monarchie* ‘monarchy’...

The beginning of the collocation list is as follows:

Třetí říše ‘Third empire’, *Římská říše* ‘Roman empire’, *Osmanská říše* ‘Ottoman empire’, *Německá říše* ‘German empire’, *Svatá říše* ‘Holy empire’, *Velkomoravská říše* ‘Great Moravian empire’, *vládce říše* ‘ruler of the empire’, *zánik říše* ‘fall of the empire’...

Anyone who is familiar with Czech culture will agree that these terms do have a strong semantic link with the word *říše* ‘empire’.

5.3 Conclusion and further work

As we can see, the use of a text corpus and “smart” corpus tools allow to automatically extract syntagmatic and paradigmatic relations and create rather reasonable content of a term system. Moreover, obtained lists significantly expand the existing lexicographic guides. However, the question is where are the limits of the field “empire”. We see in the periphery of the summary distributional thesaurus such words as *посол* ‘ambassador’, *отечество* ‘Motherland’, *воевода* ‘battlemaster’, *воин* ‘soldier’, etc. that hardly belong to the field “empire”. This encourages us to repeat the experiments with “stricter” parameters of corpus tools. At the same time the work to identify elements semantically related to terms included in the nucleus of the field “empire” will be conducted, i.e. a task to create second-stage thesauri (minifields) and form a single list made, if possible, as a semantic network.

One can note that the concept “empire” in Russian had different connotations during different periods in the Russian culture defined by different parameters. For example, significant specific feature of the 18th century texts catches the attention. This is evident in the contents of the vocabulary – see Section 4.2: 32 words of 79 words “unique” for only one period belong to the 18th century and only 8 words from the field nucleus are listed in 18th century minithesaurus. In general, it can be carefully concluded that despite the fact that the empire existed *de facto* in the 18th century, the very concept of the empire did not form in the Russian culture at the time.

It is interesting to get interpretation of the results linked to historical or cultural aspects concerning different languages. That is why we have started selecting texts for a parallel English-Russian, Czech-Russian and English-Czech corpora. It appears that the volume of the corpora will be small due to the difficulties in selecting parallel texts, but we think that it is important to do that because the elements of the field will be in the same temporal and historical paradigm in these texts. It is also interesting to see which words (and why) will prevail in translation of the same concept: for example, the Czech *říše* can be translated into Russian as *империя* ‘empire’, *королевство* ‘kingdom’, *царская власть* ‘reign’, *рейх* ‘Reich’. The Russian *империя* is translated

into Czech as *impérium* ‘empire’, *říše* ‘empire’, *císařství* ‘empire’, *država* ‘domain’, etc. The same goes for other terms and other language pairs.

Finally, it can be stated that the task of building one small semantic field reflects the peculiarities of the lexico-semantic system of a language as well as opportunities and barriers in automation of semantic processing.

ACKNOWLEDGMENTS

This work was implemented with the financial support of the Russian Foundation for Basic Research, Project No. 18-012-00474 «Semantic field “empire” in Russian, English and Czech» and partly with the Project No. 17-04-00552 “Parametric modeling of the lexical system of the modern Russian literary language”.

References

- [1] Admoni, V. G. (1973). Syntax of modern German: The system of the relations and the system of construction, [Sintaksis sovremennogo nemeckogo jazyka: Sistema otnoshenij i sistema postroenija], Leningrad.
- [2] Akhmanova, O. S. (1966). Dictionary of Linguistic Terminology [Slovar' lingvisticheskikh terminov]. Moscow.
- [3] Arapov, M. V. (1964). Some principles of creation of the “thesaurus” dictionary NTI Serie 2(4), pages 40–46.
- [4] Askoldov, S. A. (1980). Concept and word, [Koncept i slovo]. Moscow.
- [5] Gamallo, P., Gasperin, C., Augustini, A., and Lopes, G. P. (2001). Syntactic-Based Methods for Measuring Word Similarity, In Text, Speech and Dialogue: Fourth International Conference TSD–2001. LNAI 2166, pages 116–125. Springer-Verlag.
- [6] Jones, K.S. (1965). Experiments in semantic classification, Mechanical Translation and Computational Linguistics, 8(3–4), pages 97–112.
- [7] Khokhlova, M.V. (2010). Development of the grammatical module of Russian for the specialized system of processing of corpus data [Razrabotka grammaticheskogo modulja ruskogo jazyka dlja specializirovannoj sistemy obrabotki korpusnyh dannyh], Bulletin of St. Petersburg State University [Vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta], Series 9, Philology, oriental studies, journalism, 2(9), pages 162–169.
- [8] Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus, In Proceedings of the 13th EURALEX International Congress. Spain, July 2008, pages 425–432. EURALEX.
- [9] Kilgarriff, A., Rychlý, P., Jakubiček, M., Rundell, M. et al.: Sketch Engine [Computer Software and Information Resource]. Accessible at: <http://www.sketchengine.co.uk>.
- [10] Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwel, D. (2004), The Sketch Engine, In Proceedings of the XIth Euralex International Congress, pages 105–116. Lorient, Université de Bretagne-Sud.
- [11] Kilgarriff, A., and Rychlý, P. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments), In Proceedings of the 45th Annual Meeting of the ACL. Interactive Poster and Demonstration Sessions. Czech Republic, June 2007, pages 41–44. ACL.

- [12] Kilgarriff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pages 7–36.
- [13] Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proc. COLING-ACL*, pages 768–774. Montreal.
- [14] Pekar, V. (2004). Linguistic Preprocessing for Distributional Classification of Words. In *Proceedings of the COLING-04 Workshop on Enhancing and Using Electronic Dictionaries*, pages 15–21, Geneva.
- [15] Rychlý, P. (2008). A lexicographer-friendly association score, In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9. Brno.
- [16] Shaykevich, A. Ya. (1982). Distributive and statistical analysis of texts [Distributivno-statisticheskij analiz tekstov], PhD thesis. Leningrad.
- [17] Shaykevich, A.Ya. (1963). Distribution of words in the text and allocation of semantic fields [Raspredelenie slov v tekste i vydelenie semanticheskikh polej], In *Foreign languages in higher education*, 2, pages 14–26, Moscow.
- [18] Shchur, G.S. (1974). Field theory in linguistics, [Teoriya polja v lingvistike], Moscow-Leningrad.
- [19] Smrž, P., and Rychlý, P. (2001). Finding Semantically Related Words in Large Corpora, In *Text, Speech and Dialogue: Fourth International Conference (TSD-2001)*, LNAI 2166, pages 108–115. Springer-Verlag.
- [20] Statistics Used in Sketch Engine. Accessible at: <https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine>.
- [21] Wierzbicka, A. (2001). Understanding of cultures through keywords, [Ponimanie kul'tur cherez posredstvo kljuchevyh slov]. Moscow.

ANALYSIS OF THE LEMMA MATEŘSTVÍ (MOTHERHOOD)

ZUZANA ČERNÁ¹ – RADEK ČECH²

^{1,2} Faculty of Arts, University of Ostrava, Czech Republic

ČERNÁ, Zuzana – ČECH, Radek: Analysis of the lemma *Mateřství* (Motherhood). *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 244 – 253.

Abstract: The paper presents results of analysis of the lemma *mateřství* ‘motherhood’. The authors applied methods of corpus linguistics and discourse analysis – the corpus assisted discourse studies approach – in order to survey representations of the lemma in Czech journalistic texts published from 2010 to 2014, sorted the results into discourse categories on the basis of collocation and concordance analysis, and found out that chief referential discourse-of-motherhood categories were surrogate motherhood, relationship of motherhood and career, delight from motherhood, family relationships, financial and time aspects of motherhood, changes brought by motherhood, and active motherhood. Surrogate motherhood was presented as a solution to women who cannot have a baby themselves, nevertheless also as a complicated issue, in which case emphasis was put on relevant legislation. Motherhood was presented as a danger for a woman’s career, however also as a source of joy, an essential relationship within a family, a right for financial support from the state, a life change, an activity, and an entity closely connected to time factors.

Keywords: motherhood, discourse, mass media, corpus, CADS

1 INTRODUCTION

Motherhood is specific – each of us has a mother, some of us are mothers, motherhood has been discussed in various spheres of humanities (cf. family psychology, family sociology) or pictured in multiple works of art (randomly Whistler’s *Arrangement in Grey and Black No. 1*, Michelangelo’s sculptural group *Madonna, Dvořák’s Stabat Mater*).

Models of motherhood have changed in Czech society in recent decades. What was a dominant norm, i.e. mother taking care both of children and work and father going to work, but not much involved in the daily care of children, is not viewed as the only model any more. Women endeavor to have both children and a job, men participate in the care of their children and sometimes stay on parental leave, families lead dual-career lives and employers offer flexible working hours and part-time jobs to help parents balance work and life. On the other hand, there is a growing number of couples who decide to postpone starting a family or not to have it at all, birth-rate is insufficient and new forms of expecting a child (artificial insemination or surrogate motherhood) have appeared.

We inquired how the phenomenon of motherhood was reflected in Czech journalistic texts, which are generally considered to be a powerful tool to influence public opinion. In order to survey what leading issues related to motherhood they present and what discourse of motherhood they produce, we implemented the corpus assisted discourse studies approach (CADS), being inspired mainly by Baker [1] and [2]. We surveyed representations of the lemma *mateřství* ‘motherhood’ in a corpus of journalistic texts available in the Czech Corpus of Contemporary Written Czech SYN2015 [3]; we observed the frequency of the lemma with regard to a medium, journal title, text type and genre, made its collocation analysis, and classified the collocates into discourse categories; we searched for repeated discourse patterns that the lemma appeared in by means of concordance analysis and attempted to interpret them.

The research submitted is a partial probe within a broader project aimed at surveying Czech mass media discourse of motherhood. Analysis of lemmas such as *rodičovství* ‘parenthood’, *těhotenství* ‘pregnancy’ or *porod* ‘childbirth’, which emerged from the research by way of neural networks (Kubát, Hůla [4]) as close to the lemma *mateřství* (motherhood), as well as analyses of data from older and spoken corpora, will follow.

2 THEORETICAL BACKGROUND

The term *discourse* has been widely used in humanities, however with various meanings: for example, Stubbs [5] defines it as a language above a sentence or clause, Fairclough [6] as an element of social life which is closely interconnected with other elements and effects power relations in the society, Blommaert [7] as a language in action which forms objects and produces a particular version of events. Baker [1] understands it as a practice which systematically forms objects of which it speaks; there are many discourses, each with a different story to tell about the world. We follow mainly Blommaert’s and Baker’s concept in the paper.

Discourse of motherhood has been studied with various aims, from multiple perspectives and by way of various methods and techniques. Feminism, which states that mothers are oppressed and constrained to home environment and care for family members by dominant patriarchal discourses of motherhood and ideal womanhood (Buchanan [8], Rich, McClatchy [9], Vincent [10] and others) and that women are stressed due to the discrepancy between those discourses and their own view of motherhood and womanhood (Haratyan [11], Butler [12], Parker [13], and others), belongs to the most influential research-on-motherhood streams.

Considering the *discourse of motherhood in media*, e.g. Woodward [14] and Tomešová [15] are worth mentioning: Woodward searched the emergence of a figure of New Mother, independent and paid for her work, in women’s magazines; Tomešová found out that the image of a pregnant woman is constructed solely from the perspective of maternity, ignoring others components of her life such as sex or job.

This paper of ours aims at surveying the issue of motherhood presentation in Czech journalistic texts. Unlike the aforementioned researchers, we made use of a corpus-driven approach to discourse in order to describe how motherhood is treated in the corpus used.

3 METHODOLOGY AND MATERIAL

We made use of the qualitative-quantitative strategy of the CADS, considering – in compliance with Baker [2] – both parts equal. Corpus linguistics helps researchers minimize the interpretation bias and make use of large real life data sets; discourse analysis enables them interpret what is beyond the numbers.

First, we surveyed the frequency of the lemma *mateřství* ‘motherhood’ from different aspects (medium; journal title; text type; genre) and made its collocation analysis by means of the T-score association measure. We decided on the T-score with regard to the fact that we were looking for typical collocations of the lemma and, according to Brezina [16], T-score is highly suitable for detection of frequent non-exclusive collocations.

We set the collocation span L5–P5 (i.e. 5 positions to the left, 5 to the right) from the KWIC (key word in context), the minimum frequency of a collocate in the subcorpus at 3 and the minimum frequency of a collocate in context also at 3.

We treated autosemantic collocates only in our follow-up analysis. We determined the main discourse categories on the basis of the T-score results and on the contextual meaning of individual collocates. This step was implemented manually. Subsequently, by way of concordance analysis, we searched for repeated discourse patterns that motherhood appeared in.

As mentioned above, we analyzed journalistic texts from the Czech Corpus of Contemporary Written Czech SYN2015 [3], namely the text group NMG: journalism (NMG: publicistika) of approximately 33 mil. tokens published between 2010 and 2014. It consists of NEW: traditional journalism (NEW: tradiční publicistika) and LEI: leisure time journalism (LEI: volnočasová publicistika). We made use of the KonText application.

4 RESULTS

The lemma *mateřství* ‘motherhood’ appeared 236 times in the aforementioned subcorpus, with the value of instances per million words (hereinafter i. p. m.) of 5.94, which indicates that motherhood was not a moving topic (lemmas with this frequency have a rank laying in the interval $r = \langle 11086; 11126 \rangle$ in rank-frequency distribution of lemmas in SYN2015, the corpus contains 363 819 individual lemmas).

Compared with the NFC: oborová literatura (NFC: expert literature; 5.89 i. p. m.) and FIC: beletrie (FIC: fiction; 3.63 i. p. m.) subcorpora, the i. p. m. was not considerably higher in NMG: publicistika (NMG: journalism; 5.94 i. p. m.).

The average reduced frequency (hereinafter ARF) of the lemma *mateřství* ‘motherhood’ in the NMG subcorpus was 52.26, indicating that the lemma was distributed slightly unevenly. According to Křen and Cvrček [3], the value of ARF for a certain expression is its corrected frequency, based on the distribution of its occurrence in a corpus: the closer the ARF is to frequency, the more even the distribution is; expressions that occur in a single cluster within a corpus, reach the ARF value of 1.

We found out that the lemma was more than twice as frequent in magazines (5.85 i. p. m.) as in newspapers (2.77 i. p. m.). In compliance with our expectations, the lemma occurred most frequently in the genre of lifestyle (29.38 i. p. m.), social life (13.57 i. p. m.), and society (10.25 i. p. m.). It appeared mostly in women’s magazines – the highest i. p. m. of 148.15 being in *Maminka* (Mummy), which was not surprising. The i. p. m. in newspapers ranked towards the end of the imaginary ladder (*Hospodářské noviny* 1.7 i. p. m., *Mladá fronta Dnes* 1.35 i. p. m.); for details, see Table 1.

	Journal title	i. p. m.
1	Maminka (Mummy)	148.15
2	Marie Claire	72.49
3	Playboy	55.52
4	Rytmus života (Life Rhythm)	55.19
5	Svět ženy (A Woman's World)	54.41
6	Esprit	53.21
7	Ona dnes (Her Today)	48.24
8	Žena a život (Woman and Life)	39.13
9	Story	32.13
10	Katka (Kate)	32.12
11	Elle	30.91
	...	9.39
54	Hospodářské noviny (Economic News)	1.70
55	Mladá fronta Dnes (Young Frontline Today)	1.35
	...	
58	Sport	0.50

Tab. 1. Frequency of the lemma *mateřství* in individual title

These numbers might result from the fact that the issue of motherhood is naturally closer to women, especially to mothers, than men and that there is an absence of stimulating issues related to motherhood which newspapers could be interested in.

Based on T-score, the lemma made strong collocations with autosemantic lemmas presented in Table 2. It is remarkable that the first 16 autosemantic collocates covered 50% of all autosemantic tokens modified by *mateřství* ‘motherhood’.

	Lemma	T-score		Lemma	T-score
1	náhradní (surrogate)	4.90	34	plodnost (fertility)	1.73
2	žena (woman)	4.44	35	porodit (deliver)	1.73
3	kariéra (career)	4.24	36	odkládat (postpone)	1.73
4	užívat (enjoy ¹)	3.60	37	legální (legal)	1.73
5	dítě (child)	3.26	38	upravovat (treat)	1.73
6	matka (mother)	3.14	39	zvládat (manage)	1.73
7	život (life)	3.10	40	upozorňovat (warn)	1.73
8	peněžitý (financial)	3.00	41	užít (enjoy ²)	1.73
9	pomoc (help)	2.81	42	mateřský (maternal)	1.73
10	změnit (change)	2.63	43	zkrátka (simply)	1.72
11	hodně (a lot)	2.55	44	přinášet (bring)	1.72
12	velký (big)	2.51	45	povinnost (duty)	1.72
13	role (role)	2.43	46	skvělý (great)	1.72
14	teď (now)	2.38	47	tradiční (traditional)	1.72
15	nárok (right)	2.23	48	Česko (Czechia)	1.71
16	aktivní (active)	2.23	49	řešit (solve)	1.71
17	spojený (connected)	2.22	50	manžel (husband)	1.71
18	práce (job)	2.17	51	láska (love)	1.71
19	doba (period)	2.14	52	rozhodnout (decide)	1.69
20	hnutí (movement)	1.99	53	nyní (now)	1.69
21	pozdní (late)	1.99	54	rodina (family)	1.69
22	radost (joy)	1.99	55	snažit (strive)	1.68
23	téma (topic)	1.98	56	mluvit (speak)	1.68
24	zákon (law)	1.97	57	poslední (last)	1.65
25	pracovat (work)	1.97	58	muž (man)	1.65

¹ Imperfectum.

² Perfectum.

	Lemma	T-score		Lemma	T-score
26	věc (thing)	1.94	59	čas (time)	1.65
27	rok (year)	1.92	60	začít (begin)	1.62
28	myslet (think)	1.92	61	druhý (second)	1.61
29	český (Czech)	1.91	62	jiný (other)	1.59
30	dobrý (good)	1.89	63	celý (whole)	1.58
31	surogátní (substitutive)	1.73	64	dva (two)	1.57
32	odsouvání (postponing)	1.73	65	jít (go)	1.55
33	skloubit (harmonize)	1.73			

Tab. 2. Autosemantic collocates to the lemma *mateřství*

In compliance with Baker [1] and [2], we conducted a discourse analysis, determining the chief **discourse-of-motherhood categories** on the basis of the aforementioned T-score results and the contextual rather than the dictionary meaning of individual collocates (see Table 3). This step was implemented manually and its results are therefore subject to the researchers' decision.

Discourse category	Representing lemmas
surrogacy	surrogate, law, Czechia, legal, treat
career	career, woman, job, work, harmonize, manage
joy	joy, enjoy
family relations	child, mother, life, role, husband, family, love, man
finances	financial, help, right
changes	change
activity	active, movement
time	period, late, year, postponing, delaying, now, last, time

Tab. 3. Discourse-of-motherhood categories

Through a follow-up manual analysis of all concordance lines, we found out what emerged from the discourse categories:

Surrogacy: Motherhood and surrogacy are closely related; their collocation was the strongest one of all that emerged on the basis of T-score. *Náhradní* 'surrogate' was never used with anything else than *mateřství* 'motherhood'. The discourse of surrogacy was concentrated mainly round the collocates *náhradní* 'surrogate', *zákon* 'law', *Česko* 'Czechia', *legální* 'legal', *upravovat* 'treat' – it discussed mostly the legality / illegality of surrogacy in the Czech Republic, or in other countries. Surrogacy is perceived as a help to women who cannot have a baby themselves: *Náhradní matka se nechá oplodnit spermatem muže, pokud je jeho žena neplodná.*

(‘A surrogate mother undergoes artificial insemination when an ovum of a woman who cannot have a baby herself is placed into her womb.’) What is viewed as problematic is the immoral business with surrogacy and the unclear status of the biological and the “target” mother because *komplikace mohou nastat, když se náhradní matka po porodu dítěte nevzdá. Podle zákona je totiž matkou žena, která dítě porodí, nikoli dárkyně vajíčka!* (‘complications may turn up when a surrogate mother does not hand over the child. In compliance with the Act, mother is a woman who gives birth, not the donor of the ovum!’)

Career: This discourse category was formed mainly round the collocates *kariéra* ‘career’, *žena* ‘woman’, *práce* ‘job’, *pracovat* ‘work’, *skloubit* ‘harmonize’, and *zvládat* ‘manage’ and it states that motherhood endangers a woman’s *career* and women are to make a pro-motherhood/pro-career choice, others postpone motherhood to later life periods, and others somehow try to balance both. Mothers experience discrimination in the labour market because *muži se méně starají o domácnost a děti. Ženy kvůli mateřství odsunují kariéru, nevezmou náročnější, tudíž lépe placenou, pozici.* (‘men take care of household and children less. Women postpone their career because of motherhood, they do not take a more demanding, therefore a better paid, job.’) The category of time could be well included in this category because time was often referred to in connection with postponed motherhood but as it fits into the category of time as well, we decided to discuss it there.

Joy: Motherhood was perceived as delight in women-mothers, judging from the frequently used expressions *radost* ‘joy’ and *užívat, užít* ‘enjoy’. The collocation *užívat / užít si mateřství* ‘enjoy motherhood’ appeared 15 times, *joy* made collocations with *motherhood* only 4 times (out of which 3 times it went together with *tíha / strast / povinnost mateřství* – ‘burden / duties of motherhood’). The collocate *skvělý* ‘great’ could also be included in this category as it expresses a high degree of pleasure related to motherhood.

Family relations: Motherhood is set into a context of family relations; this discourse category was formed chiefly round the collocates *dítě* ‘child’, *matka* ‘mother’, *život* ‘life’, *role* ‘role’, *manžel* ‘husband’, *rodina* ‘family’, *láska* ‘love’, *manželé* ‘husband / man and wife’³), that is those generally considered typical for nuclear family positions, roles and relationships. Motherhood is presented as a significant component (directly a role) of a woman’s life, of the relationships within a family and of a family as a whole. Although surrogacy and employment of women-mothers might also be included in this category, they were already discussed above and are therefore omitted here.

Finances: This discourse category was formed chiefly round the collocates *peněžitý* ‘financial’, *pomoc* ‘help’, and *nárok* ‘right’. It implies that motherhood is

³ Czech *manžel* corresponds with English *husband*, *manželé* with *man and wife*. With the collocate *manžel*, both meanings are included.

financially demanding and the financial help provided to mothers by the state to partially counterbalance their lower income, is their right. *Podmínkou toho, aby žena mohla pobírat peněžitou pomoc v mateřství, je, že nesmí mít žádný jiný příjem.* ('To be granted a motherhood benefit, women must not have any other income.') Motherhood gives women a right to help: *Ano, máte nárok na peněžitou pomoc v mateřství.* ('Yes, you have a right to financial help in motherhood.') It was interesting that the collocate *pomoc* 'help' was related to the financial aspect of motherhood only.

Changes: The discourse of change emerged mainly round the collocate *změna, změnit* '(to) change'. Motherhood means a change, it has a potential to change women: *V čem Vás mateřství změnilo?* ('What has motherhood changed in you?')

Activity: Motherhood is active; activity was discussed in relation to the Movement for Active Motherhood, the aim of which was to make conditions for delivering women better, giving them voice to implement their own needs and wishes when raising a child. The discourse of activity was concentrated mainly round the collocates *aktivní* 'active' and *hnutí* 'movement': *Před jedenácti lety založila Hnutí za aktivní mateřství, aby se zlepšily podmínky ve zdejších porodnicích.* ('She established the Movement for Active Motherhood to improve conditions in local maternity hospitals.')

Time: Time factors play an important role in motherhood; the discourse of time emerged mainly round the collocates *dobu* 'period', *pozdní* 'late', *rok* 'year', *ted'* 'now', *odsouvání* 'postponing', *odkládat* 'postpone', *poslední* 'last' and *čas* 'time'. It referred mostly to postponed motherhood – *odsouvání mateřství na pozdější dobu* 'postponing motherhood to a later age'; *Patříte k ženám, které mateřství odkládaly kvůli kariéře?* ('Are you among women who postpone motherhood for career?') or to recent changes – *Pozdní mateřství je in. Poslední dobou si potomky "na stará kolena" pořizuje stále více celebrit.* ('Late motherhood is in. In recent years, celebrities have been having offspring at a later age.');

Ted' je kult mateřství jiný a často až extrémní. ('Nowadays, the cult of motherhood is different, often even extreme.')

Late motherhood brings benefits (woman's maturity) as well as risks – *nebezpečí neplodnosti* 'threat of infertility'.

5 SUMMARY

In order to survey what leading issues related to motherhood Czech journalistic texts present and what discourse of motherhood they produce, we decided to analyze the lemma *mateřství* 'motherhood' in them. We applied the CADS approach in compliance with Baker and commented on the frequency of the lemma with regard to a medium, journal title, text type and genre; we made its collocation analysis on the basis of the T-score association measure, and classified the collocates into discourse categories. Afterwards, we searched for repeated discourse patterns that the lemma appeared in by means of a concordance analysis.

We found out that the lemma was more frequent in magazines than in newspapers, most common in the genre of lifestyle, social life and society, and in terms of magazines, it occurred mostly in women's magazines. The chief discourse-of-motherhood categories were: surrogate motherhood, relationship between motherhood and career, delight from motherhood, family relationships, financial and time aspects of motherhood, changes due to motherhood, and active motherhood. Surrogate motherhood was presented as a solution for women who cannot have a baby themselves, nevertheless also as a complicated issue due to the unclear status of the biological and "target" mother; emphasis was put on relevant legislation. Motherhood was further presented as a danger for a woman's career, as a result of which women were forced to make a pro-motherhood / pro-career choice, postpone motherhood to later life periods, or somehow try to balance both. Mothers experience discrimination in the labor market as it is mostly them who bear the burden of childcare. On the other hand, motherhood was perceived also as a source of joy, an essential relationship in a family, as a right to financial support, as a change in a woman's life, as an activity, and as an entity closely connected to time factors.

Being aware of research limits and challenges, e.g. lack of possibility to compare the discourse of individual titles, need for data from older corpora, need for analyzing lemmas corresponding with motherhood, and need for analyzing spoken language, we are going to treat them in our further research.

ACKNOWLEDGMENTS

This study was supported by grant No. SGS02/FF/2019.

References

- [1] Baker, P. (2006). *Using Corpora in Discourse Analysis*. London, Continuum.
- [2] Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T., and Wodak, R. (2008). *A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the UK*. *Discourse and Society*, 19(3), pages 273–306.
- [3] Křen, M., Cvrček, V. et al. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2522–2528, Portorož. ELRA.
- [4] Kubát, M., Hůla, J. et al. (2018). *The Lexical Context in a Style Analysis: A Word Embeddings Approach*. *Corpus Linguistics and Linguistic Theory*. Accessible at: <https://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2018-0003/cllt-2018-0003.xml>.
- [5] Stubbs, M. (1983). *Discourse Analysis: The Sociolinguistic Analysis of Natural Language*. Chicago, IL: The University of Chicago Press.

- [6] Fairclough, N. (2004). *Analysing Discourse. Textual Analysis for Social Research*. New York, London, Routledge.
- [7] Blommaert, J. (2005). *Discourse: A Critical Introduction*. Cambridge, Cambridge University Press.
- [8] Buchanan, L. (2013). *Rhetorics of Motherhood*. Carbondale, Southern Illinois University Press.
- [9] Rich, A. C., and McClatchy, J. D. (2013). *Adrienne Rich*. Columbia, University of Missouri.
- [10] Vincent, C. (2010). *The Sociology of Mothering*. In *The Routledge International Handbook of the Sociology of Education*, pages 109–120, New York, London, Routledge.
- [11] Haratyan, F. (2012). *Contradictory Discourses of Motherhood as Institution and Experience*. *Southeast Asian Review of English*, 51(1), pages 40–47.
- [12] Butler, J. (2010). *Performative Agency*. *Journal of Economy*, 3 (2), pages 147–161.
- [13] Parker, R. (1997). *The Production and Purposes of Maternal Ambivalence*. In *Mothering and Ambivalence*, pages 17–36, Blackwell, Psychology Press.
- [14] Woodward, K. (1994). *Discourses of Motherhood in Women's Magazines in Contemporary Britain*. Doctoral thesis. Sheffield Hallam University.
- [15] Tomešová, K. (2008). *Mediální obraz těhotenství ve vybraných časopisech pro rodinu [Media Discourse of Pregnancy in Chosen Family Magazines]*. Bachelor work. Brno, MU Brno.
- [16] Brezina, V. (2018). *Statistics in Corpus Linguistics. A Practical Guide*. Cambridge, Cambridge University Press.

CORPUS-SUPPORTED SEMANTIC STUDIES: PART/WHOLE EXPRESSIONS IN RUSSIAN

IGOR BOGUSLAVSKY^{1,2} – LEONID IOMDIN¹

¹Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

²Technical University of Madrid, Madrid, Spain

BOGUSLAVSKY, Igor – IOMDIN, Leonid: Corpus-supported semantic studies: Part/Whole expressions in Russian. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 254 – 266.

Abstract: We investigate valency properties of partials – words and constructions that express the Part/Whole relation, primarily in Russian, offering new observations largely based on the Russian National Corpus. Special attention is given to such lexical units as *bol'sinstvo* 'majority', *men'sinstvo* 'minority', *čast'* 'part', *prosent* 'percentages', *v bol'sinstve svoem* 'in its <their, etc.>majority', *po bol'sej časti* 'for the most part', etc.

Keywords: corpus, semantics, valency, part-whole

1 INTRODUCTORY REMARKS

Various aspects of words and constructions that express the Part/Whole relation (henceforth referred to as partial expressions or partials) have attracted much attention. Yet one extremely complicated aspect of these expressions, namely their valency properties, was not investigated sufficiently until recently. Several years ago, the authors of this paper introduced important concepts and proposed solutions for some of the challenging issues (see Boguslavsky [1], [2], [3]). In this paper, we will supplement the findings of these papers, and offer a number of new observations, largely based on the material of the Russian National Corpus (RNC, www.ruscorpora.ru).

The paper is arranged as follows. Section 2 is focused on the valency structure of the most typical Russian nominal partials – *часть* 'part', *большинство* 'majority', *меньшинство* 'minority' and *процент* 'per cent/percentage'. We start with the regular means of valency instantiations and proceed to non-canonical cases. In Section 3, we present the argument properties of a few other partials. Then, in Section 4, we go on to characterize certain microsyntactic constructions belonging to the class of partials.

2 PROTOTYPICAL PARTIALS

2.1 Valency instantiation options

If we view the syntactic representation of a sentence as a dependency structure, we can state that an argument A of a predicate P may occupy one of the three logically possible positions with regard to P: (a) P directly subordinates A; (b) P directly depends on A; (c) A and P are not related to each other by any direct dependence. In the (a) case, we will say that the respective valency of P is filled by A by an active pattern, in case (b) we have a passive valency pattern, while case (c) displays a discontinuous pattern. The valencies of verbs are mostly filled actively, the valencies of nouns may follow an active or, to a lesser extent, a discontinuous pattern, while adjectival and adverbial predicates accept their valencies primarily in the passive manner, even though active or discontinuous patterns are also possible ([4]).

2.2 The main valencies of partials

The minimal argument structure of partials consists of two arguments: Whole and Part. In the most typical case, the valency of Whole is instantiated actively – by an NP in the genitive or a prepositional phrase formed by prepositions ИЗ ‘from’, ОТ ‘from, of’, or, less frequently, СРЕДИ ‘among’, while the valency of Part remains unfilled, cf. *часть (половина, 30%) урожая* ‘part (a half, 30%) of crops’, *часть из нас* ‘a part of us’, *10% от суммы* ‘10 percent of the sum’.

Less prototypical filling options for these valencies are illustrated by cases when, in addition to the valency of Whole of a partial, its valency of Part is also filled, via a support verb that represents a value of a lexical function (in the sense of Igor Mel’čuk’s Meaning-Text theory, see [9]), namely, Func_i, Oper_i or Labor_{ij}, or through a copula. In these cases, taken from RNC, we deal with the discontinuous pattern of valency filling:

(1) *Ряд улиц, пересекающихся под прямым углом, образовывал [Oper₁(часть)], так сказать, старинную часть города* ‘**A number of streets crossing at right angles** formed, so to say, the ancient part of the city’.

In (1), the noun phrase printed in bold fills the valency of Part of the noun *часть*. Similarly, boldface words instantiate the valency of Part of partials in sentences (1–5).

(2) *Женщины составляют [Oper₁(процент)] сегодня 57 процентов общей численности этой возрастной группы* ‘**The women** make up 57 per cent of the totality of this age group today’.

(3) *Вратарь – [zero copula] это половина команды* ‘**The goalkeeper** is half of the team.’

(4) *Лютеране составляют [Oper₁(меньшинство)] меньшинство среди всех протестантов* ‘**Lutherans** are a minority among all Protestants’.

(5) *Большая часть Вселенной состоит [Func1(часть)] из загадочного невидимого вещества, называемого тёмной материей* ‘The greater part of the Universe consists of **mysterious invisible substance, called dark matter**’.

In addition, the valency of Part may be instantiated by a dependent genitive NP, which seems quite exotic as normally this pattern is characteristic of the valency of Whole. Compare the following two sentences:

(6) *Руководство компании перевело за границу 20% дохода* ‘The company’s leadership transferred abroad 20% of their income’.

(7) *Вложение принесло 20% дохода* ‘The investment yielded 20% income’.

The two sentences (unlike their English equivalents) appear to be very similar. Yet in (6) and (7) the noun *доход* ‘income’ fills the different valencies of the partial *процент* ‘per cent’. In (6), this is the valency of Whole: 20% of the total income were taken and transferred abroad. In (7), this is the valency of Part: the income brought up 20% of the total investment. Note that in either case one of the two valencies remains unsaturated.

Below is another example illustrating the instantiation of the valency of Part by a dependent genitive NP with the unfilled valency of Whole:

(8) *Воссозданию класса рантье способствует и наше либеральное налоговое законодательство: 13% подоходного налога – это немного, к тому же многие не платят вообще ничего* ‘Our liberal tax legislation helps the recreation of the rentier class: 13% of income tax is not much, besides, many pay nothing at all’.

Here, income tax is 13% of the total income, which is not specified in the sentence.

There is one more type of constructions in which both the valency of Part and the valency of Whole display non-prototypical instantiations. The valency of Part is filled by the dependent genitive NP, while the valency of Whole is represented by an expression made up by the VP:

(9) *Большинство жителей башкирским не владеют, ведь русских в городе живет больше половины, татар – процентов тридцать, а башкир ... всего пятнадцать.* ‘The majority of residents do not speak Bashkir, since Russians living in the city constitute more than half of the population, Tatars, about thirty percent and Bashkirs, only fifteen.’

The relevant part of (9) – *Русских в городе живет больше половины* lit. ‘Of Russians, in the city live more than half’ demonstrates the fact that the valency of Part of the word *половина* ‘half’ is filled in by the noun *русских* ‘Russians’ (in the genitive) and the valency of Whole is presented with the VP *в городе живёт* ‘live in the city’: ‘more than half of those who live in the city are Russians’. Similarly, sentence

(10) *На Самотлорском месторождении только 4% нефти, а все остальное – это вода* ‘At the Samotlor oil field there is only 4% of oil, and the rest is water’.

says that 4% of what there is at the oil field is oil.

2.3 Additional valencies of partials

So far we have been discussing the partials with two valencies – Part and Whole. However some of the partials have at least three valencies. These include the words *большинство* ‘majority’, *меньшинство* ‘minority’, and *часть*² (the latter will be discussed in Section 4.2).

The nouns *большинство* and *меньшинство*, in addition to the above mentioned valencies, have a third valency of Property. Consider the following definition:

- *большинство* <*меньшинство*> (Q,R,P) = ‘Q is part of a set R such that this Q possesses the property P; this part is greater <smaller> than the remaining part of R’.

This definition can be illustrated by the following example:

(11) *Ныне же большинство загадок, ранее казавшихся непостижимыми, благополучно разрешены наукой* ‘Now the majority of mysteries that earlier appeared to be incomprehensible, are successfully solved by science’.

This sentence means that among the mysteries that seemed unsolvable (**Whole**) such mysteries that have the **Property** of being solved by science constitute a greater part of mysteries than those that do not possess this property.

As a rule, the valency of Property is expressed by the VP, as in (11). However, curious variations can be observed here, too. For instance, this valency is very often represented by an adjectival modifier, as in *русскоязычное большинство* ‘Russian-speaking majority’, which refers to the part of the population for which Russian is the native tongue and which exceeds the remaining part of the population, or in *национальное меньшинство* ‘national minority’¹. Cf., for example,

(12) *На юге жили кхмерские народы; они и сегодня составляют здесь значительное национальное меньшинство* ‘Khmer people lived in the south, they constitute a considerable national minority’.

3 OTHER PARTIALS

3.1 Voting majority

The noun *большинство* ‘majority’ has another meaning in addition to the one discussed above: for convenience, we will call it the voting majority and denote as *большинство*². This lexical unit has a number of features that distinguish it from the

¹ Interestingly, this last phrase is in fact an idiom as it denotes a part of the population that belongs to a particular nationality while all other inhabitants of the respective country or territory (not necessarily belonging to one and the same nationality!) outnumber this part. It is worth adding that this idiom may be used in plural: *национальные меньшинства* ‘national minorities’ and even has a special abbreviation: *нацменьшинство*. Another example of such phrases is a recently emerged *сексуальное меньшинство* ‘sexual minority’: surprisingly, its first occurrence in RNC dates to 1991 and remained exceedingly rare until 2000.

regular *большинство*. First of all, its valencies have a different content: (i) Who has majority? (ii) Over whom? (iii) Which number of votes does the majority party have? These properties can be illustrated by sentences like

(13) *Лейбористы добились большинства над консерваторами* ‘The Labor Party achieved a majority over the conservatives’.

(14) *Для решения такого вопроса требуется большинство в две трети голосов <семидесятипроцентное большинство>* ‘To resolve this issue, a two-thirds majority <a seventy-percent majority> is needed’.

(15) *Альенде сменил президента Эдуардо Фрея Монтальву, который, в отличие от него, был избран небывалым в истории страны большинством в 56% голосов.* ‘Allende replaced President Eduardo Frei Montalva, who, by contrast, was elected by a majority of 56% of votes, unprecedented in the history of the country’.

Interestingly, the quantitative expression introduced by the preposition *в* (which has in this case a loose English equivalent *of*) may refer to two different things: it can either denote the number (or share) of votes constituting the majority, as in (14)–(15), or the difference between the larger and the smaller part of the votes, as in (16):

(16) *Правление большинством, правда, всего в один голос (четыре против трех), проголосовало за вас.* ‘The board voted for you, albeit by a majority of only one vote (four against three)’.

The difference between the larger and the smaller part may also be expressed by the preposition *на* ‘by’, which is rather rare and obsolete but attested by the corpora:

(17) *И тогда у Ленина образовалось большинство на один голос.* ‘Then Lenin achieved a majority of one vote’.

It is curious that the same fragment of meaning – the difference between the larger and smaller parts – may be conveyed by an adjective. *Незначительное большинство* ‘slim <narrow> majority’ does not mean that the majority has a small (or negligible) number of elements but that its difference from the minority is small. So we have here a non-compositional phrase, almost an idiom. Amusingly enough, the phrase *незначительное меньшинство* ‘small minority’ has the meaning that should be expected: it includes very few elements.

Another distinctive feature of *большинство*² is the fact that it can be used in the instrumental case, denoting a mode of action, as in (15) – (16) above.

Finally, the word has a very specific lexical combinatorics; cf. such expressions as *простое большинство* ‘simple majority’ or *квалифицированное большинство* ‘qualified majority’.

3.2 Percentage

The noun *процент* ‘percentage’ has a specific use which emerges when it stands in the singular and has no dependent numeral: we will refer to it as *процент*². For example,

(18) *Процент недовольных растет* ‘The percentage of discontented people is growing’

does not mean that the hundredth part of those discontented is increasing but that the number of such people, expressed as per cent of the whole population is growing. Similar to (9)–(10) above, the NP in the genitive (*недовольных*) instantiates here the valency of Part, and not Whole.

Thus, the neutral usages of expressions like *большинство студентов* ‘the majority of students’ and *часть студентов* ‘part of students’, on the one hand, and *процент студентов* ‘percentage of students’, on the other hand, need different interpretations: in the first case, students are the enveloping set and in the second case they are a part of the enveloping set. Hence, the expression *процент студентов в городе* ‘percentage of students in the city’ implies that the city also has non-students, while the phrase *большинство студентов в городе* ‘the majority of students in the city’ does not presuppose any involvement of non-students.

3.3 The larger part

the expression *большая часть* ‘the larger part’ has two essentially different interpretations: (a) part of a whole that is greater than some other part of this whole, and (b) part of a whole that is greater than the remaining part of the whole. The first interpretation is seen in

(19) *Каждый год ситуация будет ухудшаться – либо придется повышать налоги, отнимая у людей все большую часть заработной платы, либо придется уменьшать пенсии.* ‘With every year the situation will worsen – either one will have to raise taxes, taking away a bigger and bigger part of the salary from people, or one will have to cut pensions’.

Here, *большая часть* has a standard interpretation of the comparative (a): one will take away a part of the salary which will be greater than the part taken away previously.

The second interpretation (b) may be illustrated by the sentence:

(20) *У меня отняли большую часть зарплаты* ‘They took away the larger part of my salary’ (i.e. took away the part which is larger than the remaining part).

In interpretation (b), the expression cannot be viewed as a comparative (note that it cannot accept a comparative conjunction *чем* ‘than’). In this case *большая часть* ‘the larger part’ is idiomatic and represents a microsyntactic construction, largely synonymous to the word *большинство* ‘majority’ (with the exception that *большинство* requires a whole that can be counted, cf. *большая часть наследства* ‘the larger part of the heritage’ but not **большинство наследства* ‘the majority of heritage’).

3.4 By half

Russian has several units whose meaning is related to ‘half’. These are the noun *половина* ‘half’, adverbs *пополам*, *наполовину*, *вполовину* ‘by half’, the numeral

пол and prefix *полу*-². Below, we will compare *наполовину* and *полу*-, which in some contexts manifest curious contrast.

At first sight, these units are very close – cf. pairs (21)–(22) and (23)–(24) where they are synonymous.

(21) *Мой отец отличался мягкостью сердца, легкостью нрава – и целым винегретом из генов: был швейцарский гражданин, полуфранцуз-полуавстриец, с Дунайской прожилкой.* ‘My father was a gentle, easy-going person, a salad of racial genes: a Swiss citizen, of mixed French and Austrian descent, with a dash of the Danube in his veins’ (Nabokov, *Lolita*), lit. ‘half-French, half-Austrian’; ‘half’ expressed by prefix *полу*-.

(22) *Мой отец был наполовину француз, наполовину австриец.* ‘My father was half French and half Austrian; ‘half’ expressed by the adverb *наполовину*.

(23) *Она посмотрела на него полуиспуганно-полуудивленно* ‘She looked at him half-frightened and half-surprised’; ‘half’ expressed by prefix *полу*-.

(24) *Она посмотрела на него наполовину испуганно, наполовину удивленно* ‘She looked at him half frightened and half surprised’; ‘half’ expressed by adverb *наполовину*.

It can be noted however that the adverb and the prefix behave somewhat differently. The examples above show that they both form two-element strings: *полуфранцуз-полуавстриец, наполовину француз, наполовину австриец* ‘half French and half Austrian’. However, *наполовину* cannot form longer chains, cf.

(22’) **наполовину француз, наполовину австриец, наполовину китаец* ‘half French, half Austrian, half Chinese’,

while the prefix *полу*- is not bound by this constraint:

(23’) *Она посмотрела на него полуиспуганно, полуудивленно, полувопросительно* ‘she looked at him half-frightened, half-surprised and half-questioningly’.

(25) *Полу-милорд, полу-купец, полу-мудрец, полу-невежда, полу-подлец, но есть надежда, что станет полным наконец* ‘half-milord, half-merchant, half-wise man, half-illiterate, half-scoundrel...’ (an epigram by Pushkin).

In our opinion, this difference can be explained as follows. When we use *наполовину* in (23) to say that she looked at him half-frightened and half-surprised, we mean that among the properties of her look, one half is a manifestation of fright and the other one is a manifestation of surprise. This enumeration cannot be extended because nothing can have more (or less!) than two halves. Sentence (23’) with the prefix carries a different message. We do not describe two halves of the look. We characterize different components mixed in her look, claiming that the look manifests different feelings and nothing can prevent it from manifesting several feelings at a time. On the other hand, each of them is not presented in its entirety but only

² The semantics of this prefix was described in Boris Iomdin ([4]) in much detail. However, the phenomena discussed below were not accounted for.

partially. It is not real fright, it is only half-fright. Thus, the difference between *наполовину* and *полу-* is related to what the whole is from which halves are extracted.

However, prefix *полу-* can have both interpretations, and it may be difficult to decide which of them is preferable in a concrete case. For example, sentence (26) can mean both (a) ‘not quite jokingly, not quite lovingly’, and (b) ‘each of two pure elements – joke and love – constitutes one half of the whole attitude’:

(26) *Извини, Васъвасъ, – так, полушутя-полулюбовно, она называла своего мужа, – но это тебя* (А. Якунин) ‘Excuse me, Vas’vas’, – this is how half-jokingly and half-lovingly she called her husband – but this is for you’.³

4 CERTAIN MICROCONSTRUCTIONS

The expressions discussed below can be viewed as microsyntactic constructions in the sense of Iomdin ([5], [7]) as they have specific syntactic and semantic properties. We will look at two such constructions in sufficient detail.

4.1 In the majority of cases

The expression discussed has two variants differing in word order: (a) *в большинстве своём* and (b) *в своём большинстве*, where the reflexive adjective *своём* precedes the noun; (a) and (b) are fully synonymous. Basically, this is an adverbial derivate of the noun *большинство* ‘majority’, which inherits its major semantic valencies, in particular, the valencies of the Whole and Property. Prototypically this adverbial is dependent on the predicate P (often expressed by a verb) and semantically refers to its subject X, stating the fact that most of the individual entities of the (collective) subject X are affected by the predicate:

(27) *Математики, в большинстве своём, не замечают, что слово «неподалёку» означает нечто большее, чем малость расстояния.* ‘Mathematicians, in their majority, do not notice that the word ‘nearby’ means something more than smallness of distance’ (Russian mathematician Vladimir Uspensky on the barrier between mathematics and humanitarian science).

In (27), *математики* ‘mathematicians’ is X, *не замечать* ‘not to notice’ is P, and the adverbial refers to the part of mathematicians which is greater than the part who do not fall under the category of non-noticers. In this case, X instantiates the valency of Whole, while P instantiates the valency of Property of this adverbial.

By default, X is expressed by a count noun in plural denoting humans, but corpus data show that these restrictions can be overridden quite easily. Cf.

(28) *Человек в большинстве своем слаб* (А. Розенбаум) ‘Man is for the most part weak’,

where *человек* ‘human’ is used generically; or

³ We thank Valentina Apresjan for bringing our attention to this example.

(29) *Современные приборы в большинстве своем страшно дороги и сложны* ‘Modern appliances, in their majority, are terribly expensive and complicated’.

What cannot be overridden, though, is the fact that the adverbial must be subject-oriented – probably due to the fact that it contains the reflexive adjective *своём* ‘one’s own’, whose syntactic behavior in Russian obeys the rule that it should be anaphorically related to the subject of a predicate (and, typically, even to the grammatical subject) – even though the presence of this reflexive adjective is, in our opinion, hardly justified semantically⁴. So, sentences like

(30) *Больные уважают врачей в своем большинстве* ‘Patients respect doctors, for the most part’,

where the property P (‘respect’) refers to the majority of patients and not to the majority of doctors, notwithstanding the linear sequence of nouns that could, hypothetically, claim to be fulfilling the adverbial’s valency of Whole. In (30) this valency is instantiated by the noun that plays the role of the grammatical subject.

A similar restriction seems to hold in the Russian National corpus example

(31) (*Согласно статистике, сегодня в России более шести миллионов неполных семей.*) *В большинстве своём, конечно, детей воспитывают матери-одиночки.* ‘According to the statistics, there are over six million one-parent families in Russia today). In the majority of cases, naturally, it is single mothers who raise children’:

In (31), the adverbial *в большинстве своём* refers to the majority of single mothers (also the grammatical subject of the verb) rather than to the majority of children⁵. However, the semantics of this sentence is very different from that of (30). In order to explicate this difference, let us compare two sentences (32) and (33), essentially produced with the lexical material of (31):

(32) *Матери-одиночки, в большинстве своём, воспитывают детей без помощи бабушек и дедушек* ‘Single mothers, in their majority, raise children with no assistance from grandparents’;

(33) *Воспитывают детей без помощи бабушек и дедушек, в большинстве своем, матери-одиночки* lit. ‘raise children with no assistance from grandparents, in their majority, single mothers’ ≈ ‘In the majority of cases, it is single mothers who raise children with no assistance from grandparents’.

⁴ The etymology of the adverbial and especially the use of the reflexive adjective here require special research, which we cannot undertake now. It is noteworthy however that a similar reflexive pronoun occurs in the equivalents of the microconstruction *в большинстве своём* in other Slavic languages: Ukrainian (*в більшості своїй*), Belarussian (*у большасці сваёй*), Polish (*w większości swojej*), and Bulgarian (*в по-голямата си част*).

⁵ Of course, common sense expectations may also bring us to the conclusion that the majority of children from one-parent families are brought up by single mothers – but this is explained by logical inference and not by the lexicographic definition of the adverbial.

These sentences have identical syntactic structures but different word order and communicative structures, and it turns out that the adverbial *в большинстве своём* makes essentially different contributions into their semantics. Sentence (32) states that the majority of single mothers raise their children alone. In contrast, sentence (33) says that among all people who raise children with no grandparental assistance the majority are single mothers. This means that the valencies of our adverbial are expressed differently in (32) and (33). As a first approximation, the rule defining the valency instantiation looks as follows: the sentence's Theme fills in the valency of Whole, while its Rheme fills in the valency of the Property. Indeed, in (32) the NP *single mothers* constitutes the Theme and the remaining VP makes up the Rheme. Accordingly, (32) means 'the majority of single mothers have the property of raising children alone'. In other words, single mothers are the enveloping set (from which a subset of those receiving no help from grandparents is chosen).

In (33), the Theme is formed by the VP (those who raise children alone) and the Rheme is *single mothers*. By our rule, the meaning of (33) is 'the majority of those who raise children alone have the property of being single mothers'.

Looking back at sentence (31), we find the same situation: *single mothers* fill in the valency of Property of our adverbial – a different valency that is instantiated by the same NP in (30).

Of special interest is the case where our adverbial refers to a predicate expressed by an adjective, which in its turn plays the role of a modifier of a noun:

(34) *Религиозные в большинстве своем граждане Соединенных Штатов уделяют ритуальным услугам особое внимание* 'Religious, in their majority, citizens of the United States pay special attention to ritual services' (RNC).

Syntactically, the adverbial is clearly subordinated by the adjective. Yet, semantically, the adjective *религиозные* 'religious' fills its valency of Property, thus manifesting a passive pattern of valency filling. It should be added that the valency of Whole of our adverbial is instantiated by the noun *граждане* 'citizens', exemplifying another valency following a passive pattern. Most interestingly, (34) does not specify the valency of Part of the adverbial in any way: there is no evidence that something is peculiar to the religious (or non-religious, for that matter) part of the citizens – their interest in ritual services is stated for the whole body of them, notwithstanding the religiosity of particular persons. It should be added that in constructions like (34) the adjective always acts as a qualificative and not as a restrictive modifier in the sense of Jespersen ([8]).

Note that it is not the only microconstruction to reveal this kind of behavior. The same valency distribution and the qualificative status of the adjective can be seen in the largely synonymous adverbials *по большей части* 'for the most part', *в основном* 'mainly', *по преимуществу* 'predominantly, par excellence' and the latter's one-word adverb variant *преимущественно* 'predominantly'; cf.

(35) *Его понтификат проходил на фоне Первой мировой войны ..., когда развалилась Австро-Венгерская империя с католическим по преимуществу населением* ‘His pontificate took place during World War I... when the Austro-Hungarian empire with its predominantly Catholic population collapsed’ (newspaper subcorpus of RNC).

To conclude the discussion of the microconstruction *в большинстве своём* we would like to make four more comments.

(i) The adverbial *в большинстве своём* itself tends to be unstressed and precedes the predicate that instantiates its valency of Property. Normally, however, it does not belong to the Theme of the utterance (and never belongs to the Rheme thereof).

(ii) There are a few other adverbials derived from *большинство* which do not include the reflexive adjective (e.g. *в большинстве случаев* ‘in the majority of cases’, *в большинстве ситуаций* ‘in the majority of situations, or simply *в большинстве* with no extending words – lit. ‘in majority’, the latter occurring relatively rarely:

(36) *Люди в большинстве не знали, куда вложить ваучер* ‘The people, in their majority, did not know where to invest their voucher’.

These adverbials do not share the requirement to be oriented to the subject, cf.

(37) *Внешняя атрибутика для турков в большинстве важнее сути* ‘For Turks, in their majority, external attributes are more important than the essence’

(iii) In singular cases, the microsyntactic construction may have an additional adjectival modifier of the word *большинство*: cf.

(38) *Архитекторы в абсолютном своём большинстве сторонники европейского типа застройки* ‘Architects in their absolute majority are supporters of the European type of housing’.

Interestingly, such modifiers need not even express the high degree of majority, although in most cases they do. However, in

(39) *Но в своих поступках люди, в своём нормальном большинстве, все же руководствуются не снами* lit. ‘But in their actions people, in their normal majority, are not guided by dreams after all’ (newspaper subcorpus of RNC),

the adjective *нормальный* ‘normal’ introduces the third valency of the word *большинство* and hence of the adverbial under study, namely, the valency of Part. Indeed, in (39) the valency of Whole is expressed by the word *люди* ‘people’, the valency of Property is instantiated by the predicate complex *руководствоваться не снами* (lit. ‘not be guided by dreams’), and this property is used to select the subset of people that we want to call normal.

(iv) On the periphery of the lexical system, antonymic constructions for our adverbials can be found, such as *в меньшинстве своём* ‘in their minority’. They are for the most part used ironically, as part of language play, implicitly referring to the situation described by the original construction, as in

(40) *Женская половина в меньшинстве своём следит за своим внешним видом.* ‘The female half, in their minority, care about their looks’

The occurrences of such potential lexical units are rare and their representation in corpora is negligible.

4.2 Constructions of the type *tret'ja čast'*

In Section 2, we discussed the partial *часть*¹ 'part'. This noun, however, has a different meaning *часть*² which also belongs to the class of partials but is better considered here as it never occurs independently and forms a special microsyntactic construction. This meaning can be illustrated by (41):

(41) *Мэри может рассчитывать на пятую часть наследства* 'Mary can count on a fifth of the inheritance'.

The word *часть*² 'part' may be given the following definition: *Х есть N-ая часть² Y-а = 'X is one of N equal parts (часть¹) of Y, which, taken together, constitute Y'*. *Часть*² differs from *часть*¹, in particular, in the fact that it has an additional syntactically obligatory valency N, denoting the number of equal parts into which the whole is divided, which has to be expressed by an ordinal adjective. The obligatory character of valency N does not imply that sentences from which such an ordinal adjective is absent will be ungrammatical. It only means that without such an adjective the meaning *часть*² cannot be realized. If we omit the adjective in (41), the idea of the whole being divided into equal parts will disappear and the word *часть* will be interpreted as *часть*¹:

(42) *Мэри может рассчитывать на часть¹ наследства*. 'Mary can count on a part of the inheritance'.

The word *часть*² is related with a whole set of words who share the following property: in the meaning of these words the variable N is filled in by a concrete number: *половина*, *вторая* (N=2) 'a half', *треть*, *третья* (N=3) 'a third', *четверть*, *четвертая* (N=4) 'a quarter', *пятая* (N=5), 'a fifth', *шестая* (N=6) 'a sixth' etc. The nominalized adjectives belonging to this list may be presented as the result of omission⁶ of the noun *часть*² in phrases like *пятая часть*². An interesting difference, however, can be observed: in contrast to phrases like *пятая часть*², the nominalized adjectives require a numeral:

(43) *Мэри может рассчитывать на одну пятую <две пятых, три пятых> наследства*. 'Mary can count on a fifth <two fifths, three fifths> of the inheritance'.
but not

(44) **Мэри может рассчитывать на пятую наследства*.

Despite the fact that this numeral is syntactically compulsory, it does not form a new valency, but is a normal modifier.

To complete the account of this microsyntactic construction, we should add that it has a closely synonymous construction formed with the word *доля* 'share' which is gradually becoming obsolete: *третья доля* 'a third', *миллионная доля* 'a millionth'

⁶ In much the same way, constructions like *пятое марта* (and its English equivalent the fifth of March) can be viewed as omissions of the word *число* or day, resulting in the nominalization of the ordinal adjectives.

etc. It has curious peculiarities differing it from the word *часть*² (mainly of selectional restriction nature) but they fall out of scope of this paper.

5 CONCLUSION

We have studied a class of expressions with the meaning of parts of a whole, involving individual words and microsyntactic constructions. We were primarily interested in valency properties of these expressions, choosing the material from the Russian National Corpus and a few other resources. Due to this approach, we were able to find and explain a number of interesting phenomena which eluded researchers' observation so far.

ACKNOWLEDGMENTS

This work was partly supported by a grant No. 16-18-10422-P from the Russian Scientific Foundation, which is greatly appreciated.

References

- [1] Boguslavsky, I. (2005). Valencies of Quantifying Words. [Валентности кванторных слов.] // *Kvantitativnyj aspect jazyka*. Moscow, pages 139–165. (In Russian).
- [2] Boguslavsky, I. (2013). Adverbial partials in Russian (vdvoe ‘twice as much/half’, napolovinu ‘half’ and others). In *Proceedings of the 6th International Conference on Meaning-Text Theory Prague, August 30–31, 2013*. Eds. V. Apresjan, B. Iomdin, and E. Ageeva. Available at: <http://meaningtext.net/mtt2013/proceedings>
- [3] Boguslavsky, I. (2018). Partial expressions in Russian. [Парциальные выражения в русском языке.] *Voprosy jazykoznanija*, 2, pages 29–52. (In Russian).
- [4] Iomdin, B. (2003). The semantics of the Russian Prefix ПОЛУ-. [Семантика русской приставки ПОЛУ-] // *Rusistika na poroge XXI veka: problem I perspektivy*. In *Proceedings of an international scientific conference (Moscow, 8–10 June 2002)*, pages 109–113. Moscow, Institute of Russian Language. (In Russian.)
- [5] Iomdin, L. (2017a). Between a syntactic idiom and a syntactic construction. Nontrivial cases of microsyntactic ambiguities. [Между синтаксической фраземой и синтаксической конструкцией. Нетривиальные случаи микросинтаксической неоднозначности.] *SLAVIA, časopis pro slovanskou filologii*, 86(2–3), pages 230–243. (In Russian.)
- [6] Iomdin, L. (2017b). Microsyntactic annotation of Corpora and its use in Computational Linguistics Tasks. *Jazykovedný časopis*, 86(2), pages 169–178.
- [7] Iomdin, L. (2018). Once again on microconstructions formed by functional words. To i delo. [Еще раз о микроконструкциях, сформированных служебными словами: То и дело.] // *Computational Linguistics and Intellectual Technologies. International Conference (Dialog '2018)*, pages 267–283. Moscow, RGGU Publishers, 17(24). (In Russian.)
- [8] Jespersen, O. (1933). *Essentials of English grammar*. London, Allen & Unwin.
- [9] Meřčuk, I. (2014). *Semantics: From Meaning to Text*. Vol. 3. Amsterdam, John Benjamins Publishing Company.

WACKERNAGEL'S POSITION AND CONTACT POSITION OF PRONOMINAL ENCLITICS IN OLDER CZECH. COMPETITION OR COOPERATION?

RADEK ČECH¹ – PAVEL KOSEK² – OLGA NAVRÁTILOVÁ² – JÁN MAČUTEK^{2,3}

¹University of Ostrava, Ostrava, Czech Republic

²Masaryk University, Brno, Czech Republic

³Comenius University, Bratislava, Slovakia

ČECH, Radek – KOSEK, Pavel – NAVRÁTILOVÁ, Olga – MAČUTEK, Ján:
Wackernagel's position and contact position of pronominal enclitics in Older Czech.
Competition or Cooperation? *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 267 – 275.

Abstract: The paper focuses on analyzing the relationship among word order positions of pronominal enclitics in the history of Czech. Specifically, we look at the Wackernagel's position and the contact position and we try to decide whether these two positions compete, as usually taken for granted, or whether there is a certain kind of cooperation between them. The results show that the positions do not compete, at least not in the majority of cases. We used a corpus-based on selected books of the first edition of the Old Czech Bible and Kralice Bible for the analysis.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 INTRODUCTION

This article focuses on analyzing the word order of older Czech pronominal enclitics dependent on a finite verb in the corpus of selected books from a) the younger copies of the first edition of the Old Czech Bible – Olomouc Bible (Bible olomoucká) and the Litoměřice-Třeboň Bible (Bible litoměřicko-třeboňská) – and b) from the Kralice Bible (Bible Kralická). Previous research ([1], [2], [3]) shows that the word order of the older Czech pronominal (and auxiliary) enclitics follows one of the two main patterns: 1. the pronominal enclitic is in the Wackernagel's position (also called the post-initial position), i.e. the second position in a clause, the example (1) demonstrates this pattern for the enclitic pronominal form *mi* 'to me', 2. the pronominal enclitic is in the contact position, i.e. in the position that is in the immediate vicinity of its superordinate verb, its governor (hence, also called verb-adjacent position). This pattern is demonstrated in the example (2):¹

¹ Both word-order patterns occur in texts of various stages in the historical development of Czech in several variants: 1. the second position differs between the position of the pronominal enclitic after the first phrase of a clause and the position after the first word of the first phrase of a clause, 2. the contact position

- (1) a[| *Kto mi toho pojičí*, | [*aby byly popsány řeči mé?*]]
 b. Who_{NOM.SG} me_{DAT.SG} this_{ACC.N.SG} lend_{3.SG. FUT}²

BiblOI Jb 19,23

- (2) a. *Hospodin böh otevřel mi jest ucho...*
 b. Lord_{NOM.M.SG} God_{NOM.M.SG} open_{PTCP.PST.M.SG} me_{DAT.SG}
 be_{AUX.PRS.3.SG} ear_{ACC.N.SG}

BiblOI Isa 50,5

Both word-order patterns (positions) exist in modern Slavic languages [4], so that the situation of older Czech – showing the same variation – is relevant for research of contemporary Slavic languages as well.

Although enclitics are considered to be a group of heterogeneous language units [5, 6], they share some common characteristics (at least stochastic ones) manifested in their phonological and syntactic behavior. First, they are prosodically deficit, i.e. they bear no word stress, and, consequently, they are prosodically joined with the preceding word. Moreover, they have a strong tendency to appear in the second position in a clause and that is true for various languages. This is the well-known Wackernagel's position (also marked 2P in the following text) [7]. In the second position pattern, enclitic's syntactic governor does not have to be the same word that the enclitic is prosodically joined with. Thus, both prosodic and syntactic properties (and their interplay) influence enclitics' word order.

According to the well-accepted assumption, the Wackernagel's position is the original position of enclitics in Indo-European languages and, hence, also a common linguistic pattern in Proto-Slavic. The emergence of the contact position in the historical development of the Slavic languages has been interpreted as a manifestation of the grammaticalization process that transformed enclitics to inflectional affixes [8], cf. Russian *он смеялся*, where the original enclitic *ся* is a non-separable part (morpheme) of the word.

Pancheva [9] suggests that the word order of these language units and the development of their positions is more complex. First, she shows that we need a more detailed classification of particular positions in order to understand this phenomenon properly. Second, her analysis of the Old Bulgarian examples challenges the general view on the grammaticalization process substantially. Similarly, ([2], [10]) discussed other factors that influence the word order of enclitics (especially the possible variations in both the second and contact positions) in older

differs between the position of the pronominal enclitic after its governor (postverbal position) and the position of the pronominal enclitic before its governor (preverbal position). For more details, see [1], [2], [3].

² To translate the Old Czech examples completely would lengthen this paper unacceptably; hence, we only cite one example for each phenomenon and a gloss is given just for the relevant part of the example (the glossed parts of the example are indicated by a vertical line |).

Czech systematically (concerning style, length of the initial phrase, etc.). To sum up, these studies show that the problem requires further discussion.

However, there is an essential problem of the word order of enclitics, and that is the relationship between the Wackernagel's and contact patterns. As alluded to above, the relationship has not been analyzed fully yet in neither of the above-mentioned works. From recent studies, one might get an impression that the 2P and the contact position are result of different mechanisms, that, somehow, seem to compete with each other. However, the syntactically superior element (the enclitic's governor) can occur in the first and the third position in a clause, i.e. in a position adjacent to the enclitic. In this case, there is no competition between these two positions – if anything, we might talk about cooperation. To our great surprise, there is no detailed analysis of this phenomenon (except [9], where the problem is mentioned, but not analyzed thoroughly). We see the problem as crucial for the following reasons. A finer classification of the elements involved in the 2P position could shed light on the principles behind the Wackernagel's law. For instance, if – in majority of cases – an enclitic falls into the 2P and this position is also the contact position, then it means that the law influences not only the enclitic position, but also the position of the clitic's governor. Alternatively, the position of the syntactic governor can play a more important role than usually assumed even in the case the enclitic is in the 2P. More generally, it is possible to consider this problem as an instance of the least effort principle [11]. In any case, we need better empirical evidence, so that we can gain more substantial insight into the problem. Therefore, in this study, we analyze the relationship between the 2P and the contact position of the enclitic in Older Czech.

Our aim is to observe whether their relationship is competitive, cooperative or neutral (for details see Section 2). Older Czech is chosen intentionally for the following reasons: a) there is a variability of word order (cf. [1], [2], [3]), especially if compared to the contemporary Czech (the relative rigidity of clitic placement in the contemporary Czech might be sought in linguistic prescription established in the middle of the 20th century); b) we chose texts that represent both the oldest period (14th century) and younger period (16th century) with enough language material available for linguistic research. Thus, it is a proper starting point for modeling the historical development of this phenomenon.

2 LANGUAGE MATERIAL AND METHODOLOGY

We chose two Czech Bible editions translated in different periods and from different pretexts: 1. The first edition of the Old Czech Bible (2nd half of 14th century), 2. Kralice Bible³ (1579–1594). This material was chosen for the following reasons:

³ This Bible was highly valued for its brilliant language and it was re-printed repeatedly. It also served as a model (and an unattainable) ideal for the Modern Czech codification in the 19th century.

1. The first edition represents one of the oldest Old Czech prose texts⁴ (original Czech texts from an earlier stages are not suitable for the word-order analysis: it is poetry). 2. In our view, the diachronic perspective desired for our research is best brought by comparison of two different historical translations of similar texts. The texts are similar, but crucially, they are not the same: a) the first edition of the Old Czech Bible and Kralice Bible were translated by different translators, b) the first edition of the Old Czech Bible was translated from the Middle Age Latin Vulgata,⁵ whereas the Kralice Bible was translated by the members of the Unity of the Brethren (Jednota bratrská) from the Latin and Greek pretexts (New Testament) and Hebrew and Latin pretexts (Old Testament).⁶

Since the language material must be annotated manually, we restricted ourselves to the selected books both from the Old and New Testament. Intentionally, we selected texts with different styles and structure, as well as texts by different translators: The Gospels of Matthew and Luke, the Acts of the Apostles, the Revelation of John from the New Testament and Genesis, Job, Sirach, and Isaiah (chapters 14 to 66) from the Old Testament. For compiling this transcribed corpus, we used 1. the modern edition of the Olomouc and Litoměřice-Třeboň Bible, i.e. the younger copies (from the beginning of 15th century) of the original Old Czech translation (the original itself has not been preserved) [15], [16], [17], [18], [19], 2. the first edition of the Kralice Bible (1579–1594).

To observe “competition” and/or “cooperation” of the two possible word order patterns of enclitics, the language material is annotated as follows. We determine

a) The *postinitial contact position* (2PC position). In this case, the enclitic (E) occurs right after the initial phrase which is its governor (G), schematically

(3) [G] [E] []*

(the symbol []* represents zero or more syntactic units of the clause)

or the enclitic (E) occurs after the initial phrase of any type, except its governor

([]) and the enclitic is immediately followed by its governor (G), schematically

(4) [] [E] [G] []*

b) The *post-initial isolated position* (2PI position). In this case, the enclitic (E) occurs after the initial phrase of any clausal element type except its governor ([]) and it is followed by one or more syntactic element(s) of any clausal element type except its governor ([]+), schematically

(5) [] [E] []+ [G] []*

⁴ From the philological perspective, the language of the Bible is discussed in [12].

⁵ For details, see [13].

⁶ For details, see [14].

c) The *non-post-initial contact position* (NPC position). In this case, the enclitic (E) occurs anywhere except in the post-initial position and it is adjacent to its governor, schematically

(6) [] []+ [E] [G] []*

or

(7) []+ [G] [E] []*

d) The *non-post-initial isolated position* (NPI position). In this case, the enclitic (E) occurs anywhere except in the post-initial position and it is not adjacent to its governor, schematically

(8) [] []+ [E] []+ [G] []*

or

(9) []* [G] []+ [E] []*

It should be noted that the example (9) was not attested in Slavic languages [4] and should be considered ungrammatical.

The distribution of these positions is examined on the pronominal form *mi* ‘to me’. This form was a permanent enclitic already in Proto-Slavic and appears with sufficient frequency in the analyzed biblical texts. The other pronominal forms are either not documented at all *si, ti* ‘to myself / to yourself etc., to you’, or documented in just a few examples *ho, mu* ‘him, to him’, or are not used at all for different reasons; *sě, tě* ‘myself / yourself etc., you’, for instance, could sometimes bear stress and could be used after prepositions.

Frequency of particular positions in the corpora was observed and their proportions were counted. The results are to be interpreted in the following way: a) the prevalence of the 2PC position suggests that there is a cooperation between mechanisms leading to the Wackernagel’s position and the contact position; b) the prevalence of the 2PI position means that the Wackernagel’s law is dominant and it is in competition with the contact position; c) the prevalence of the NPC position should be interpreted so that the contact position is dominant and it is in competition with the Wackernagel’s law, d) the prevalence of the NPI position means that neither the Wackernagel’s law nor the contact position influence the word order of the enclitics in any way.

3 RESULTS

The absolute and relative frequencies of particular positions are shown in Table 1, 2 and Figure 1.

	2PC	2PI	NPC	NPI	Σ
Gn	83	14	11	0	108
%	76.85	12.96	10.19	0	
Jb	33	4	3	1	41
%	80.49	9.76	7.32	2.44	
Ecc	8	2	2	0	12
%	66.67	16.67	16.67	0	
Iz	7	0	4	0	11
%	63.64	0.00	36.36	0	
Mt	22	6	0	0	28
%	78.57	21.43	0	0	
Lk	19	2	1	0	22
%	86.36	9.09	4.55	0	
Sk	26	1	1	0	28
%	92.86	3.57	3.57	0	
Zj	11	2	1	0	14
%	78.57	14.29	7.14	0	
Σ	209	31	23	1	264
%	79.17	11.74	8.71	0.38	

Tab. 1. Absolute and relative frequencies of particular positions in the Olomouc Bible

	2PC	2PI	NPC	NPI	Σ
Gn	74	12	16	1	103
%	71.84	11.65	15.53	0.97	
Jb	32	7	9	0	48
%	66.67	14.58	18.75	0.00	
Ecc	8	2	2	0	12
%	66.67	16.67	16.67	0.00	
Iz	25	4	10	0	39
%	64.10	10.26	25.64	0.00	
Mt	25	4	2	0	31
%	80.65	12.90	6.45	0.00	
Lk	15	3	5	0	23
%	65.22	13.04	21.74	0.00	

	2PC	2PI	NPC	NPI	Σ
Sk	21	5	4	1	31
%	67.74	16.13	12.90	3.23	
Zj	22	1	2	0	25
%	88.00	4.00	8.00	0.00	
Σ	222	38	50	2	312
%	71.15	12.18	16.03	0.64	

Tab. 2. Absolute and relative frequencies of particular positions in the Kralice Bible

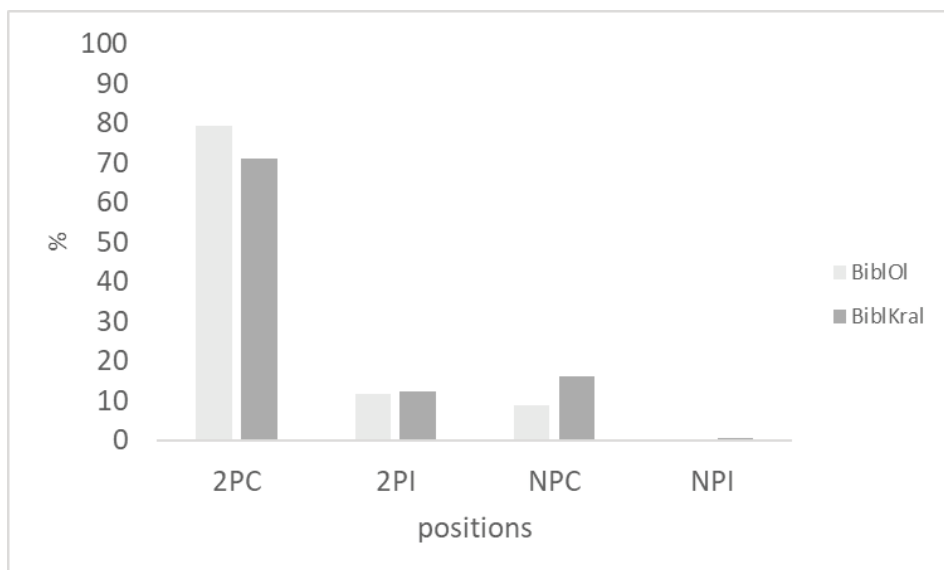


Fig. 1. Relative frequencies of particular positions in the Olomouc Bible (BiblOl) and the Kralice Bible (BiblKral)

The results show that the 2PC position is clearly dominant in all the cases. It means that the Wackernagel's position and the contact position are not in competition in the majority of cases. Furthermore, this result is not influenced by the style of the pretext or the translation. Moreover, a comparison of the Olomouc Bible and the Kralice Bible shows the same tendency in both corpora. Even though there are some differences (a higher proportion of the NPC position accompanied with a lower proportion of the 2PC position in the Kralice Bible), the application of simulate chi-square test reveals that the result is on the border of significant difference (for the significance level $\alpha = 0.05$), $\chi^2 = 7.47$, p-value = 0.058. This means that despite a) the time span of 200 years, b) the different pretexts, and c) different translation "strategy" [14], we identify a stable language behavior for the clitic placement phenomenon.

As for the 2PI and the NPC positions, the differences in their proportions in various biblical books are striking. However, absolute frequencies are too small, thus, it would be wrong to interpret these results. Nevertheless, the above-mentioned higher proportion of the NPC position in the Kralice Bible (for all the books in the corpus), can be interpreted as pointing towards an increasing competition between these two positions. However, only further research can reveal whether it is a manifestation of the historical development, or specificity of the translators of the Kralice Bible, or, eventually, whether it is only a random fluctuation (cf. above-mentioned result of the statistical test). Finally, the minimal frequency of the NPI position shows that there had been a very strong tendency to avoid such word order patterns that do not follow the main enclitic placement strategies, i.e. in the Wackernagel's and/or the contact position (even if the position might be grammatical).

4 CONCLUSION

Our results show that the word order of the selected Czech enclitic pronominal form *mi* 'me' in the chosen two historical Czech Bibles is by and large limited to the two dominant word-order patterns: the Wackernagel's position and the contact position. Surprisingly, these two positions do not compete with each other but rather cooperate: most examples in our study are clauses in which the post-initial and the contact position of an enclitic merge. A question that requires further research is whether this situation is specific to the language of biblical translation or whether it manifests a general mechanism and, further, whether this situation has changed during the following development of Czech.

ACKNOWLEDGMENTS

This study was supported by the project Development of the Czech pronominal (en)clitics (GAČR GA17–02545S).

References

- [1] Kosek, P., Navrátilová, O., and Čech, R. (2018). Slovosled staročeských pronominálních enklitik závislých na VF ve staročeské bibli 1. redakce. *SLAVIA časopis pro slovanskou filologii*, 87(1–3), pages 189–204.
- [2] Kosek, P., Navrátilová, O., Čech, R., and Mačutek, J. (2018). Word Order of Reflexive 'se' in Finite Verb Phrases in the First Edition of the Old Czech Bible Translation. (Part 2). *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 135(3), pages 189–200.
- [3] Kosek, P., Čech, R., and Navrátilová, O. (2018). Starobylá dativní enklitika *mi*, *si*, *ti* ve staročeské bibli 1. redakce. In Malčík, P. (ed.). *Vesper Slavicus. Sborník k nedožitým devadesátinám prof. Radoslava Večerky*. *Studia etymologica Brunensia* 23, Praha, Nakladatelství Lidové noviny, pages 137–151.

- [4] Franks, S., and King, T. H. (2000). *A Handbook of Slavic Clitics*. Oxford.
- [5] Zwicky A. M. (1994). What is a Clitic. In Nevis, J. A., Joseph B. D. et al. (eds.), *Clitics. A Comprehensive Bibliography 1892–1991*, pages 12–20.
- [6] Uhlířová, L., Kosta, P., and Veselovská, L. (2017). Klitika. In Karlík, P., Nekula, M., Pleskalová, J. (eds.): *Nový encyklopedický slovník češtiny*.
- [7] Wackernagel, J. (1892). Über ein Gesetz der indogermanischen Wortstellung. *IF* 1, pages 33–436.
- [8] Zwicky, A. (1977). *On Clitics*. Bloomington, Indiana University Linguistics Club.
- [9] Pancheva, R. (2005). The Rise and Fall of Second-position Clitics. *Natural Language & Linguistic Theory*, 23, pages 103–167.
- [10] Čech, R., Kosek, P., Navrátilová, O., and Mačutek, J. On the impact of the initial phrase length on the position of enclitics in the Old Czech. (to appear)
- [11] Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Addison-Wesley.
- [12] Kosek, P., Navrátilová, O., Čech, R., and Mačutek, J. (2018). Word Order of Reflexive ‘se’ in Finite Verb Phrases in the First Edition of the Old Czech Bible Translation. (Part 1). *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 135(3), pages 177–188.
- [13] Kyas, V. (1997). *Česká Bible v dějinách národního písemnictví*. Vyšehrad.
- [14] Dittmann, R. (2012). *Dynamika textu Kralické bible v české překladatelské tradici*. Refugium Velehrad-Roma.
- [15] Kyas, V. (ed.). (1981). *Staročeská bible drážďanská a olomoucká: kritické vydání nejstaršího českého překladu bible ze 14. století. I. Evangelia*. Praha.
- [16] Kyas, V. (ed.). (1985). *Staročeská bible drážďanská a olomoucká: kritické vydání nejstaršího českého překladu bible ze 14. století s částmi Bible litoměřicko-třeboňské. II. Epištoly. Skutky apoštolů. Apokalypsa*. Praha.
- [17] Kyas, V. (ed.). (1988). *Staročeská bible drážďanská a olomoucká: kritické vydání nejstaršího českého překladu bible ze 14. století. III. Genesis – Esdráš*. Praha.
- [18] Kyas, V., Kyasová, V., and Pečirková, J. (eds.). (1996). *Staročeská bible drážďanská a olomoucká: kritické vydání nejstaršího českého překladu bible ze 14. století. IV. Tobiaš – Sirachovec. Padeborn*.
- [19] Pečirková, J. (ed.). (2009). *Staročeská Bible drážďanská a olomoucká s částmi Proroků rožmberských a Bible litoměřicko-třeboňské, V/1 Izaiáš – Daniel, V/2 Ozeáš – 2. kniha Makabejská*. Praha.

FREQUENCY DICTIONARY OF 16th CENTURY CYRILLIC WRITTEN MONUMENT

OKSANA NIKA¹ – SVITLANA HRYTSYNA²

¹Institute of Philology of Taras Shevchenko National University of Kyiv, Ukraine

²Educational and Scientific Center “Institute of Biology and Medicine” of Taras Shevchenko National University of Kyiv, Ukraine

NIKA, Oksana: Frequency dictionary of 16th century Cyrillic written monument. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 276 – 288.

Abstract: The article presents the algorithm of the frequency dictionary to an original ancient text, “Otpys” (“Response”) by Kliryk Ostrozkyi (the Cleric of Ostroh) of the late 16th century. Until now, no historical corpus of text of the Ukrainian language has been created; therefore the drafting of metagraphical texts with their subsequent processing in accordance with linguistic tasks can fill this gap. The peculiarity of creating a frequency dictionary based on one written monument is in using the model of frequency dictionaries and describing the specifics of processing the ancient text. These specifics is based on a deep understanding of the state of language in the end of the 16th century and consists in the unification of graphic and spelling variants, as well as in the formation of stems and lemmas. Work results are presented in the form of a Frequency Dictionary of Word Forms of “Otpys” by Kliryk Ostrozkyi according to the frequency decrease and a Frequency Dictionary of “Otpys” by Kliryk Ostrozkyi according to the frequency decrease.

Keywords: frequency dictionary, tokenization, stemming, lemmatization, hapax legomena, written monument of the late 16th century

1 INTRODUCTION

The text of “Otpys” by Kliryk Ostrozkyi, the subject of this analysis, is not common for this type of research, having been written in the end of the 16th century. In 2016, monograph “Dialogism in Historical Dimensions of the Old Ukrainian Time: “Otpys” (“Response”) by Kliryk Ostrozkyi” was published, where along with the photocopies of 1598 old-printed book, metagraphical text and “Index of words and word forms” to this text were provided [5, pp. 1–44]. In particular, the purpose of the publication was the “preparation of a Historical Corpus of the Ukrainian language” [5, p. IV].

In Slavic studies, frequency dictionaries of written monuments often became the basis of the research of historical lexicology [6]. O. Tvorohov developed “Frequency Dictionary of ‘The Tale of Bygone Years’ according to the Lavrentiev List” and showed

the significance of such dictionary for historical and linguistic research as well as the description of the type of dictionary to a literature monument ([7], [8]).

In Ukrainian studies, frequency dictionaries have been elaborated based on texts of the same style or genre, as well as individual texts (primarily literary ones) ([1], [2]). These dictionaries are mainly focused on the texts of the late 20th – early 21st centuries. Some of them are based on texts of the 19th century by Taras Shevchenko, Lesia Ukrainka, Ivan Franko.

In this article, frequency dictionary of an original Cyrillic monument of the late 16th century, “Otpys” by Kliryk Ostrozkyi, is created for the first time.

Taking this into account, it is necessary to clarify the usage of metagraphical text and its variants in the process of creating a frequency dictionary. In the article, the old printings of the late 16th century are called the “authentic text”, the metagraphical text with lowered superscripts, the decoding of some contracted words and sometimes simplified spelling are referred to as the “secondary text”, and the processed text after tokenization is the “working” one.

“Index of words and word forms” of the “Otpys” text was important for the analysis [5, pp. 1–44].

2 CREATION PROCEDURE OF THE WORD FORMS FREQUENCY DICTIONARY OF “OTPYS” BY KLIRIK OSTROZKYI ACCORDING TO THE FREQUENCY DECREASE

Tokenization was applied in the first stage of the text processing. For the purposes of analysis of modern languages, the task of tokenization is “to separate words from syntactic characters, numbers, complexes of letters and numbers, Internet addresses, nicknames, symbols %, +, -, //, etc.” [4, p. 35]. Taking into account historical nature of the analyzed text, tokenization of “Otpys” by Kliryk Ostrozkyi had some specifics and was carried out already with the secondary text.

Thus, tilde “[^]” (a superscripted diacritical mark used to reduce frequently repetitive words) was removed from the text. With the help of a computer program (developed by O. Malin, engineer, mechanic, researcher) according to the section “Words with tildes” [5, pp. 63–65], an algorithm of the correlation of words with the tilde to the corresponding words without the tilde was developed. As a result, they were replaced by those with a similar meaning, but complete in their form in the index, and then, for the accuracy of calculation, the respective words were replaced also in the “Otpys” text.

Parentheses in the authentic text serving as punctuation marks were also removed. Remarkably, parentheses “()”, that were introduced in the process of working on the secondary text, remained (they contained superscript letters, and some of ligatures were decoded with their help (^)), as they did not prevent the recognition of individual words that a sequence of letters between spaces was considered to be. In addition, the text was checked for the presence of square brackets that neither were a part of the

authentic text, but contained research information. Among the signs removed there were also dash, hyphen, comma, full stop, semicolon. For the accuracy of calculation, explanatory words in modern Ukrainian language were also removed from the text (for example, “gloss”). Moreover, foliations in the authentic and in the secondary text were removed. In addition, letters with a numerical value were not taken into account, as their sequence did not form words.

The removal of homonymy was no less important. A graphic symbol was added to either word from a homonym pair, thus helping the program to consider these two words as different ones.

With the help of a specially computer program developed by engineer, mechanic, and researcher O. Malin, a working text was prepared and automatic calculation of absolute frequency of each word form was carried out. As a result, the frequency dictionary of word forms of the “Otpys” was developed according to the frequency decrease.

2.1 The lexicographic database of the GDC

The lexicographic database of the GDC is a separate section of the corpus. This platform has functions such as collecting, processing and converting lexicographic data into dictionaries.

The lexicographic database includes:

- Digitized online dictionaries from earlier printed dictionaries
- Various lexicographic data collected from lexicographic fieldwork
- Lexicographic data published by various authors
- Lexicographic data extracted from the existing linguistic and/or ethnographic studies.

The lexicographic database grows continually, with new texts being added over time. The database covers over 10 dialects and lists about 60 000 entries. This database has been developed based on the traditional lexicographic principles and methods. The research team will follow this methods in compiling comprehensive online dictionary of other Georgian dialects. Overall, four online dictionaries has been published so far. The published dictionaries are: Fereydanian, Ingiloan, “Chveneburebi”, and Laz dictionaries. New dictionaries with corresponding lexicographic data will be added to the corpus interface.

3 CREATION PROCEDURE OF THE FREQUENCY DICTIONARY OF “OTPYs” BY KLIRIK OSTROZKYI ACCORDING TO THE FREQUENCY DECREASE

Lemmatization is a traditional stage of creating a frequency dictionary. V. Jongejan and H. Dalianis define it as “the process of reducing a word to its base form, normally the dictionary look-up form (lemma) of the word” [3, p. 45].

In this study, it comprised several stages. The first step was to determine parts of speech, and then to apply the algorithm of stemming to stem words of different parts of speech. According to B. Jongejan and H. Dalian, “Stemming conflates a word to its stem. A stem does not have to be the lemma of the word, but can be any trait that is shared between a group of words” [3, p. 45].

On the basis of the “Index of words and word forms”, words were reduced to their stems by removing endings, and sometimes suffixes. For example, the word forms of the noun **ГОЛОВА**/holova ‘head’ (**ГОЛОВАХЪ**/holovah, **ГОЛОВОЮ**/holovoju, **ГОЛОВЪ**/holov, **ГОЛОВЫ**/holovy, **ГОЛОВѢ**/holovi) were combined in the stem **ГОЛОВ** (holov). With the old Ukrainian language having variations of the graphic writing of the same word, the stemming of adjectives **ДОУХОВНОЙ**/duchovnoj, **ДОУХОВНОМЪ**/duchovnom, **ДОУХОВНОМЪ**/duchovnom, **ДОУХ(В)НЫМИ**/duchovnymy ‘spiritual’ resulted in two separate stems **ДОУХОВН**, **ДОУХ(В)Н**/duchovn. In the example above, the writing of **ОВ** (o), **В(В)** (v) is parallel. In addition, different writing of letters identical in sound can be referred to as graphic unifications: **ЗЗ** (z) (**ЗАКОН**/zakon – **ЗАКОН**/zakon ‘law’), **УҀ** (u) (**МОУДР**/mudr – **МДР**/mudr ‘wise’) and others.

It is worth noting that the same lemma could comprise words with the interchange of consonants, for example: **ГѢ** (g/z) (**ДОРОГҀ**/dorohu – **ДОРОЗѢ**/dorozi ‘way’), **ГЗ** (g/z) (**НОГИ**/nohy – **НЗѢ**/nozi ‘legs’); different degrees of adjectives comparison (**МЕНША**/menša ‘less’ – **НАМЕНШОГО**/namenšoho ‘the least’); different forms of pronouns: **А**/ja ‘I’, **ІА**/ja ‘I’, **МЕНЕ**/mene ‘me’, **МА**/mia ‘me’, **МНѢ**/mni ‘to me’, **МНОЮ**/mnoju ‘by me’, **МИ**/mi ‘me’. All of them formed different stems, for instance: **ДОРОГ**/doroh, **ДОРОЗ**/doroz; **МЕНШ**/menš, **НАМЕНШ**/namenš etc.

Different parts of speech were analyzed separately. Word forms of nouns and adjectives with the same meaning, but different in terms of gender, number, and case were referred to nouns and adjective lemmas. Pronoun lemmas also included word forms, divided in two groups according to the type of declension: those comparable with nouns and others comparable with adjectives. The first group had case forms, and in some instances, indications of gender.

Verbal lemmas with the same lexical meaning united word forms different in tense, mood, aspect. Verbs with prepositional **САСІА** ‘self’ were analyzed in one paradigm (**ПРИВОДИЛАСА**/pryvodylasia – **СА ПРИВОДИЛА**/sia pryvodyla ‘lead’). Verbal forms ending with **-НО**/no, **-ТО**/to were considered as separate lemmas. Participle lemmas combined word forms different in the categories of tense, aspect, voice, as well as gender, number and case, and adverbial participle lemmas – in the categories of aspect and tense. Adverbs and conjunctions were also defined by separate lemmas.

Thus, sets of stems corresponding to one lemma (with the same lexical meaning) were combined in groups. This resulted in a sequence of stem lists, each of them corresponding to one particular lemma.

On the basis of the frequency dictionary of word forms, the program merged word forms (taking into account their frequency) by stems corresponding to

a particular lemma. Thus, the total frequency of word forms which correlate with a certain lemma determined the frequency of the lemma itself. As the result of this stage, frequency dictionary of “Otpys” by Kliryk Ostrozkyi according to the frequency decrease was formed.

4 STATISTICAL REVIEW

1. In the text “Otpys” by Kliryk Ostrozkyi, containing 11 222 words (N), 2 640 lemmas (V) were recorded, which is the total amount of words reduced to the original form, and 4 677 word forms (Vf), the variations of inflexion within the same lemma.

$$N=11\ 222$$

$$V=2\ 640$$

$$Vf=4\ 677$$

2. The diversity of the dictionary (B) of “Otpys” by Kliryk Ostrozky, namely the index of diversity (the ratio of the lemma dictionary volume (V) to the text volume (N)) is 0,235.

$$B=\frac{V}{N} ; B=\frac{2640}{11222}=0,235$$

3. Average repetition of a word in the text (A), that is, the ratio of the text volume (N) to the lemmas dictionary volume (V) for “Otpys” is 4,25. This means that on average, each word is repeated in the text 4 times.

$$A=\frac{N}{V} ; A=\frac{11\ 222}{2\ 640}=4,25$$

4. Calculation of the quantity of words, called hapax legomena (Greek hapax legomena – ‘name once’), is one of the most common stages. It is generally accepted that these are the words used by the author in the text only once: “in the broad sense, hapax legomena can refer to lexical items used only in one written monument, from among all texts representing a significant period in the history of a language” [10, p.12]. Given the absence of a historical corpus of the Ukrainian language until now, it is not possible to compare the usage of hapax legomena in other ancient texts. Therefore, this article deals only with the usage of unique words in the “Otpys” by Kliryk Ostrozkyi. Consequently, the number of words with the frequency 1 (V1) in the analyzed text amounts to 1 632.

$$V1=1\ 632$$

5. The index of uniqueness for the text (Em), that is, the ratio of the quantity of words with the frequency 1 (V1) to the volume of the text (N) of “Otpys” is 0,1454.

$$Em=\frac{V1}{N} ; Em=\frac{1\ 632}{11\ 222}=0,1454$$

6. The index of uniqueness for the dictionary (E_d), that is, the ratio of the quantity of words with the frequency 1 (V_1) to the dictionary volume (V) equals 0,6181.

$$E_d = \frac{V_1}{V} ; E_d = \frac{1\ 632}{2\ 640} = 0,6181$$

7. The index of concentration is opposite to the previous results. The index of concentration in the text (E_{ct}) is the ratio of the number of the most frequent words in the text with frequency of 10 and more (V_{10_t}) to the total text volume (N). $V_{10_t} = 137$, $N = 11\ 222$, respectively, the index of concentration in “Otpys” by Kliryk Ostrozkyi is 0,012.

$$E_{ct} = \frac{V_{10t}}{N} ; E_{ct} = \frac{137}{11\ 222} = 0,012$$

8. The index of concentration in the dictionary (E_{cd}) is the ratio of the number of the most frequent words in the dictionary (with the frequency of 10 and more) ($V_{10} = 175$) to the total dictionary volume (V). In the frequency dictionary of “Otpys” it amounts to 0,066.

$$E_{cd} = \frac{V_{10}}{V} ; E_{cd} = \frac{175}{2\ 640} = 0,066$$

Results of the frequency distribution of lemmas, calculated on the basis of the data contained in the frequency dictionary of “Otpys” according to the frequency decrease, and *word forms*, calculated on the basis of the data of frequency dictionary of the “Otpys” word forms according to the frequency decrease, are represented in Table 1.

Interval of frequencies	Number of word forms	Number of lemmas
500 – 999	1	1
400 – 499	0	0
300 – 399	0	0
200 – 299	1	2
100 – 199	4	10
90 – 99	2	4
80 – 89	1	4
70 – 79	1	3
60 – 69	4	6
50 – 59	2	8
40 – 49	14	12
30 – 39	10	9
20 – 29	23	28
10 – 19	74	88
9	16	14

Interval of frequencies	Number of word forms	Number of lemmas
8	18	13
7	26	40
6	2	4
5	66	69
4	94	83
3	168	165
2	531	414
1	3575	1586

Tab. 1. Results of the frequency distribution of lemmas and word forms

The table shows that the text of “Отпыс” is characterized by 1 word form and 1 lemma with the frequency in the range of 500–999; 1 word form and 2 lemmas with the frequency of 200–299; 4 word forms and 10 lemmas with the frequency of occurrence 100–199, etc.

It is worth noting that there are no word forms or lemmas with the frequencies of 300–399 and 400–499. The largest number of lemmas and word forms has the frequency of usage 1, 2 and 3 times (see Figure 2).

To illustrate this, examples of the lemmas and word forms according to the frequency of their usage are provided below. The most frequent word forms are *и/i* ‘and’ (573), *не/ne* ‘not’ (280), *в(ъ)/v* ‘in’ (169), *на/na* ‘on’ (151), *а/a* ‘and’ (147), *до/do* ‘to’ (108), *в(т)/ot* ‘from’ (93), *ко/jako* ‘as, like’ (93), *але/ale* ‘but’ (85), *же/že* ‘that’ (70), *то/to* ‘that’ (69), *бы/by* ‘would’ (67), *што/što* ‘what, that’ (64), *его/jeho* ‘him’ (62), *ижъ/iž* ‘that’ (59), *або/abo* ‘or’ (51), *такъ/tak* ‘so’ (49), *и/i* ‘and’ (48), *по/по* ‘by’ (48), *за/za* ‘for, with, in, according to’ (46), *в/v* ‘in’ (46), *хто/čhto* ‘who’ (45), *са/са* ‘self’ (43), *ихъ/ich* ‘their, them’ (42), *църкѣви/cerkvi* ‘church’ (41), *з(ъ)/z* ‘with’ (41), *с/s* ‘with, from’ (40), *о/o* ‘oh’ (40), *есть/jest* ‘to be’ (40), *если/jesli* ‘if’ (40) and others.

The most frequent lemmas are *и/i* ‘and’ (621), *не/ne* ‘not’ (280), *въ/v* ‘in’ (207), *той/toj* ‘that’ (190), *на/na* ‘on’ (151), *а/a* ‘and’ (147), *который/kotoryj* ‘who, which’ (142), *был/byl* ‘was’ (138), *свій/svij* ‘own’ (121), *они/ony* ‘they’ (111), *до/do* ‘to’ (108), *что/čto* ‘that, what’ (107), *онъ/on* ‘he’ (101), *оний/onyj* ‘that’ (97), *ко/jako* ‘as, like’ (97), *в(т)/ot* ‘from’ (93), *есть/jest* ‘to be’ (91), *згода/zhoda* ‘agreement’ (85), *але/ale* ‘but’ (85), *ижъ/iž* ‘that’ (84), *църкѣвь/cerkov* ‘church’ (82), *о/o* ‘oh’ (79), *з(ъ)/z* ‘with’ (71), *же/že* ‘that’ (70), *то/to* ‘that’ (69), *бы/by* ‘would’ (67), *богъ/boh* ‘God’ (65), *такъ/tak* ‘so’ (64), *всеъ/vsi* ‘all’ (64), *божій/božij* ‘God’s’ (62), *хто/čhto* ‘who’ (56), *свѣтый/sviatyj* ‘holy’ (56), *за/za* ‘for, with, in, according to’ (56) etc.

Lemma	Lemma's frequency	Lemma's word forms	Word forms frequency
БО 'because'	39	БО	39
ГДЕ 'where'	38	ГДЕ ГДѢ	34 4
СЛОВО 'word'	37	СЛОВО СЛОВА СЛОВѢ СЛОВО(М) СЛОВЫ СЛОВЪ СЛОВАХЪ СЛОВА(Х)	8 18 1 2 3 3 1 1
МИЛОСТЬ 'grace'	36	МИЛОСТЬ МИЛОСТИ МИЛОСТЬЮ	15 20 1
ПЕТРЪ 'Peter'	34	ПЕТРЪ ПЕ(Т)РЪ ПЕТРА ПЕТРОУ ПЕТРОМЪ ПЕТРОВИ ПЕТРОВЪ	22 2 4 1 1 2 2

Tab. 2. Fragment of the frequency dictionary

The results of the analysis showed that among word forms with the frequency of 40 and higher, auxiliary parts of speech, pronouns (*хто/čto* 'who', *ихъ/ich* 'their, them'), linking verb (*єсть/jest* 'to be') and nouns (*църкѣви/cerkvi* 'church') prevail.

Among the most frequent lemmas, there are various parts of speech: in addition to the auxiliary parts of speech, there are also pronouns (*той/toj* 'that', *онъ/on* 'he', *мы/my* 'we'), adjectives (*божій/božij* 'God's', *святый/sviatyj* 'holy'), nouns (*църкѣвъ/cerkov* 'church'), numerals (*один/odyn* 'one'), copular verb (*єсть/jest* 'to be').

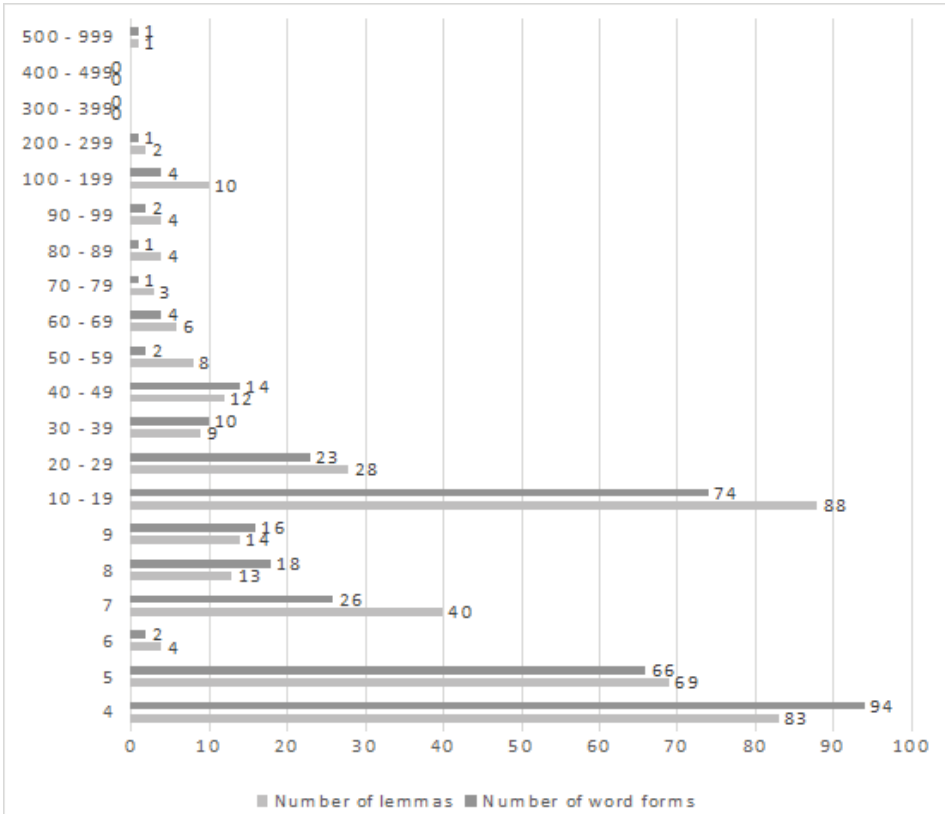


Fig. 1. Graphic representation of the distribution of lemmas and word forms according to the frequency decrease ranged from 999 to 4 based on Table 1

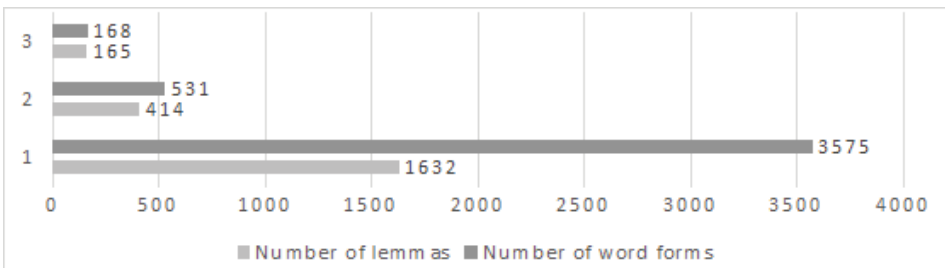


Fig. 2. Graphic representation of the distribution of lemmas and word forms according to the frequency decrease ranged from 3 to 1, according to Table 1

The next step was the calculation of ratio between word ranking and the volume of text which they cover (C). Further analysis requires to define the “ranking” of

a lemma. A lemma with the first ranking has the highest frequency, a lemma with the second ranking is the one following, based on the frequency decrease, and so on.

$$C = \frac{F}{N} \times 100\%;$$

$$F = F_1 + F_2 + \dots + F_n = \sum_{k=1}^n F_k$$

where **n** is the ranking of a lemma, **F** is the sum of frequencies from the 1st to the n-ranking.

For example, if $n = 1$, **F** is equal to the frequency of lemma with the first ranking, that is, the lemma with the highest frequency; if $n = 2$, then **F** equals the sum of frequencies of the first and second rankings; if $n = 3$, then **F** is equal to the sum of frequencies of the first, second and third rankings, and so on.

Number of first rankings (n) in calculation of the sum of frequencies (F)	% of coverage, (C)
1	6
5	13
10	18
25	30
50	41
75	48
100	52
200	63
300	68
400	72
500	75
1000	85
1500	89
2000	94
2640	100

Tab. 3. Percentage of text coverage by lemmas by sequences of rankings

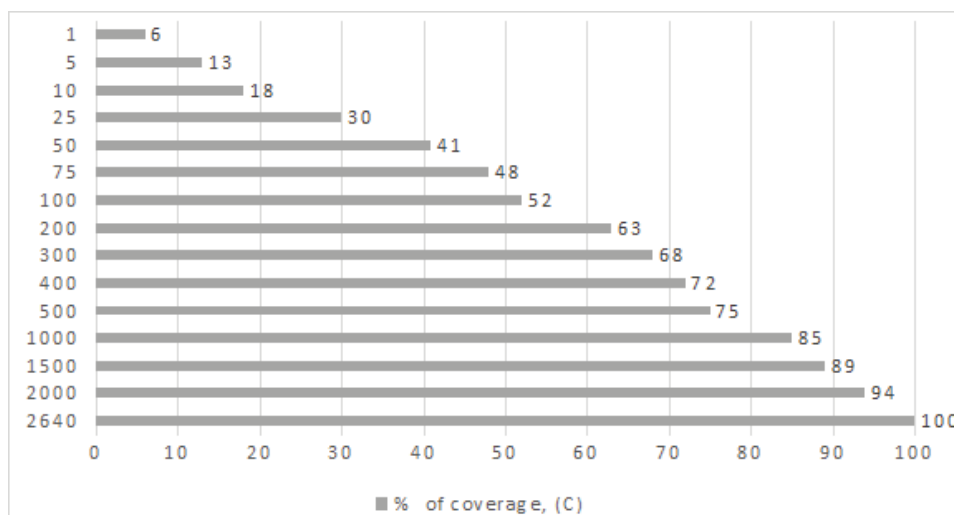


Fig. 3. Graphic representation of percentage of text coverage by lemmas by sequence of rankings according to Table 2

The table and the figure show that *i* ‘and’ as a lemma with the first ranking, covers 6% of the text. The calculations were carried out as follows: a lemma with the first rankings *i* ‘and’ has the frequency of usage 621, that is:

$$F = F_1 = 621.$$

the percentage of text coverage (C) with the first ranking lemma is:

$$C = \frac{F}{N} \times 100\% = \frac{621}{11\ 222} \times 100\% = 5,53\ \% \approx 6\%$$

Calculation of the percentage of text coverage with the lemmas ranking from 1 to 5, namely: first ranking lemma *i* ‘and’ with the frequency of 621 (F_1), second ranking lemma *не/ne* ‘not’ with the frequency of 280 (F_2), third ranking lemma of *въ/v* ‘in’, with the frequency of 207 (F_3), fourth ranking lemma *той/toj* ‘that’ with the frequency of 190 (F_4), fifth ranking lemma *на/na* ‘on’ with the frequency of 151 (F_5). The sum of their frequencies (F) is equal to:

$$F = F_1 + F_2 + F_3 + F_4 + F_5 = 621 + 280 + 207 + 190 + 151 = 1449$$

So, the text coverage with the first 5 ranking lemmas is:

$$C = \frac{F}{N} \times 100\% = \frac{1\ 449}{11\ 222} \times 100\% = 12,9\ \% \approx 13\%$$

Therefore, the lemmas *i* ‘and’, *не/ne* ‘not’, *въ/v* ‘in’, *той/toj* ‘that’, *на/na* ‘on’ cover 13% of the text.

5 CONCLUSIONS

Thus, the article presents the algorithm of creating the frequency dictionary of the Cyrillic written monument of the 16th century, the analysis of the text “Otpys” by Kliryk Ostrozkyi according to the frequency of its elements. The research resulted in creation of the Frequency Dictionary of Word Forms of “Otpys” by Kliryk Ostrozkyi according to the frequency decrease and the Frequency Dictionary of “Otpys” by Kliryk Ostrozkyi according to the frequency decrease.

Research perspectives lie in the ambit of developing frequency dictionaries on the basis of other written monuments of that time and comparison of the obtained results according to the frequency of the usage of words, analysis of the obtained results in lexicology, morphology, stylistics, corpus linguistics and computer lexicography.

References

- [1] Buk S., and Rovenčák A. (2007). Častotnyi slovnyk romanu Ivana Franka “Perechresni stežky”. [Frequency dictionary of Ivan Franko’s novel “Cross Paths”]. In Stežkamy Frankovoho tekstu (komunikatyvni, stylistyčni ta leksyčni vymiry romanu “Perechresni stežky”), pages 138–369, Lviv.
- [2] Častotni slovnyky. [Frequency dictionaries.] Accessible at: <http://www.mova.info/Page.aspx?l1=57>.
- [3] Jongejan, B., and H. Dalianis. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In the Proceeding of the ACL-2009, Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, August 2–7, pages 145–153, Singapore.
- [4] Matvejeva S. A. (2018). Tokenizacija jak sposib obrobky korpusnogo tekstu [Tokenization as a method of text corpus processing] I Mižnarodna naukoivo-prykladna konferencija “Prykladna i korpusna linhvistyka: rozrobleňna tehnolohij novoho pokoliňňa”, pages 35–36, Kyiv.
- [5] Nika O. I. (2016). Dialohizm v istoryčnych vymirach staroukrajinskoho času: “Otpys” Kliryka Ostrozkoho [Dialogism in Historical Dimensions of the Old Ukrainian Time: “Otpys” by Kliryk Ostrozkyi (the Cleric of Ostroh).] Kyiv, Education of Ukraine Publishers.
- [6] Tvorogov O. V. (1984). Leksičeskij sostav “Povesti vremennyh let” (slovoukazateli i častotnyj slovnyk) [The lexical structure of the “The Tale of Bygone Years” (the word pointers and frequency dictionary).], pages 211–217, Kyiv. Accessible at: <http://litopys.org.ua/lavrlet/lavrdod20.htm>.
- [7] Tvorogov O. V. (1967). O primenenii častotnyh slovarj v istoričeskoj leksikologii russkogo jazyka [About the usage of frequency dictionaries in the historical lexicology of the Russian language.] In Voprosy jazykoznanija, 2, pages 109–117.
- [8] Tvorogov O. V. (1961). O tipe slovaria k literaturnomu pamiatniku [The type of dictionary to the literary monument.] In Vestnik Leningradskogo universiteta, 14(3). Serija istorii, jazyka i literatury, pages 119–129.

- [9] Tvorogov O. V. (1962). Slovar'-komentarij k "Povesti vremennyh let" [Dictionary-commentary on the "Tale of Bygone Years"] Author's abstract. dis. candidate of Philological Sciences, 20 p.
- [10] Tvorogov O. V. (1995). Gapaksy "Slova" [Hapaxes of "Word"] In Enciklopedija "Slova o polku Igoreve", In 5 Vol., Vol. 2. G – I, p. 12–15, St. Petersburg. Accessible at: <http://feb-web.ru/feb/slovinc/es/es2/es2-0121.htm>.

KINSHIP TERMINOLOGY IN WESTERN SLAVIC LANGUAGES BASED ON CORPORA ANALYSIS

JANA KOBZOVÁ

Department of Slavonic Studies, Faculty of Arts, Masaryk University, Czech Republic

KOBZOVÁ, Jana: Kinship terminology in Western Slavic languages based on corpora analysis. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 289 – 298.

Abstract: This paper is discussing kinship arrangements and more generally families of Western Slavs based on linguistic and corpora data. It is argued here that we can find correlation between lexicon and society, and that studying of lexicon can provide supportive data for society examination. In this paper we used corpora data that provides us with reliable information about lexicon that is truly used by speakers of Western Slavic languages and provided possible explanations for changes occurring in this part of vocabulary. Paper is divided into three main parts, one discussing relations between social reality and kinship terminology, while the second is discussing data from corpora. Third part is devoted to drawing conclusions.

Keywords: kinship terminology, corpora linguistics, social reality, family, Western Slavic languages

1 INTRODUCTION

Research in lexicology has experienced a great shift with evolution of language corpora. Linguists does not need to blindly believe in dictionaries and language atlases anymore – they can instantly explore the written or spoken language of specific era and find out more about the real usage of lexemes.

This is the case of this paper as well. Research in kinship terminology tends to stick to assumptions about stable character of this part of lexicon and does not explore the patterns of change. Based on the study of Western Slavic languages we must conclude that change is clearly visible in naming relatives in these languages and that further changes will follow in the future.

This paper is based on examination of language corpora of Slovak, Czech and Polish, and contrasting the results with the standard, inherited kinship-terminology lexicon. Languages are being put into contrast with the reconstructed Proto-Slavic language and with each other as well.

The aim here is to challenge traditional views of kinship terminology, to define parts of the lexicon that are most prone to change, and to outline possible future changes to it. Our starting assumption is that kinship terminology is being simplified

in the modern era, based on the change of family (from multigenerational to nuclear) and society itself (increasing level of non-married couples, children born to such relations, increasing level of divorces and more).

Article is divided into three main parts, one focusing on connections between social reality and kinship terminology, while the second part discusses kinship terminology in Western Slavic languages in more detail. This second part is then divided into three subchapters, one devoted to ancestors, one to coeval relatives and one to descendants. The last chapter tries to draw conclusions.

2 KINSHIP TERMINOLOGY AND SOCIAL REALITY

If we want to contrast kinship terminology with social reality, we need to focus on the possible links between these two objects for a moment. Kinship terminology surely is very stable. Even a short look at the lexicon will show us many similarities between Indo-European languages (e.g. English *mother*, Slovak *matka* and Italian *madre* are very similar and can be traced to one Proto-Indo-European etymon). Even the general assumption among many linguists is that social reality is not mirrored in the language and therefore kinship terminology does not change with changed social conditions. As stated by Thomas R. Trautmann, “Joining the anthropological study of kinship terminology with a rich historical record leads us to think that the structures of kinship terminology may be very slow to change and resistant to effects of changed political, economic or social circumstance [...]” [1]

On the other hand, the fact that many of the terms used in kinship terminology are still evident in the modern Indo-European languages (and therefore are used for several thousand years), however, does not imply that these cannot change. Even the assumption that political, economic and social circumstance does not mirror in terminology is not enough. In fact, we can see many terms being replaced, deleted or even being added to the lexicon. The real question therefore is not whether this is happening, but how and why, and whether this is really linked to the changing social arrangements.

For example, in studying and reconstructing Proto-Indo-European society, we are left with no more than reconstructed language, where kinship terms play a significant role. According to J. P. Mallory and Douglas Q. Adams, “the reconstructed lexicon offers us our best hope of glimpsing the world of the speakers of Proto-Indo-European,” [2] what might be the reason for some scientists giving a bigger credit to terminology when reconstructing society. Heinrich Hettrich agrees with this: “Therefore, if the PIE kinship terminology, or at least parts of it, can be reconstructed, we can also conversely expect to obtain some indication about the social organization of the speakers.” [3] This utilitarian attitude is shared by author of this article as well, with two reservations. First, the gained data (in this case terminology systems) should not be viewed as the only source of society

reconstruction (explanation), as there might be more reasons for terminology change than the change in society. We do believe, however, that this data can provide us with interesting theories that should be verified (or falsified) in next stages of research. Second, we should not expect the society to have immediate effect on the vocabulary. As said before, kinship terminology is very stable and therefore a change in society that would influence the terminology must be a major one that is ongoing for several generations. As said by Sergey Kullanda, to think that “kinship terms reflect synchronous social conditions [...] would be absurd.” [4] Having in mind these two reservations, looking at the changed systems of Slavic kinship terminology might provide interesting insights into changing society patterns.

3 KINSHIP TERMINOLOGY IN WESTERN SLAVIC LANGUAGES

3.1 Ancestors in Western Slavic languages

There is not much in the terminology of parents and grandparents to prove changes in social structure of the areas in question – these languages follow the old Indo-European and later Proto-Slavic patterns. In the areas of parents, the only change can be traced in wide-spread usage of terms *tato*, *táta*, *tata* and derivations, their incidence in corpora however makes only up to one quarter of the less affectionate terms *otec*, *ojciec*.

Very interesting (even though not for social structure) is changing nature of Slovak and partly Polish to two-word descriptive terminology in case of grandparents. There is no evidence for distinguishing grandparents from mother’s and father’s side, even though it is assumed that there might have been such distinction in Proto-Indo-European [5], even if some authors do not agree [3]. The same tendency to two-word terminology is evident in godparent terminology as well as in naming stepparents. This outlines the tendency to replace zero-equation terms with descriptive ones, occurring in more areas, as will be seen below.

The closest relations and terminology used for them seems to be stable very much, however this changes significantly when we examine other relations more deeply. First, let us have a look at aunt and uncle terminology. Original Proto-Slavic terminology strictly distinguished gender of parent’s sibling, consanguinity, as well as gender of the intermedator (agent of this relation). This combination would provide us with eight possibilities in total. In Proto-Slavic we reconstruct term **teta* for both mother’s and father’s side, and possibilities for naming her husband are so diverse that we can expect early dissolution and regional specialization (but we do not assume there would be different names for father’s sister’s husband and mother’s sister’s husband if the term for both these sisters is the same). This undifferentiation in case of parent’s sister causes there are only six terms available in this area of relation left for Proto-Slavic, including term for mother’s brother (**ujb*) and father’s brother (**stryjb*), where names for their wives are clearly derived

from these terms, what can still be evident in Western Slavic languages (e.g. *stryná* as a wife of *stryč* in Slovak).

Nonetheless, comparing theory to corpora evidence shows that this six-term strategy is not valid in Western Slavic languages in full anymore. Even though these languages are still aware of the terms (as can be seen in dictionaries and elsewhere), they might not be aware of the original meaning. As the Czech language clearly shows, the only used term for father's or mother's brother is *stryč*, as occurrence of *ujec* is less than 1% compared to *stryč*. Terms *stryna* and *ujčina* for their wives are marginal (we can expect term *teta* to spread to these meanings as well). On the other hand, the same is true for Slovak and Polish, but with the other term (originally meaning "mother's brother"). The share of *stryj* (Polish) and *stryč* (Slovak) is just 20% (in case of Polish) or 50% (in case of Slovak) of the full naming capacity for this term (terms), while *wuj* and *ujec* (and derivatives) are clearly dominant. Derivates for naming wives of these uncles are marginal here as well. In naming parent's sister's husband, we can conclude, that all the specialized terms are disappearing from the vocabulary as well what might cause speakers to reach out to consanguineal terminology even for this affinal relation.

It is not of big surprise that simplification of kinship terms is ongoing in other fields as well. One of such fields is naming parents in law that originate in Proto-Indo-European and are reconstructed in Proto-Slavic as **svetry* (for husband's father) and **tstb* (for wife's father, even though etymology is not clear [6]), while terms for husband's and wife's mother are again derived from the one for father.

Modern Western Slavic languages (as documented in language corpora) tend to simplify these terms to one term for both husband's and wife's father and one for such mother. As in the terminology for aunts and uncles, we can see, that even here the preferred term is not the same among all the examined languages. Slovak is particularly fond of terms *svokor* (masculine) and *svokra* (feminine), differing in this area from both Czech and Polish giving priority to *tchán* and *tešć* (masculine) and *tchyně* and *teściowa* (feminine). Differences in usage of different terms are great in all three languages. Slovak *test'* and *testiná* does not have higher i.p.m. than 11 for both terms, while *svokor* and *svokra* have i.p.m. more than 2 for masculine form and more than 7 for feminine. Similar situation is in Polish, while in Czech the difference in favour of *tchán* and *tchyně* is even more significant.

Interesting part of vocabulary is without a doubt the one of stepparents. The changing society with increasing number of divorces and new partners of parents raises a question of naming new types of relations. This is well described by Ondrejkoivič and Majerčiková, where among changes in natality, marriage and divorce rate and abortion rate several others are listed – changes in education and work patterns, self-reflexion of women, rising economic standard, liberalization, rationalization, technical advancement or social state system [7]. While terms as *macocha* 'stepmother' and *otčim* 'stepfather' are still applicable, new types of

naming, based on description are gaining more weight. Simple two-word *nevlastná matka* and *nevlastný otec* are accompanied by more detailed description terms as *nový partner mojej mamy* ‘new partner of my mother’ and similar ones, mainly in cases where parent is not married to their new partner. As this is a new territory, it would be unreasonable to draw any conclusions at this time, however, it is useful to have this in mind. We will focus on more emerging terms connected with social changes in the last decades in the following chapters.

3.2 Coevals in Western Slavic languages

Again, there is not much evidence of change in the closest family terms (brother, sister). Although, a whole new world opens when we come to cousin terminology. As is generally perceived, the inconsistency in naming cousins among Slavic languages suggests that these terms were formed only after dissolution of Slavic unity. In Polish we can see ongoing change from descriptive terms (such as *brat cioteczny/stryjeczny/wujeczny* what is attitude like the one held in Serbian for example) to two gendered terms – *kuzyn* and *kuzynka*, obviously take-overs from other European languages (most probably French). Descriptive terms still make around 10 % share of the capacity, but their potential is very low to expect a revival. Even though there are still slight traces of descriptive terms in Czech and Slovak, these are so small that we can perceive the process of change as finished in this area (preferred terms in both these languages are *bratranec* and *sestřenica, sestřenice*), therefore derived terms from names for brother and sister respectively.

All three languages in question share their attitude to naming siblings-in-law. Even though at least two terms of the terminology (PIE **daiwer-* “husband’s brother” and **jenāter-* “husband’s brother’s wife”) are reliably reconstructed in Proto-Indo-European, and though these are still preserved e. g. in (some) Southern Slavic languages, all three examined Western Slavic languages replaced them again with two gendered terms (*švagor* a *švagriná* in Slovak, *švagr* and *švagrová* in Czech and *szwagier* and *szwagierka* in Polish), all clearly take-overs from German, as in f. e. Slovenian. All even despite the fact that we can see possibility of 10 relations names in total here (12 if we distinguish gender of the speaker: husband’s brother and his wife, husband’s sister and her husband, wife’s brother and his wife, wife’s sister and her husband, brother’s wife and sister’s husband – from the point of woman or man). This great simplification is another important point in the research.

Before we get to naming husbands and wives, we would like to point to one quite rarely used term for children’s partner’s parents (*svat* or *swat* for such father and *svatka* or *swatowa* for such mother). Even though their usage is not very high (partly probably because of the distance of this relation), this is still a living part of the vocabulary in Western Slavic languages. This is especially important in the light

of many terms disappearing and being replaced by descriptive ones around. For reasons that will need to be stated in future research, terms used for children's partner's parents seem to be more resistant than some others.

In naming one's partner, we again see a dynamically changing environment. Although names for legally wedded partners stay the same and there is no significant difference between the three Western Slavic languages, the changing society discussed above has an influence here as well. Terms that define not wedded partners, even living in one household and having children together (even though not necessarily) will need to be (re)invented. Official terms, such as Slovak *druh* and *družka* might be challenged by another terms as *partner* (*partnerka*), general *muž* 'man' and *žena* 'women' – possibly accompanied by possessive pronoun, or other terms. Whether they will be take-overs from another language, or we will witness semantic extension, will be decided in the future.

3.3 Descendants in Western Slavic languages

In case of closest family, the terms in Western Slavic languages seems to be untouched again. This is in compliance with the pattern already sketched above with parent and grandparent terminology, as well as sibling terminology. There is one standard term for both daughter, son, granddaughter and grandson, with only diminutives extending the basic vocabulary. All these terms are derived from Proto-Slavic and Proto-Indo-European respectively.

In case of naming partners of children, we cannot find any changes that would point to change of social arrangement. Even though all three languages stick to different terms for daughter-in-law (*snacha* in Czech, continuing the tradition, *nevesta* in Slovak, derived from bride-name, and *synowa* in Polish, deriving from name for son), there is no evidence of changing nature of these relations.

However, in the lateral line a great deal of simplification took place. Even though there are still some traces of distinguishing between brother's and sister's son and daughter in Czech and Slovak, these two languages adopted system of two gendered terms (*synovec* and *neter* or *neteř*), even though each is of different origin (derivation from name for son and Proto-Indo-European origin). On the other hand, Polish still distinguishes all four types of relation (*bratanek* and *bratanica*, *siostrzeniec* and *siostrzenica*).

Finally, the topic of currently changing social conditions and naming partner's children from the previous marriages or relations is reflected in descendant line too. This is the field where descriptive terms gained a huge popularity and almost edged the original zero-equation term (*pastorok* and *pastorkyňa* in Slovak, *pastorek* and *pastorkyně* in Czech). The only active part seems to be Polish, where *pasierb* and *pasierbica* are more frequent than descriptive terms. Interestingly, Polish (and surprisingly Czech) sticks to zero-equation terms also in case of godchildren, where Slovak has fully developed two-word descriptive terms.

4 CONCLUSIONS

Our research has provided several outcomes that we would like to point to at this place. First, nuclear family terminology seems to be untouched and unchanged (terms for parents, grandparents, siblings, children and grandchildren). Second, in the lateral line (aunts and uncles, cousins, nephews) a great deal of simplification has taken place (except of nephew terminology in Polish). Third, affinal terminology seems to undergo different processes – the children-in-law and children's partner's parents terminology is untouched, in case of parents-in-law and siblings-in-law a major simplification has occurred, and in case of stepparents, stepchildren and partner names the situation became more complicated, mainly due to ongoing changes in the society that has not been finished yet.

Based on the corpora data, we can conclude that nuclear family, as the prevalent organization of family, has been natural in the areas of nowadays Poland, Czechia and Slovakia for a longer time that enabled change in the vocabulary. We can also conclude that extended family has lost part of its significance what enabled great deal of simplification of terminology. [8] This is also supported by the fact that some terms, previously reserved to consanguine relations only, are now used even for affinal relations. In the affinal relations, some simplification took place where possible – original inherited two gendered terms (for children-in-law and children's partner's parents) stayed untouched, whereas where there used to be more terms (parents-in-law and siblings-in-law) the simplification to two gendered terms is clearly visible.

Special area seems to be the one linked to current social changes (marriage, divorce, new partners etc.) where new terms are popping out regularly and we cannot predict which one will prevail yet.

When comparing attitudes of Western Slavic languages, all of them show similar patterns. However, Polish seems to be the most stable or even most conservative, what is shown on the nephew and children-in-law terminology. This opposes the general view of Polish as the most prone to branch away from fusional type of language [9], what should be characterized – among other things – by a rise of two-word descriptive terms.

It is thanks to language corpora data that we can outline Western Slavs societies here, as a simple search for the terms in dictionaries and in kinship terminology lists would not represent all the changes that we described above.

Full data

Table below does not include those terms that were not found in the corpora. In some places alternative meanings might be included (e.g. not just *druh* as one's partner, but also as "type") – these are in italic. Also, descriptive terms including more than two words are not included (e.g. *nový partner mojej mamy*).

All the used corpora were accessed via Sketch Engine application, where the most recent and most extensive corpus was selected for each of the languages (Slovak Web 2011 – skTen11, Czech Web 2017 – csTenTen17, Polish Web 2012 – plTenTen12) [10], [11].

Term	Slovak	i.p.m.	Czech	i.p.m.	Polish	i.p.m.
Father	otec	230.63	otec	110.00	ojciec	192.6
	tato	32.03	táta	23.88	tata	23.09
	tatino	1.51	taťka	4.69	tato	20.12
Grandmother	stará mama	3.32	babička	44.54	babcia	38.87
	stará matka	0.37	stařenka	1.58	staruszka	4.82
	stará mať	0.24			stara matka	0.04
	babka	21.5			stara mama	0.01
	babička	8.97				
	babina	0.47				
	starká	3.41				
Grandfather	dedko	9.08	děda	12.91	dziadek	28.42
	dedo	9.61	dědeček	10.44	dziad	5.66
	starý otec	6.85	děd	5.57	staruszek	5.05
	starký	2.39	stařeček	0.68	stary ojciec	0.13
Aunt	teta	17.14	teta	12.85	ciotka	7.62
	stryná	0.07	stryna	<0.01	stryjenka	0.09
	ujčína	0.05	ujčina	<0.01	wujenka	0.04
Uncle	strýko	6.46	strejda	3.61	stryj	1.17
	strýc	0.63	strýc	6.06	stryjek	0.34
	ujó	12.00	ujec	0.07	wuj	1.93
	ujko	0.41			wujek	7.16
	ujec	0.51				
Father-in-law	svokor	2.04	tchán	2.23	teść	2.57
	teš'ť	0.11			świekr	0.01
Mother-in-law	svokra	7.68	tchyně	2.13	teściowa	7.4
	testiná	0.07	švekruše	<0.01	świekra	0.04
Stepmother	macocha	1.26	macecha	0.91	macochą	0.18
	nevlastná matka	0.09	nevlastní matka	0.21	przybrana matka	0.03
	nevlastná mama	0.02	nevlastní máma	0.04		
Stepfather	otčím	0.47	otčím	0.50	ojczym	1.14
	nevlastný otec	0.48	nevlastní otec	0.21	przybrany ojciec	0.05
			nevlastní táta	0.02		
Godmother	kmotor	0.55	kmotr	9.81	kum	0.99
	krstný otec	2.48	křestní otec	<0.01	kmotr	0.01
					ojciec chrzestny	2.06
Godfather	kmotra	0.50	kmotra	1.81	kuma	0.79
	krstná mama	0.39	křestní matka	<0.01	kmotra	<0.01
	krstná matka	0.07			matka chrzestna	0.46

Term	Slovak	i.p.m.	Czech	i.p.m.	Polish	i.p.m.
Cousin (male)	bratranec	5.35	bratranec	4.59	brat cioteczny	0.31
	strýčny brat	<0.01	strýčenec	<0.01	brat stryjeczny	0.15
			tetěnc	<0.01	brat wujeczny	0.01
					kuzyn	8
Cousin (female)	sesternica	3.61	sestřenice	2.23	siostra stryjeczna	0.01
	strýčna sestra	<0.01	tetěnice	<0.01	siostra cioteczna	0.12
					siostra wujeczna	<0.01
					kuzynka	5.13
Co-parents- in-law	svat	1.25	svat	0.31	swat	0.91
	svatka	0.09	svatka	0.30	swatka	0.19
Brother-in- law	švagor	2.89	švagr	2.13	szwagier	4.05
			švára	0.14	dziewierz	0.01
Sister-in-law	švagriná	1.44	švagrová	1.72	szwagierka	0.83
	švagrinka	0.05	zelva	0.02	jątrew	<0.01
	zolvica	0.01				
Husband (partner)	manžel	91.89	manžel	93.93	moj/twoj/jej mąż	5.16
	môj muž	5.71	můj/tvůj/ její muž	4.66	małżonek	8.36
	partner	154.02	partner	114.79		
	druh	259.97	druh	273.79		
Wife (partner)	manželka	80.18	manželka	58.84	moja/twoja/ jego żona	11.76
	moja žena	7.28	moje/tvoje/ jeho žena	8.65	małżonka	9.01
	partnerka	13.37	partnerka	12.70		
	družka	2.07	družka	1.87		
Godson	kmotrenec	<0.01	kmotřenec	0.11	syn chrzestny	0.01
	krstňa	0.09			chrześniak	0.49
	krstný syn	0.09				
Goddaughter	kmotrenka	<0.01	kmotřenka	0.07	córka chrzestna	0.01
	krstná dcéra	0.03			chrześnica	0.23
					chrześniaczka	0.06
Stepson	pastorok	0.45	pastorek	1.18	pasierb	0.65
	nevlastný syn	0.23	nevlastní syn	0.10	przybrany syn	0.03
Stepdaughter	pastorkyňa	0.12	pastorkyně	0.11	pasierbica	0.22
	nevlastná dcéra	0.09	nevlastní dcera	0.26	przybrana córka	0.02

Tab. 1. Full data used for this paper

References

- [1] Trautmann, T. R. (2001). The Whole History of Kinship Terminology in Three Chapters. *Anthropological Theory* 1(2), pages 268–287.
- [2] Mallory, J. P., and Adams, D. Q. (2006) *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. Oxford, Oxford University Press.
- [3] Hettrich, H. (1985). Indo-European Kinship Terminology in Linguistics and Anthropology. *Anthropological Linguistics* 27(4), pages 453–480.
- [4] Kullanda, S. (2002). Indo-European “Kinship Terms” Revisited. *Current Anthropology*, 43(1), pages 89–111.
- [5] Friedrich, P. (1966) Proto-Indo-European Kinship. *Ethnology* 5(1), pages 1–36.
- [6] Králik, L. (2016). *Stručný etymologický slovník slovenčiny*. Bratislava, Veda.
- [7] Ondrejko, P., and Majerčíková, J. (2006). Zmeny v spoločnosti a zmeny v rodine – kontinuita a zmena. *Slovak Sociological Review* 38(1), pages 5–30.
- [8] Parkin, R. (2015). Indo-European Kinship Terminologies in Europe: Trajectories of Change. *Journal of the Anthropological Society of Oxford* 7(2), pages 205–233.
- [9] Lotko, E. (1986). *Čeština a polština v překladatelské a tlumočnické praxi*. Ostrava, Profil.
- [10] Jakubiček, M. et al. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference CL 2013*. Lancaster, 2013, pages 125–127.
- [11] TenTen Corpus Family. Sketch Engine. Available at: <https://www.sketchengine.eu>.

GENDER-SPECIFIC ADJECTIVES IN CZECH NEWSPAPERS AND MAGAZINES

ADRIAN JAN ZASINA

Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

ZASINA, Adrian Jan: Gender-specific adjectives in Czech newspapers and magazines. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 299 – 312.

Abstract: This study is one of the few studies dealing with gender in the Czech language using corpus methods. It focuses on the issue of gender in Czech journalistic texts from the years 2010–2014. The main goal was to investigate the extent of stereotypical images of men and women in the press. This analysis is based on adjectival collocations of the lexemes *muž* ‘man’ and *žena* ‘woman’ and their semantic categorization. The research uses a journalistic part of the SYN2015 corpus. First, gender-specific adjectival collocates were identified. Second, adjectival collocates were classified into semantic categories and analyzed within journalistic genres. This study has shown that certain adjectives tend to co-occur with one of the examined lexemes and project a gender-stereotypical image of men and women within particular journalistic genres. It was confirmed that men are strongly associated with age specification, strength, appearance, and negative situations as a subject of crime, whereas women are related to motherhood, attractiveness, ethnicity, nationality, and are more often seen as victims of crime.

Keywords: gender studies, language and gender, discourse analysis, corpus linguistics, sociolinguistics

1 INTRODUCTION

Using corpus methods to analyze gender in language does not have a long tradition in the Czech environment. Previous studies carried out in this area are of a rather qualitative nature ([8], [10], [23]). Nevertheless, studies using quantitative methods ([21], [9], [26]) have also appeared in recent years. Inasmuch as the corpus analysis of gender in the Czech language has no long-term tradition, the present paper was inspired by a great deal of corpus-based research on gender in English-language discourse ([18], [5], [1], [15], [22], [17], [2]). Moreover, this study is based on earlier pilot analyses of collocation profiles of the nouns *man* and *woman* in Czech ([24], [25]).

The paper focuses on the issue of gender in Czech journalistic texts from the years 2010–2014. In particular, it aims to systematically describe a profile of premodifying adjectives co-occurring exclusively or predominantly with the nouns *muž*, ‘man,’ or *žena*, ‘woman’. It also emphasizes the occurrence of the analysed

adjectival collocations within the journalistic genres, revealing gender stereotypical tendencies.

Gender “stereotypes” typically refer to the identification of the desirable identities of “hegemonic masculinity” and “preferred femininity” [23, pp. 8–10]. According to psychological studies [11, p. 121], men are described as aggressive, tough, and assertive, whereas women are viewed as kind, gentle, warm, and passive. This stereotypical view of men and women has also been noted in the Czech language [10]. However, a complex empirical study considering the collocational patterns of adjectives that co-occur with the lexemes *muž* and *žena* is still missing. Therefore, the objective of the present paper is to fill this gap and propose a comprehensive framework to study gender in the Czech language.

First, I describe the data applied in my research and the research questions. Second, I focus on the analysis of premodifying adjectives collocating with the lexemes *muž* and *žena*. Finally, I summarize the results and conclude.

2 DATA AND RESEARCH QUESTIONS

My research was provided on the material of the SYN2015 corpus ([12], [13]), a collection of contemporary written Czech texts of the last five-year period (2010–2014). SYN2015 is divided at the topmost level into three groups: fiction, non-fiction, and newspapers and magazines (NMG).

In my investigation, I used the NMG subcorpus that contains only journalistic texts with total number of 39,744,419 tokens. The subcorpus is further categorized into eight genres: national press (NTW), regional press (REG), home & garden, hobby (HOU), lifestyle (LIF), tabloids (SCT), sport (SPO), interesting facts (INT), supplements, Sunday magazines (MIX).

In the analysis of data, I used two different interfaces: the KonText [16] and the paradigmatic query interface [7]. The latter made it possible to compile a classification of examined adjectives which collocate exclusively, almost exclusively or predominantly with the lexemes *muž* and *žena*. A detailed discussion concerning the categorisation is presented in paragraph 3.

The main goal of my study is to systematically describe a profile of premodifying adjectives collocating with the nouns *muž* or *žena*¹ and reveal the gender stereotypical tendencies in journalistic texts. I posed three research questions while investigating examined collocations:

- Are certain adjectives exclusive to one of the examined lexemes?
- Do journalistic texts reflect in collocations with the lexemes *muž* and *žena* gender stereotypes?

¹ These two lexemes were chosen as they provide the most generic reference in comparison to more specific terms such as *divka* ‘girl’, *chlapec* ‘boy’, *dáma* ‘lady’, *mládenec* ‘young man’. As my aim is to cover a maximally general view of man and woman in Czech journalistic texts.

- Is there any connection between certain meanings of semantic categories and the concrete text genres in journalistic texts?

Thanks to the two different interfaces, it was possible to apply a quantitative approach to the data (paradigmatic query interface) as well as a more detailed qualitative approach to identify semantic categories of examined adjectives based on a context (KonText). The next section zooms in on the analysis of adjectival collocations.

3 ANALYSIS

The first step in my analysis was to establish the list of “gender-specific” adjectives collocating with the lexemes *muž* and *žena*. By “gender-specific” adjectives, I mean adjectives that co-occur exclusively or predominantly only with a noun referring to a female (ex. *pretty woman*) or male individual (ex. *handsome man*). I am therefore using the terms feminine, masculine, and neutral adjectives in this work in the above described context but not in the meaning of conventional grammatical gender.

Using the paradigmatic query interface, I looked for collocations by lemma. Adjectival collocates were identified within a span of one word on the left side to capture attributive adjectives which mainly go before a noun [6, pp. 303–304]. The minimum collocate frequency was arbitrarily set to five hits.

In the next step, I divided the identified adjectives into four groups according to a scale of gender-specificity. The boundary of gender-specificity for my research was assumed arbitrarily. I considered an adjective to be masculine or feminine when it collocated in more than 70% of cases with one of the examined lexemes over the other. In addition, masculine and feminine adjectives were separated into collocates which exclusively, almost exclusively, and predominantly co-occur with the lexemes *muž* or *žena*. The remaining words formed a group of gender-neutral adjectives. The established groups are as follows:

1. exclusively feminine or masculine adjectives – occurring in 100% of cases only with the lexeme *muž* or *žena*;
2. almost exclusively feminine or masculine adjectives – occurring in more than 90% and less than 100% of cases with one of the examined lexemes over the other;
3. predominantly feminine and masculine adjectives – occurring in more than 70% and less than 90% of cases with one of the examined lexemes over the other;
4. neutral adjectives – occurring 70% of cases and less with one of the examined lexemes than the other.

The table below presents the adjectival collocates for each group. Only the top 20 results are given for the neutral adjectives. The number in brackets represents the absolute frequency of collocates in the collocation with a noun in NMG.

Category	Feminine	Masculine
Exclusively (100%)	těhotný (184) 'pregnant', praktický (51) 'practical', vdaný (29) 'married', emancipovaný (15) 'emancipated', kočičí (11) 'cat-like', nádherný (9) 'gorgeous', půvabný (9) 'graceful', lesbický (8) 'lesbian', lehký (7) 'loose', rodičí (6) 'giving birth', vystrašený (6) 'frightened', rázný (6) 'spirited', bezbranný (6) 'defenseless', obřezaný (5) 'circumcised', indiánský (5) 'Indian', inspirativní (5) 'inspirational', sociálnědemokratický (5) 'social-democratic'	klíčový (40) 'key', ženatý (28) 'married', ledový (21) 'ice', maskovaný (19) 'masked', železný (16) 'iron; strong', netopýří (15) 'bat-like', pavoučí (13) 'spider-like', 30letý (12) '30-year-old', otevřený (11) 'open', zmrzlý (9) 'frozen', trestaný (8) 'punished', čestný (8) 'honest', urostlý (8) 'shapely', smutný (8) 'sad', homosexuální (7) 'homosexual', holohlavý (6) 'bald', svalnatý (6) 'muscular', hlavní (6) 'head', zavalený (6) 'collapsed', respektovaný (6) 'respected', zlomený (6) 'broken', sedmadesátiletý (6) '77-year-old', 69letý (5) '69-year-old', prošeďivělý (5) 'graying', galantní (5) 'gallant', bělovlasý (5) 'white-haired', 56letý (5) '56-year-old', 43letý (5) '43-year-old'
Almost exclusively (> 90%)	kojící (27) 'breastfeeding', týraný (25) 'abused', znásilněný (14) 'raped', padlý (12) 'fallen', ambiciózní (10) 'ambitious'	svatý (34) 'holy', třiatřicetiletý (32) '33-year-old', agresivní (28) 'aggressive', ozbrojený (26) 'armed', rozhněvaný (14) 'angry', 34letý (14) '34-year-old', statný (13) 'burly', uniformovaný (12) 'uniformed', 29letý (11) '29-year-old', nenápadný (10) 'discreet', oběšený (10) 'hanged'
Predominantly (> 70%)	krásný (173) 'beautiful', bývalý (68) 'ex', sebevědomý (31) 'confident', obyčejný (23) 'ordinary', zavražděný (14) 'murdered', tehdejší (13) 'former', normální (11) 'casual', pracující (10) 'working', rozvedený (10) 'divorced', tamní (10) 'local', aktivní (9) 'active', hezký (9) 'pretty', kolemjdoucí (9) 'passersby', zahalený (9) 'veiled', úžasný (8) 'amazing', elegantní (7) 'elegant', oslovený (7) 'addressed', samotný (6) 'lonely', evropský (5) 'European', konkrétní (5) 'major league', ruský (5) 'Russian', saúdský (5)	čtyřicetiletý (43) '40-year-old', obviněný (41) 'accused', podezřelý (39) 'suspect', devětatřicetiletý (32) '39-year-old', šedesátiletý (30) '60-year-old', velký (30) 'big', bílý (28) 'white', čtyřiatřicetiletý (28) '34-year-old', hledaný (27) 'wanted', pětadvacetiletý (27) '25-year-old', osmadvacetiletý (26) '28-year-old', pohřešovaný (24) 'missing', důležitý (23) 'important', pětapadesátiletý (23) '55-year-old', zadržený (20) 'detained', devětadvacetiletý (18) '29-year-old', dvačtyřicetiletý (18) '42-year-old', sněžný (17) 'snow', správný (17) 'right', třiačtyřicetiletý (17) '43-year-old', dvaapadesátiletý (16) '52-year-old', sedmapadesátiletý (16) '57-year-old', osmačtyřicetiletý (15) '48-year-old', tajemný (15) 'mysterious', čtyřiačtyřicetiletý (14) '44-year-old', 35letý (14) '35-year-old', šestapadesátiletý (12) '56-year-old', 33letý (12)

Predominantly (> 70%)	'Saudi', talentovaný (5) 'talented'	'33-year-old', 41letý (12) '41-year-old', bezpečný (11) 'safe', jednadvacetiletý (11) '21-year-old', ležící (11) 'lying', podnapilý (11) 'tipsy', třiapadesátiletý (11) '53-year-old', jedenatřicetiletý (10) '31-year-old', vousatý (10) 'beared', 27letý (10) '27-year-old', 38letý (10) '38-year-old', 45letý (10) '45-year-old', záhadný (9) 'mysterious', 44letý (9) '44-year-old', hovořící (8) 'speaking', charismatický (8) 'charismatic', spící (8) 'sleeping', tvrdý (8) 'tough', 36letý (8) '36-year-old', heterosexuální (7) 'heterosexual', pětasedmdesátiletý (7) '75-year-old', sledovaný (7) 'watched', zdatný (7) 'able-bodied', 32letý (7) '32-year-old', 49letý (7) '49-year-old', 62letý (7) '62-year-old', 40letý (6) '40-year-old', 48letý (6) '48-year-old', stoletý (5) '100-year-old', vhodný (5) 'suitable', 23letý (5) '23-year-old', 53letý (5) '53-year-old', 59letý (5) '59-year-old', 66letý (5) '66-year-old'
Neutral adjectives – Top 20	mladý (1067) 'young', starý (407) 'old', známý (208) 'famous', jiný (183) 'other', český (160) 'Czech', další (137) 'next', mrtvý (116) 'dead', jediný (113) 'only', opilý (102) 'drunk', zraněný (100) 'injured', mocný (94) 'powerful', zlý (89) 'evil', dospělý (89) 'adult', bohatý (74) 'rich', dobrý (73) 'good', nový (69) 'new', nemocný (67) 'ill', vysoký (67) 'tall', cizí (65) 'foreign', silný (64) 'strong'	

Tab. 1. Feminine, masculine, and neutral adjectives in Czech journalistic texts

It is noticeable that the group of masculine adjectives is constituted of a larger number of collocates than the group of feminine adjectives (100 adjectives vs. 46 respectively), albeit the occurrence of lexemes *žena* and *muž* is comparable, making it respectively 22 949 and 22 397 hits in NMG. The lexeme *žena* even has theoretically more possibilities to collocate with adjectives due to a slightly higher occurrence, however, masculine adjectives are twice as frequent. Analyzing masculine adjectives, men are often described by age (*30letý* '30-year-old', *sedmasedmdesátiletý* '77-year-old'), strength (*železný* 'strong'), appearance (*urostlý* 'shapely', *svalnatý* 'muscular', *statný* 'burly'), adjectives evoking negative (*trestaný* 'punished') and positive (*otevřený* 'open', *čestný* 'honest', *galantní* 'gallant') emotions. On the other hand, almost exclusively and exclusively feminine adjectives are related to motherhood (*těhotná* 'pregnant', *rodící* 'giving birth', *kojící* 'breastfeeding'), attractiveness (*nádherná* 'gorgeous', *půvabná* 'graceful') and adjectives evoking negative emotions or conditions (*vystrašená* 'frightened', *týraná* 'abused', *znásilněná* 'raped', *padlá* 'fallen').

The examined adjectives reveal certain semantic preferences [20, p. 65] that make it possible to categorize them into common semantically related groups of words. Moreover, a few adjectival collocations relate to positive and negative prosody, discussed in more details in paragraph 3.1.

3.1 Sematic categories of gender-specific adjectives

Based on the apparent semantic characteristics of examined adjectives and inspired by Caldas-Coulthard and Moon's study [5, p. 111], it was possible to introduce semantic categories for analyzed collocates. In comparison with Caldas-Coulthard and Moon's categorization, the categorization presented in this study is reduced to the ten following groups: age, strength/supernatural power, appearance/attractiveness, positive and negative character/social/emotional states², maternity, nationality/ethnicity, action, material status, sexual orientation, and others. Collocates from the Table 1 were categorized into semantic groups to the best of my knowledge by considering their individual context and functions. Table 2 below presents feminine and masculine adjectives by semantic categories.

Category	Feminine	Masculine
Age ³		77- (6)/69- (5)/30- (12)/56- (5)/43- (5)/33- (32)/34- (14)/29- (11)/40- (43)/39- (32)/60- (30)/34- (28)/25- (27)/28- (26)/55- (23)/29- (18)/42- (18)/43- (14)/52- (16)/57- (16)/48- (15)/44- (14)/35- (14)/56- (12)/33- (12)/41- (12)/21- (11)/53- (11)/31- (10)/27- (10)/38- (10)/45- (10)/44- (9)/36- (8)/75 (7)/32- (7)/49- (7)/62- (7)/40- (6)/48- (6)/100- (5)/23- (5)/53- (5)/59- (5)/66letý (5) '77-/69-/30-/56-/43-/33-/34-/29-/40-/39-/60-/34-/25-/28-/55-/29-/42-/43-/52-/57-/48-/44-/35-/56-/33-/41-/21-/53-/31-/27-/38-/45-/44-/36-/75-/32-/49-/62-/40-/48-/100-/23-/53-/59-/66-year-old'
Strength and supernatural power	kočičí (11) 'cat'	železný (16) 'iron/strong', netopýří (15) 'bat', pavoučí (13) 'spider', tvrdý (8) 'tough', zdatný (7) 'efficient',

² This category in early study [24] was labelled as *character, psysical state, and adjectives evoking positive/negative emotions*. However, to avoid possible confusion it was relabelled (for the discussion see [26, pp. 173–174]).

³ To save space in the table all age specifications were converted to numeric values.

Appearance and attractiveness	nádherný (9) ‘gorgeous’, půvabný (9) ‘graceful’, krásný (173) ‘beautiful’, hezký (9) ‘pretty’, úžasný (8) ‘amazing’, elegantní (7) ‘elegant’	urostlý (8) ‘shapely’, holohlavý (6) ‘bald’, svalnatý (6) ‘muscular’, prošeďivělý (5) ‘graying’, bělovlasý (5) ‘white-haired’, statný (13) ‘burly’, vousatý (10) ‘bearded’, bílý (28) ‘white’, velký (30) ‘big’
Character, social, and emotional states	positive	klíčový (40) ‘key’, otevřený (11) ‘open’, čestný (8) ‘honest’, respektovaný (6) ‘respected’, hlavní (6) ‘head’, galantní (5) ‘gallant’, svatý (34) ‘holy’, bezpečný (11) ‘safe’, důležitý (23) ‘important’, správný (17) ‘right’, sledovaný (7) ‘watched’, vhodný (5) ‘suitable’, charismatický (8) ‘charismatic’
	negative	trestaný (8) ‘punished’, zlomený (6) ‘broken’, ozbrojený (26) ‘armed’, oběšený (10) ‘hanged’, obviněný (41) ‘accused’, podezřelý (39) ‘suspect’, hledaný (27) ‘wanted’, pohřešovaný (24) ‘missing’, zadržovaný (20) ‘detained’, podnapilý (11) ‘tipsy’, smutný (8) ‘sad’, agresivní (28) ‘aggressive’, rozhněvaný (14) ‘angry’, maskovaný (19) ‘masked’
Maternity	těhotný (184) ‘pregnant’, rodičí (6) ‘giving birth’, kojící (27) ‘breastfeeding’	
Nationality and ethnicity	indiánský (5) ‘Indian’, ruský (5) ‘Russian’, saudský (5) ‘Saudi’, evropský (5) ‘European’	
Action	pracující (10) ‘working’, kolemjdoucí (9) ‘passing by’	ležící (11) ‘lying’, hovořící (8) ‘speaking’, spící (8) ‘sleeping’
Material status	vdaný (29) ‘married’, rozvedený (10) ‘divorced’	ženatý (28) ‘married’
Sexual orientation	lesbický (8) ‘lesbian’	homosexuální (7) ‘homosexual’, heterosexuální (7) ‘heterosexual’

Others	emancipovaný (15) ‘emancipated’, obřezaný (5) ‘circumcised’, sociálnědemokratický (5) ‘social democratic’, bývalý (68) ‘ex’, obyčejný (23) ‘ordinary’, tehdejší (13) ‘former’, normální (11) ‘casual’, tamní (10) ‘local’, oslovený (7) ‘addressed’, zahalený (9) ‘veiled’	ledový (21) ‘ice’, uniformovaný (12) ‘uniformed’, nenápadný (10) ‘discreet’, sněžný (17) ‘snow’, tajemný (15) ‘mysterious’, záhadný (9) ‘mysterious’, zmrzlý (9) ‘frozen’, zavalený (6) ‘collapsed’
--------	--	---

Tab. 2. Feminine and masculine adjectives by semantic categories

While most categories of adjectives modify both *muž* and *žena*, fewer categories modify only one of examined lexemes. The categories of maternity and of nationality and ethnicity are represented only by feminine adjectives. It is obvious that in the case of maternity, the lack of masculine adjectives is conditioned largely by biology and in that case we suppose to talk about sex-specific rather than gender-specific adjectives⁴. In the category of nationality and ethnicity, there is a greater interest in discussions of women in sociological discourse that is commensurate with Pearce’s findings [18, p. 12].

The age-specification category is only represented by masculine adjectives. The close reading has revealed that the adjectives appear predominantly in crime reports and *muž* is more often seen as a perpetrator of crime. These findings are similar to Pearce’s conclusion that “men are more strongly associated with crime, violence and the criminal justice system than woman” [18, p. 9]. On the other hand, Caldas-Coulthard and Moon observed that “adjectives indicating age are common” [5, p.116]; this might be caused by including gender-neutral adjectives in their analysis as well (ex. *young, old* etc.).

The other categories are represented by both feminine and masculine adjectives that show different tendencies. In terms of appearance and attractiveness the feminine adjectival collocates emphasize attractiveness, while the masculine adjectival collocates underline appearance. The same conclusion was made by Pearce [18, p. 17] and Caldas-Coulthard and Moon [5, p. 117]. On the contrary, Baker [1, p. 138] observes in the contemporary English language some changes where “men can be now represented in terms of caring about or looking after their appearance.”

⁴ Buttler [4, pp. 9–11] makes clear binary distinction between the terms *gender* and *sex*. Whereas *sex* is based on biological assumptions and refers to sexual body, *gender* is a social construct set by society.

Regarding the strength and supernatural power category, the adjectives mostly refer to superheroes such as *kočičí žena* ‘Catwoman’, *železný muž* ‘Ironman’, *netopyří muž* ‘Batman’, *pavoučí muž* ‘Spiderman’, direct translations from English. These proper names mainly express the image stereotypes and are closely related to other gender-specific proper nouns such as: Barbie and Ken, Adam and Eve etc. Within this category, the correlation between man and strength was also highlighted in the collocation *železný muž* ‘strong man’.

It is also worth noting that the character, social, and emotional states category reveals semantic prosody ([3], [14], [19]). Examined adjectives have positive or negative prosody and represent different gender specifications. Adjectives with positive connotation describe man mainly as a *key*, *open*, *honest*, *respected*, and *head*, underlining his importance and high social status. Whereas feminine adjectives such as *practical*, *inspirational*, *ambitious*, *confident*, and *active* emphasise the effort of woman who has to constantly prove her abilities. In the term of negative adjectives, further gender stereotypes are revealed. Man is often presented as a perpetrator of crime (*trestaný* ‘punished’, *podezřelý* ‘suspect’) and woman is a victim of violence (*týraný* ‘abused’, *znásilněný* ‘raped’) as is consistent with Pearce’s [18, p. 19] and Zasina’s observations [26, p. 186]. Moreover, there are stressed unfavourable female qualities expressed by euphemisms (*lehký* ‘loose’, *padlý* ‘fallen’), and there is expressed male aggressiveness (*agresivní* ‘aggressive’, *rozhněvaný* ‘angry’).

Furthermore, the sexual orientation category reveals that the adjective *heterosexual* is masculine because in a journalistic discourse both straight and gay men are compared, and it is essential to stress it. Within the material status category *divorced* occurs strongly only with woman, pointing to the focus of discussion in the media.

3.2 Gender-specific collocations within journalistic genres

In this subsection, I focus on the relation between certain meanings of semantic categories and the concrete text genres in NMG. I zoomed in on the frequency distribution of adjectival collocations with the nouns *muž* and *žena* in NMG genres (see Section 2). To compare particular NMG genres, each of which consisted of different number of texts, I used relative frequencies expressed in instances per million (ipm). I took into account all exclusive adjectival collocations from Table 1 and collocations of the two following semantic categories: appearance and attractiveness, and character, social, and emotional states from the Table 2. I chose these two categories because they require a detailed discussion in relation to NMG genres.

First, I examined the exclusive masculine and feminine collocations to investigate general trends in using gender-specific adjectives collocating with *muž* and *žena*. Figure 1 shows the frequency distribution across NMG genres.

The NMG genres differ by gender. Results show that feminine collocations are more often in LIF, HOU, and MIX. The LIF and HOU genres are predominantly intended for female readers; therefore articles topics are targeted at women's issues such as pregnancy, marriage, emancipation etc. Hoffmannová [10] also notices that gender stereotypes are presented in women magazines even though they are predominantly edited by women. Masculine collocations, on the other hand, are the most frequent in SPO, MIX, INT. Sport newspapers and magazines seem to be addressed to men and discuss men's issues evidence of which supported by the low occurrence of feminine collocations. The lowest occurrence of masculine collocations in HOU indicates greater focus on female readers.

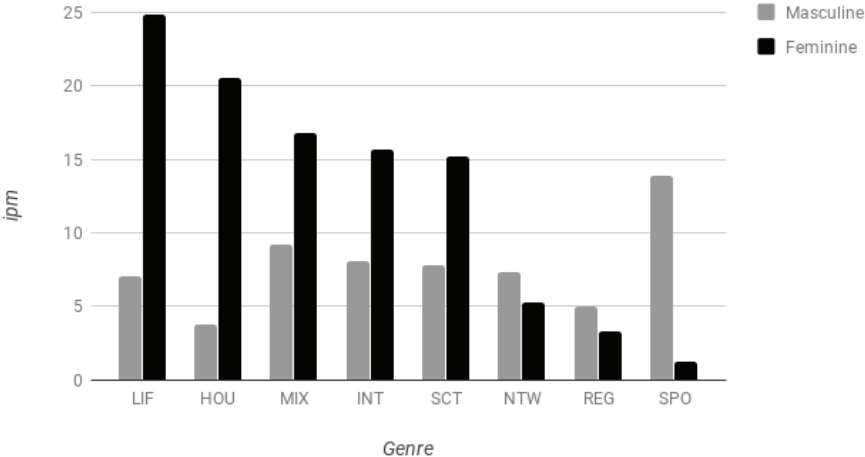


Fig. 1. Distribution of exclusive masculine and feminine adjectives in collocation with the lexemes muž and žena by text genres in NMG

Second, I focus on chosen semantic categories: appearance and attractiveness, and character, social, and emotional states. Figures 2 and 3 show the frequency distribution across NMG genres for each semantic category.

Analysing the data in Figure 2, there is a significantly higher occurrence of feminine collocations within SCT, LIF, and MIX genres. This shows a great interest in female beauty as it is a crucial topic in tabloids and lifestyle magazines. On the other hand, male physical strength and appearance is underlined the most in genres such as INT, SCT, and MIX. Although masculine collocations predominate in INT genre, there is also an interest in women that indicates this genre is dealing with both genders but in a different range. The lowest attention is given to appearance and attractiveness in HOU within both genders and in REG within masculine collocations.

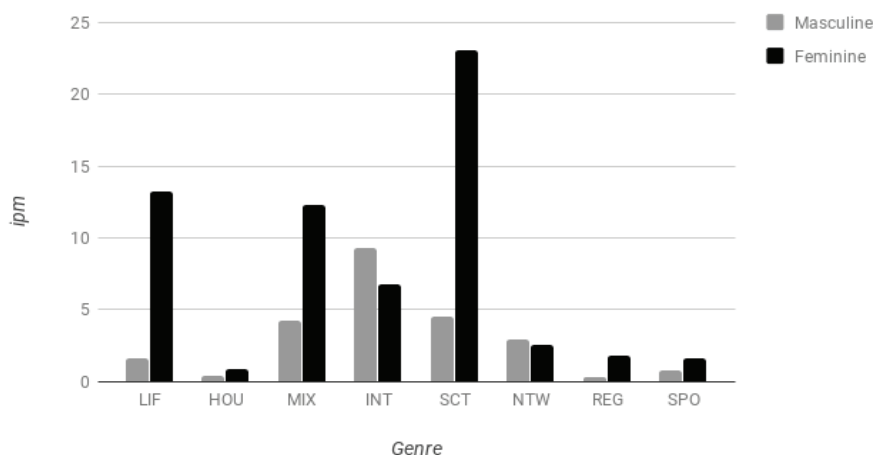


Fig. 2. Distribution of appearance and attractiveness adjectives in collocation with the lexemes *muž* and *žena* by texts genres in NMG

Figure 3 presents the character, social, and emotional states category divided into positive and negative subcategories. The positive masculine collocations are the most represented in INT, and SPO. It is commensurate with my previous statements that these genres are aimed at male readers and describe them mostly in a positive way. In the case of the INT genre, more attention is focused on men than women. On the other hand, the negative masculine collocates are strongly associated with SCT, REG, and NTW. It supports previous findings discussed in subsection 3.1 that negative meaning and men correlate quite often in crime reports. There were no negative masculine collocations in HOU, and the positive ones were rare.

Moreover, analyzing the positive feminine collocations I confirmed that women are strongly seen in a positive light in the HOU and LIF genres. It is notable that these two genres are targeted at a female audience, while the SPO genre is not. Furthermore, the negative feminine collocations are prominent in the SCT and LIF genres. It was expected that tabloids provide for the most part a negative view of women as well as men. Lifestyle magazines seem to build a more positive picture of men and women rather a negative one. There were no negative feminine collocations in HOU genre, which presents only positive qualities.

This section using empirical methods proves that there is a strong relation between gender-specific collocations and journalistic genres. Some genres are more associated with women (LIF, HOU), and others with men (SPO, INT). The analysis of semantic categories confirmed that Czech journalistic texts still project a stereotypical image of men and women. It might be related to readers' expectations of traditional masculine and, feminine roles; however, the further investigation in this area is needed.

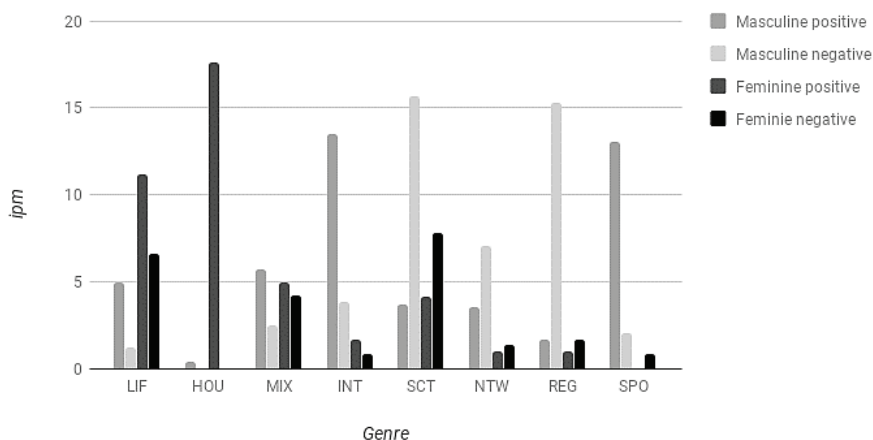


Fig. 3. Distribution of character, social, and emotional states in collocation with the lexemes *muž* and *žena* by texts genres in NMG

4 CONCLUSION

The present study focused on gender-specific adjectives in Czech newspapers and magazines. The analysis of adjectival collocations with the lexemes *muž* and *žena* has confirmed that it is possible to specify adjectives that are exclusively or almost exclusively associated with one of the examined lexemes. Moreover, it has been proven that Czech journalistic texts reflect gender stereotypes in the adjectival collocations. Men pattern strongly with age specification, strength, appearance, and negative situations as a perpetrator of crime. On the other hand, women are portrayed as being mothers, attractive, and victims of crime. Women's nationality or ethnicity is also underlined. Within the positive adjectives of character, social, and emotional states category, both men and women have positive attributes; however, they highlight gender differences. In comparison to the existing literature, this study shares findings with previous English works while bringing new insight in studying gender in Czech; it establishes a comprehensive methodological framework to analyse gender using corpus methods and identifies the gender-specific adjectives with their semantic categories that might be a base for further analyses.

In my work, I also examined the gender-specific collocations within journalistic genres. The analysis has shown that the exclusive masculine and feminine collocations occur in different genres that project stereotypical roles of men and women. The feminine collocations are much more frequent in lifestyle and hobby magazines, whereas the masculine collocations are more frequent in sport magazines and newspapers. Moreover, it was confirmed that there is a connection between certain meanings of semantic categories and concrete journalistic genres. Attractiveness of

women is for the most part stressed in tabloids and lifestyle magazines, while the appearance of men is underlined in magazines about interesting facts. Further, the investigation of positive and negative character, social, and emotional states has revealed that women are strongly positively portrayed in lifestyle and home & garden, and hobby magazines, whereas men are predominantly represented in a positive way in magazines about interesting facts, and sport magazines and newspapers. As for negative meaning, both men and women were associated with tabloids. Additionally, men also had negative connotation in regional press.

Contemporary Czech journalistic texts still present a stereotypical image of man and woman. The present study contributes new insights into previous qualitative gender studies ([8], [10], [23]) and complements previous corpus-based gender analyses ([21], [9], [26]). However, it is necessary to carry out further research in this area, which ought to tell us more about the character of man and woman in the Czech language.

References

- [1] Baker, P. (2010). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language*, 4(1), pages 125–149.
- [2] Baker, P. (2014). *Using corpora to analyze gender*. London, Bloomsbury.
- [3] Bednarek, M. (2008). Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory*, 4(2), pages 119–139.
- [4] Butler, J. (1999). *Gender trouble: Feminism and the subversion of identity*. London, Routledge.
- [5] Caldas-Coulthard, C. R., and Moon, R. (2010). ‘Curvy, hunky, kinky’: Using corpora as tools for critical analysis. *Discourse & Society*, 21(2), pages 99–133.
- [6] Cvrček, V. et al. (2010). *Mluvnice současné češtiny*. Praha, Karolinum.
- [7] Cvrček, V. (2017). Paradigmatické korpusové dotazy a moderní diachronie. In M. Stluka & M. Škrabal (eds.), *Ljčka a czban – Sborník příspěvků k 70. narozeninám prof. Karla Kučery*, pages 117–130. Praha, Czech Republic: Nakladatelství Lidové noviny.
- [8] Čmejrková, S. (2003). Communicating gender in Czech. In M. Hellinger, and H. Bußmann (Eds.), *Gender across languages: The linguistic representation of women and men*, pages 27–58. Amsterdam, John Benjamins Publishing Company.
- [9] Elmerot, I. (2017). *These women’s verbs: a combined corpus and discourse analysis on reporting verbs about women and men in Czech media 1989–2015* (Master’s thesis, Stockholm University). Accessible at: https://www.researchgate.net/publication/322539185_These_women’s_verbs_-_a_combined_corpus_and_discourse_analysis_on_reporting_verbs_about_women_and_men_in_Czech_media_1989-2015_Master’s_thesis
- [10] Hoffmannová, J. (2004). Ženy a muži v časopisech pro ženy: Role, perspektivy, výrazové stereotypy. *Stylistyka*, XIII, pages 27–34.
- [11] Huddy, L., and Terkildsen, N. (1993). Gender stereotypes and the perception of male and female candidates. *American Journal of Political Science*, 37(1), pages 119–147.

- [12] Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářiková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondříčka, P., and Zasina, A. J. (2015). SYN2015: reprezentativní korpus psané češtiny. Praha, Ústav Českého národního korpusu FF UK. Available at: <http://www.korpus.cz>.
- [13] Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářiková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondříčka, P., and Zasina, A. J. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2522–2528. Portorož, Slovenia: ELRA.
- [14] Louw, B. (1993). Irony in the text or insincerity in the writer? – The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, and E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair*, pages 157–176. Amsterdam, John Benjamins.
- [15] Macalister, J. (2011). Flower-girl and bugler-boy no more: Changing gender representation in writing for children. *Corpora*, 6(1), pages 25–44.
- [16] Machálek, T., and Křen, M. (2013). Query interface for diverse corpus types. In K. Gajdošová, and A. Žáková (eds.), *Natural language processing, corpus linguistics, e-learning*, pages 166–173. Lüdenscheid, Germany: RAM Verlag.
- [17] Moon, R. (2014). From gorgeous to grumpy: adjectives, age and gender. *Gender, and Language*, 8(1), pages 99–133.
- [18] Pearce, M. (2008). Investigating the collocational behaviour of man and woman in the BNC using sketch engine. *Corpora*, 3(1), pages 1–29.
- [19] Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London, Routledge.
- [20] Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. London, Blackwell.
- [21] Šonková, J. (2011). Genderové rozdíly v mluvené češtině. In F. Čermák (Ed.), *Korpusová lingvistika. Praha 2011–2 Výzkum a výstavba korpusů*, pages 150–165. Praha, ÚČNK – Nakladatelství Lidové Noviny.
- [22] Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora*, 8(1), 81–113.
- [23] Valdrová, J. (2006). *Gender a společnost. Ústí nad Labem, Czech Republic: Univerzita J. E. Purkyně*.
- [24] Zasina, A. J. (2016, November). Adjective collocations with the lexemes muž ‘man’ and žena ‘woman’ in Czech journalistic texts. Paper presented at the Young Linguists’ Meeting in Poznań 2016. Poznań, Poland.
- [25] Zasina, A. J. (2017, July). Premodifying female and male adjectives in journalistic texts. A gender corpus analysis in Czech. Poster presented at Corpus Linguistics 2017. Birmingham, UK.
- [26] Zasina, A. J. (2018). Image of Politicians and Gender in Czech Daily Newspapers. In M. Fidler, and V. Cvrček (eds.), *Taming the Corpus, From Inflection and Lexis to Interpretation*, pages 167–194. Chad, Springer International Publishing.

FROM THE NATIONAL CORPUS OF POLISH TO THE POLISH CORPUS INFRASTRUCTURE

MACIEJ OGRODNICZUK¹ – RAFAŁ L. GÓRSKI² –
MAREK ŁAZIŃSKI³ – PIOTR PEŹIK⁴

¹Institute of Computer Science, Polish Academy of Sciences, Poland

²Institute of Polish Language, Polish Academy of Sciences, Poland

³Institute of Polish Language, University of Warsaw, Poland

⁴Institute of English Studies, University of Łódź, Poland

OGRODNICZUK, Maciej – GÓRSKI, L. Rafał – ŁAZIŃSKI, Marek – PEŹIK, Piotr: From the National Corpus of Polish to the Polish Corpus Infrastructure. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 315 – 323.

Abstract: The National Corpus of Polish emerged as a cumulative result of many years of work on large reference corpora by computer scientists and linguists in Poland. While its impact on research in linguistics, humanities and language technology is unquestionable and highly significant, the construction of the national corpus was halted in 2011. In the paper we call for activating the research community and funding institutions around the construction of a corpus infrastructure with the national corpus at its heart. It is claimed that on the verge of an artificial intelligence revolution the envisaged Polish Corpus Infrastructure would provide reliable language data, combine available resources and allow easy integration of new ones.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 THE NATIONAL CORPUS OF POLISH IN THE CONTEXT OF POLISH CORPUS RESEARCH

The first edition of the National Corpus of Polish (Pol. Narodowy Korpus Języka Polskiego – NKJP; [19]) has found extremely diverse scientific and technological applications. NKJP is still the main reference corpus in lexicography (see e.g. Żmigrodzki et al. [26]), applied linguistics and psycholinguistics (e.g. Riegel et al. [20]) and language modeling (e.g. Mykowiecka et al. [12]). It has been used to boost the accuracy of natural language processing on various tasks, and to develop many tools and resources for Polish such as the Concraft disambiguating tagger (Waszczuk [22]), Polish Dependency Corpus (Wróblewska [25]), Polish Coreference Corpus (Ogrodniczuk et al. [13]), Hask collocation databases (Peżik [15]), SEJF phraseological dictionary (Czerepowicka [1]) or Walenty valence dictionary (Hajnicz et al. [6]). The National Corpus is cited as the basic resource of linguistic research in hundreds of publications. The NKJP search

engines serve more than one million distinct corpus user queries every year, 11% of which originate from outside of Poland. The corpus is used both by national research infrastructures (e.g. CLARIN-PL¹) and in international projects (e.g. PARSEME²).

Parallel to NKJP, a number of independent reference corpora of Polish exist, spanning the period from the early days of the language to the modern era, including the corpus of pre-1500 Old Polish texts (Twardzik and Górski [21]), Electronic Corpus of 17th and 18th century Polish texts (Pol. short Korpus Barokowy, hence KORBA; Gruszczyński et al. [5]), the corpus of the 19th century Polish texts (Derwojedowa et al. [3], Kieraś and Woliński [8]) and the MoncoPL monitor corpus of web-based Polish (Pęzik [18]). However, each of these resources resulted from a separate project and operates independently using custom-made standards of presenting linguistic information in a variety of user interfaces. This fragmentation has naturally given rise to the idea of linking all related corpora through a common federated infrastructure as recently discussed in papers outlining the development of NKJP (Ogrodniczuk et al. [14]) or proposing the Diachronic Corpus of Polish (Król et al. [10]). Similarly, the first steps towards a common representation format for the planned diachronic corpus were recently completed in the Chronoflex project³ aimed at providing a formal model of Polish inflection to represent historical changes in this area. There are also new developments in the area of open-source corpus search solutions, such as the MTAS-based corpus search engine⁴, which has surpassed the capabilities of the PoliQarp engine (Janus and Przepiórkowski [7]) and was successfully deployed as the main search engine for the Corpus of the 19th century Polish⁵, KORBA⁶ and NKJPI1M, the 1-million-token manually annotated subcorpus of NKJP⁷.

All these attempts aimed at unifying existing corpus resources and tools into a common infrastructure intended for synchronic and diachronic research on the Polish language. In the next sections of this paper we elaborate on this concept, outlining plans for the development of a distributed corpus framework under the umbrella name of *The Polish Corpus Infrastructure*; Pol. *Polska Infrastruktura Korpusowa* – PIK. The framework is planned to create a unified platform for corpus-based studies of Polish and establish standards for the collection, processing and distribution of Polish corpus resources.

¹ <https://clarin-pl.eu>

² <https://typo.uni-konstanz.de/parseme>

³ <http://zil.ipipan.waw.pl/Chronofleks>

⁴ <https://meertensinstituut.github.io/mtas/index.html>

⁵ <http://korpus19.nlp.ipipan.waw.pl>

⁶ <https://korba.edu.pl>

⁷ <http://nkjp.nlp.ipipan.waw.pl>

2 MOTIVATION FOR THE POLISH CORPUS INFRASTRUCTURE

The Polish Corpus Infrastructure is planned as a unique project playing a key role in further progress of research on the Polish language, both in linguistics (or more generally in the humanities) and the technology. The National Corpus of Polish was completed in 2011. It emerged as an effect of collaboration of four teams, which – prior to joining their efforts – had worked on their own corpora in the spirit of competition rather than cooperation. NKJP revealed the potential of synergy. The project which we describe in this article will cover a broader group of undertakings.

Although NKJP was one of the largest reference corpora available when it was compiled, it is a medium-sized corpus by modern standards. Moreover, the corpus is to some extent outdated, at least as a source of lexical data, which limits its applications in lexicography, but also in natural language processing: many proper and common names vital to language processing (i.e. Emmanuel Macron, Donald Trump, Brexit, Instagram, fejk/fake, fanpage, or even selfie) are absent in NKJP or occur only in outdated contexts. Straightforward consequence of such a state of affairs is increasing error of statistic language models — for example speech recognition — only because they are based on outdated linguistic data. Finally, the spoken data do not meet modern requirements, e.g. the quality of recordings is often very low, in many cases it is impossible to consult the voice, not to mention research in phonetics.

Moreover, a number of corpus related initiatives have emerged since 2011: a number of historical corpora have been compiled, covering the 16th century (Institute for Literary Research, Polish Academy of Sciences), 1600–1770 (Institute of Polish Language, Polish Academy of Sciences), and 1820–1918 (University of Warsaw). Prior to NKJP a corpus of Medieval Polish was prepared at the Institute of Polish Language, Polish Academy of Sciences. These corpora cover a large portion of historical Polish, however, there remain some gaps. What is more important, having been compiled by diverse researchers, they are not entirely compatible. Recently, there have been attempts to unify such corpora and establish a common format of metadata and a tagset based on a unified theoretical approach (Król et al. [10]). Morphosyntactic taggers for post-medieval Polish have also been created (e.g. Waszczuk et al. [23]).

As for the development of spoken Polish corpora, the original datasets of spoken-conversational language included in NKJP were expanded, time-aligned with the original recordings and exposed through the Spokes search engine (Pęzik [16]). In 2019, a large corpus documenting the dialect of Spisz, comprising ca. 2 million running words of transcripts of spoken language was launched (Grochola-Szczepanek et al. [4]) and a corpus of Corpus of Polish Teenage Talk was compiled in 2014⁸. Despite these efforts, the level of representation of spoken Polish in the form of multimedia databases leaves much to be desired.

⁸ <http://www.laboratoriumjezykowe.uw.edu.pl>

With regard to parallel corpora, major parallel corpora compiled after 2011 include a Polish-Russian⁹, Polish-German, and Polish-English (Paralela; Pezik [17]). Additionally a Polish component of the International Comparable Corpus was developed (Kirk et al. [9]).

Many of these projects employed a shared set of tools, e.g. a morphological analyzer Morfeusz (Woliński [24]) which is used in virtually all Polish corpora.

Perhaps most importantly, Polish is still one of the few large European languages with an outdated national corpus. There is little doubt that the proposed infrastructure could bridge a number of gaps in the availability and interoperability of corpus resources, thus advancing research on Polish — one of the biggest Slavic languages — in the forthcoming decades of the digital era.

3 THE OUTLINE OF THE INFRASTRUCTURE

The National Corpus of Polish will serve as the core of the proposed Polish Corpus Infrastructure and a representative and up-to-date resource whose main part covers present-day Polish starting from the year 1945. The updated corpus will have a large gender-, channel- and register-balanced component. Additionally, the reference corpus of modern Polish will be federated with various existing corpora of older Polish and Polish dialects, parallel corpora, and model training sub-corpora annotated on different semantic and syntactic levels. Federated collection of corpora will require locating individual parts of the infrastructure at individual branches of the consortium and at affiliated institutions.

One of the main goals of the infrastructure is to provide a proper level of representation of Polish for the purposes of linguistic research and language technologies in the era of big data. Reference corpora fulfilling this criterion require constant updating in order to efficiently contribute to the enhancement of linguistic technologies this facilitating the monitoring of current trends in lexical and syntactic change. For example, every five years the Institute of the Czech National Corpus issues a 100-million-word balanced sample comprising texts published during the most recent five-year period. At the same time, the corpus is supplemented with post-1990 journalistic texts, currently totalling one billion words. Similarly, distributional models of Polish, which are a basic resource in recent approaches to natural language processing can only perform optimally if they are based on a regularly updated reference corpus. A variety of language resources and technologies originally based on the 2011 edition of NKJP may soon become critically obsolete if the national corpus is not regularly updated. Also, the continuous emergence of neologisms, neosemanticisms and other aspects of language change dynamics may soon severely limit potential of NKJP in research in lexicography and linguistics.

⁹ <http://pol-ros.polon.uw.edu.pl>

Apart from simply supplementing NKJP with new data, another important goal of the proposed infrastructure is to insure the proper quality of data collected. Although in the age of the Internet, acquiring large bodies of textual data is becoming increasingly easier, the proper sanitation, balancing and classification of such texts calls for a rigorous method. There are clear benefits of a systematic and controlled approach to continuous development of reference corpus resources, as opposed to their largely uncontrolled and thus biased compilation from ad hoc internet sources.

The planned infrastructure will also ensure interoperability of tools and enable their adjustment to various types of linguistic data.

4 IMPLEMENTATION SCOPE

The PIK infrastructure will deliver a number of technical outcomes in the form of data exchange standards, reference data sets, federated corpus search and monitoring services as well as advanced corpus exploration tools. The four major technical work packages planned in the project are described below.

1. The first major set of technical tasks will focus on implementing multimodal metadata and linguistic data exchange formats for the Polish language. It will involve developing a principled approach to storing texts, including historical and dialectal ones, with potentially rich bibliographic, sociolinguistic, morphosyntactic, syntactic, semantic and discursive annotations, as well as methods for multimedia data representation: aligning texts with their sources (spoken data, video, scans) and methods of searching and managing source corpus files. This step is also necessary for integrating different formats developed within current corpus projects which will feed into the new infrastructure.
2. The second area of technical work will involve extending the balanced segment of the National Corpus of Polish with contemporary texts published after 2011. Mechanisms of continuous acquisition of densely sampled web-based corpus data will be created and deployed to monitor regular fluctuations in lexical frequencies and long-term dynamics of language change. Continuous acquisition of spoken data will have to be addressed as a separate challenge.
3. A separate work package will be aimed at establishing a federation of Polish corpora in order to provide programmatic access to existing third-party contemporary, diachronic, dialectal, spoken and parallel corpora. Mechanisms for simultaneous federated search will be implemented with special consideration of user interface experience and programmatic access, facilitating the use of the corpus infrastructure for both researchers and non-specialist users.

4. The range of dedicated tools for exploring and analyzing the core resources will be considerably larger than the federated search functionalities. The fourth technical work package of the planned infrastructure will provide corpus as well as search and analysis tools for exploring phraseology, differences in stylistic distribution, generating frequency lists, user-defined filtering of concordances, text profile analysis and creation of virtual collections from the index of reference corpus data.

5 AVAILABILITY OF THE PROPOSED INFRASTRUCTURE

In general, the proposed corpus infrastructure will be widely available for scientific research and technological applications. At the same time one should be aware of the copyright limitations on the distributability of the texts included in corpora. A number of infrastructure access models and scenarios are therefore planned to deal with this problem:

- *Access for end-users by Internet tools & services.* End-users will be able to use infrastructure through Web-based applications, i.e. corpus browsing systems and multimedia databases. This model of access will work particularly well for researchers in linguistics and humanities, because their requirements are fairly predictable.
- *Remote access for programmers.* Web applications will be accessible as programming interfaces (API) to facilitate large scale processing of data and development of client applications by the user community.
- *Full access to annotated subcorpora of samples.* Manually annotated subcorpora and training resources/data, indispensable for further progress of language processing, will be accessible in full under open-access licenses. A good example of such a resource is the 1-million-token subcorpus of NKJP which has been widely used in many NLP systems/applications for Polish.
- *Full access to public domain resources.* Whenever possible, resources acquired from public domain repositories, together with manual and automatic annotation will be fully available for offline use.
- *Full access to statistical and distributional models and other derivatives.* Statistical and distributional models, which cannot be used to reconstruct source texts, will be accessible under open licenses. An example of such data are n-gram models of language used in speech recognition systems or vector representations of words (i.e. word embeddings) used currently in many tasks in NLP, text classification, etc.
- *Custom-made models.* Researchers with special requirement will be able to order custom-made statistical models such as distributional language models computed with special tools and parameters. The operators of the infrastructure will make such models available under open licenses to other users.

We believe that the creation of the proposed corpus infrastructure will enable two-way cooperation with users and researchers. On the one hand, it will allow simple and effective inclusion of collections created for special purposes such as corpora of students' and teenagers' language, writers' idiolects, learner language and translation. On the other, it will be used through dedicated tools in research and teaching.

6 CONCLUSIONS

To sum up, it has to be stated clearly that Polish needs a large, balanced, representative national corpus. An up-to-date reference corpus is an indispensable resource for any modern language. The original National Corpus of Polish was innovative and it was even considered as an exemplary reference corpus for other languages in 2011, but for the last five years we have been getting more and more questions from our colleagues abroad about the current state of NKJP. Even a chronological update of the original corpus with samples of registers (dialects, registers, teen talk) and parallel sub-corpora would not be enough for today's challenges. The National Corpus of Polish which is truly a part of the European corpus landscape should be characterized by unrestricted availability for scientific research and innovative technical applications.

Our project addresses these expectations, but it needs funding to be realized. Two institutes of the Polish Academy of Sciences (Institute of Computer Science and Institute of Polish Language), University of Łódź and University of Warsaw have submitted an application to include the Polish Corpora Infrastructure in the The Polish Roadmap for Research Infrastructures developed by the Polish Ministry of Science and Higher Education. The inclusion of PIK in this framework would create an opportunity to obtain permanent funding which is a *sine qua non* of the initiative in the form described in this paper.

Even though it goes without saying that the development of the infrastructure will pose several scientific and technical challenges, a few of them are worth stating. The most important one seems to be related to the federated data model, requiring comparison of data obtained from many dispersed data resources and its proper evaluation, which is a research problem in its own right. Another problem is related to building complex language models adequate for highly inflectional languages. One more challenge of this kind is building a federal access system to the corpus infrastructure that meets the requirements of security and efficiency of data access. Last but not least, we envisage a logistic challenge resulting from the need to obtain consent from the copyright holders of new data, which requires convincing them that the project will bring substantial benefits to their businesses. Moreover, due to the concentration on the media market a lack of a consent of a large market player causes a significant loss of available texts. Acquiring spoken texts, which are very important for linguistic research, is an expensive and logistically complex undertaking.

However significant they may seem, these challenges must be overcome. Corpora have become a basic resource for linguistic research and language technology development. A language without an up-to-date reference corpus has limited perspectives for consideration in international research projects and language technology enterprises. It is high time that we released a new edition of the National Corpus of Polish with its full infrastructure to the public.

References

- [1] Czerepowicka M. (2014). SEJF – Słownik elektroniczny jednostek frazeologicznych. *Język Polski XCIV* (2), pages 116–129.
- [2] Čermák, F. (1997). Czech National Corpus: A case in many contexts. *International Journal of Corpus Linguistics* 2 (2), pages 181–197.
- [3] Derwojedowa M., Kieraś W., Skowrońska D., and Wołosz R. (2014). Korpus polszczyzny XIX wieku — od mikrokorpusu do korpusu średniej wielkości. *Prace Filologiczne LXX*, pages 251–256.
- [4] Grochola-Szczepanek H., Górski R. L., von Waldenfels R., and Woźniak M. (2019). Korpus języka mówionego mieszkańców Spisza. *LingVaria LV* (1), pages 165–180.
- [5] Gruszczyński W., Adamiec D., and Ogrodniczuk M. (2013). Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) *Polonica XXXIII*, pages 311–318.
- [6] Hajnicz E., Patejuk A., Przepiórkowski A., and Woliński M. (2016). Walenty: słownik walencyjny języka polskiego z bogatym komponentem frazeologicznym. In K. Skwarska and E. Kaczmarek (eds.) *Výzkum slovesné valence ve slovanských zemích*, pages 71–102. Prague, Czech Republic, Slovanský ústav AV ČR.
- [7] Janus D., and Przepiórkowski A. (2007). PoliQarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 85–88, Prague, Czech Republic.
- [8] Kieraś W., and Woliński M. (2018). Manually annotated corpus of Polish texts published between 1830 and 1918. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, S. Piperidis, and T. Tokunaga (eds.) *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3854–3859, Paris, France: European Language Resources Association.
- [9] Kirk J., Čermáková A., Ebeling S. O., Ebeling J., Kren M., Aijmer K., Benko V., Garabík R., Górski R. L., Jantunen J., Kupietz M., Simkova M., Schmidt T., and Wicher O. (2018). Introducing the International Comparable Corpus. In S. Granger, M–A. Lefer and L. Aguiar de Souza Penha Marion (eds.) *Book of Abstracts: Using Corpora in Contrastive and Translation Studies Conference (5th edition)*. CECL Papers, Louvain-la-Neuve.
- [10] Król M., Derwojedowa M., Górski R. L., Gruszczyński W., Opaliński K. W., Potoniec P., Woliński M., Kieraś W., and Eder M. (2019). Narodowy Korpus Diachroniczny Polszczyzny. *Projekt. Język Polski XCV* (1), pages 92–101.
- [11] Łaziński M. (2018). Nowe zjawiska w języku młodzieży. *Gramatyka slangu*. In B. Pędzich, M. Wanot-Miśtura, and D. Zdunkiewicz-Jedynek (eds.) *Tyle się we mnie słów zebrało. Szkice o języku i tekstach*, pages 339–356. Warsaw, Poland.
- [12] Mykowiecka A., Marciniak M., and Rychlik P. (2017). Testing word embeddings for Polish. *Cognitive Studies / Études Cognitives* 17, pages 1–19.

- [13] Ogrodniczuk M., Głowińska K., Kopec M., Savary A., and Zawislawska M. (2013). Polish Coreference Corpus. In Z. Vetulani (ed.), *Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 494–498, Poznań, Poland: Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza.
- [14] Ogrodniczuk M., Derwojedowa M., Łaziński M., and Pęzik P. (2017). *Narodowy Korpus Języka Polskiego – co dalej?* *Prace Filologiczne*, LXXI, pages 237–245.
- [15] Pęzik P. (2014). Graph-Based Analysis of Collocational Profiles. In V. Jesenšek and P. Grzybek (eds.) *Phraseologie Im Wörterbuch Und Korpus (Phraseology in Dictionaries and Corpora)*, pages 227–243. ZORA 97. Maribor.
- [16] Pęzik P. (2015). Spokes – a Search and Exploration Service for Conversational Corpus Data. In *Selected Papers from CLARIN 2014*, pages 99–109. Linköping Electronic Conference Proceedings. Linköping University Electronic Press.
- [17] Pęzik P. (2016). Exploring Phraseological Equivalence with Paralela. In *Polish-Language Parallel Corpora*, edited by Ewa Gruszczynska and Agnieszka Leńko-Szymańska, pages 67–81. Warsaw, Instytut Lingwistyki Stosowanej UW.
- [18] Pęzik P. (forthcoming, 2019). Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*.
- [19] Przepiórkowski A., Bańko M., Górski R. L., and Lewandowska-Tomaszczyk B. (eds.) (2012). *Narodowy Korpus Języka Polskiego*. Warsaw, Wydawnictwo Naukowe PWN.
- [20] Riegel M., Wierzbą M., Wypych M., Żurawski Ł., Jednoróg K., Grabowska A., and Marchewka A. (2015). Nencki Affective Word List (NAWL): The Cultural Adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. *Behavior Research Methods* 47(4), pages 1222–1236.
- [21] Twardzik W., and Górski R. L. (2003). Korpus staropolski Instytutu Języka Polskiego PAN w Krakowie. In S. Gajda (ed.) *Językoznawstwo w Polsce. Stan i perspektywy*, pages 155–157.
- [22] Waszczuk J. (2012). Harnessing the CRF complexity with domain-specific constraints: The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of COLING 2012*, pages 2789–2804. Mumbai, India.
- [23] Waszczuk J., Kieraś W., and Woliński M. (2018). Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields. In P. Sojka, A. Horák, I. Kopeček, and K. Pala (eds.) *Proceedings of the 21st Text, Speech, and Dialogue International Conference (TSD 2018)*, Brno, Czech Republic. *Lecture Notes in Artificial Intelligence* 11107, pages 188–196. Springer-Verlag.
- [24] Woliński M. (2014). Morfeusz reloaded. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.) *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1106–1111, Reykjavík, Iceland: European Language Resources Association.
- [25] Wróblewska A. (2012). Polish dependency bank. *Linguistic Issues in Language Technology* 7 (2), pages 1–18.
- [26] Żmigrodzki P., Bańko M., Batko-Tokarz B., Bobrowski J., Czelakowska A., Grochowski M., Przybylska R., Waniakowa J., and Węgrzynek K. (eds.) (2018). *Wielki słownik języka polskiego PAN*. Geniza, koncepcja, zasady opracowania. Kraków, Instytut Języka Polskiego PAN/LIBRON, 264 p.

RELEVANT CRITERIA FOR SELECTION OF SPOKEN DATA: THEORY MEETS PRACTICE

MARIE KOPŘIVOVÁ – ZUZANA KOMRSKOVÁ –
PETRA POUKAROVÁ – DAVID LUKEŠ

Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

KOPŘIVOVÁ, Marie – KOMRSKOVÁ, Zuzana – POUKAROVÁ, Petra – LUKEŠ, David: Relevant criteria for selection of spoken data: theory meets practice. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 324 – 335.

Abstract: The present paper seeks to review relevant criteria used in classifying speech events (SEs) from the perspective of spoken corpus design. The primary goal is to survey the landscape of possible types of spoken language, so as to assess in which directions the coverage of spoken Czech offered by Czech National Corpus corpora can be expanded in the future. We approach the problem from both theoretical and practical points of view, examining what the theoretical literature has to say as well as approaches implemented in practice by existing spoken corpora of various languages. We then synthesize the obtained information into a pragmatically motivated set of SE classification criteria which does not aspire to be universal or definitive but aims to serve as a useful guiding principle and conceptual framework for understanding and promoting SE diversity when collecting spoken data.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 INTRODUCTION

Ever since spoken language corpora started appearing, their authors have been trying to include different types of spoken communication in them [1]. Deciding on the criteria for the composition of a corpus which aims to reflect spoken communication is a crucial part of the entire process, as building these corpora is very time consuming and costly ([2], [3], [4]). In addition, there is no generally accepted classification of spoken language which would be similar to the library classification of disciplines in written texts (Universal Decimal Classification). Spoken communication has a number of aspects which are difficult to project into one classification and even more difficult to implement within a corpus.

It is necessary to reconsider these criteria using literature and taking into account the practical solutions chosen by the authors of previous corpora. These considerations will help us to better target the collection of those types of spoken data which take an especially great deal of effort to collect.

There are many types of communication, some more specialized or less common than others, although in practice, they all lie along a continuum. To make the discussion manageable, we exclude communication with children or in general speakers who are still learning the language, as well as communication with animals or machines, and take into consideration only communication between adult humans proficient in the given language. We also restrict our notion of spoken language to utterances that are mostly formulated on-the-fly, as they are spoken. Therefore, we exclude written-to-be-spoken communication. We use the term speech event (SE) for a stretch of speech that takes place in a particular situation and under certain conditions (e.g. lecture – formal settings, prepared etc., conversation at dinner with friends – informal, spontaneous etc. [5]).

The article is structured as follows: the first chapter surveys theoretical approaches based on selected literature. The first part of the following section gives an overview of practical, actually implemented solutions on the example of selected corpora of spoken language, the second part contains a brief summary of currently available Czech spoken data. In the third chapter, we build on the ideas presented in the previous sections to present the factors that we believe are important for the collection of spoken data in Czech. In conclusion, we present and justify the current focus of spoken data collection at the Institute of the Czech National Corpus (ICNC).

2 OVERVIEW OF LITERATURE

The basic dichotomy in language, which is more or less apparent from the lowest (morphemic) to the highest (textual) level, distinguishes between spoken and written language. These constitute the two extremes of a scale (although this scale is of course a continuum), stereotypical/prototypical representatives with mutually exclusive features. Generally, this description appears in grammars where the terms “written” and “spoken” are implicitly connected with a use of language in particular situations which require fulfilling particular “norms”. Accounts of written language usually concern the standard language (in the sense of prescriptive rules enforced in language). As for spoken language, it usually refers to spontaneous language used in informal settings among friends or family members etc. The term for this type of spoken language varies, e.g. common spoken language (e.g. [6], [7]), the language of everyday spoken dialogues [8], vernacular speech [9], [10, p. 233].¹

The vagueness and inconsistency in terminology is also reflected in the inconsistency of the notion of the term “spoken Czech” itself. According to [11, p. 46], spoken Czech is understood in various ways in linguistic papers: as a “communica-

¹ [10] mentions the form of tales and dialect texts. For other possible terms used in the Czech context, see [11, p. 46]. They are all very descriptive, capturing the external conditions of the SE or its emotional setting (emotionality, expressive speech, intimate tone etc.). Another term from English-speaking studies is intimate discourse [12].

tion form / mode of being of language” where all spoken varieties are assigned, as a synonym for “SEs in standard Czech”, typical for formal settings, or its meaning is reduced to dialects or Common Czech.

Even so, there are some conditions that are considered regarding the classification of SEs which we can generalize from (for more details, see [13]) and which could be helpful for describing the continuum between the two extreme, prototypical language forms: relationship between speaker and addressee, topic, shared context (not only regarding knowledge, but shared experience in general)², place and time³ of SE, setting (official etc.), social status of speaker and addressee. These criteria are not completely orthogonal, i.e. specifying some of them might implicitly narrow down the possible values for others – e.g. given a specific topic and an official setting for an SE, we can reasonably expect preparedness and so on.

Not all criteria have to be taken into consideration, only some of them can be chosen for the classification of SEs, depending on the research topic – specifically, only criteria which correspond to the researcher’s intention and aim can be taken into consideration, ultimately yielding not one universal classification, but various special purpose ones. As a consequence, the spectrum or continuum of SEs does not have to be defined exhaustively, but only selectively.

3 OVERVIEW OF SELECTED SPOKEN CORPORA

The *creation* of a spoken corpus is a challenge involving a number of smaller decisions on several levels. The design of the corpus should take into account the various dimensions underlying the variation that can be observed in language use. This chapter briefly summarizes the basic information about publicly available spoken corpora in six languages other than Czech (3.1) and Czech (3.2); the attention is paid to the types of SE gathered within the corpora.

3.1 Overview of selected non-Czech spoken corpora

3.1.1 Lancaster/IBM Spoken Corpus

This spoken corpus is the smallest and oldest in this overview. It contains 52,637 words of spoken British English and was released in 1987. The aim of the corpus was to collect a sizeable sample of that type of spoken English which is “suitable as a model for speech synthesis. This explains the relatively high proportion of prepared or

² [14] accentuates the relationship between the participants of an SE and the amount of shared context (lack of shared context requires adding “background information”; p. 40). In his book, attention is paid to five styles of English usage (intimate, casual, consultative, formal, frozen), which are situated on the scale of familiarity – formality.

³ For example, [15, pp. 34–35] states four functions of spoken language. The most important difference is distinguishing between situational (commonly spoken) and non-situational SEs (marked as “secondary spoken”; lectures, expert training etc.; cf. [8, p. 189]).

semi-prepared speech produced by trained broadcasters” [16, p. 6]. The length of recordings was not limited. The corpus contains the following categories of SEs: commentaries, news broadcasts, religious broadcasts (= daily services), radio discussions, propaganda, university lectures, public lectures, magazine-style reporting, fiction and poetry readings, informal dialogues among friends. All categories except the last one were produced in a public, rather formal setting and for a public audience.

3.1.2 *British National Corpus (BNC)*

The design of the BNC has been perhaps the most impactful in terms of influencing subsequent spoken corpora. The original BNC, released in 1994, consists of two components, a 10-million-word spoken one and a 90-million-word written one. The newest version called BNC2014 follows the previous design while focusing on newer data, and as of March 2019, only the spoken part has been completed [17]. The spoken BNC2014 has 11.5 million words gathered from across the United Kingdom. In contrast to the spoken BNC1994, it contains only spoken interactions in informal settings, especially at home, which took place among friends and family members.

The spoken BNC1994 includes both spontaneous, informal interactions (in the so-called demographic part) and formal context-governed encounters in four broad categories of social context: education/providing information, business, institutional/public communication, and leisure. Each category within the context-governed part of the BNC1994 was limited in size (max. 200,000–300,000 words); the range of SEs was defined, but not fixed. “The overall aim was to achieve a balanced selection within each, taking into account such features as region, level, gender of speakers, and topic. Other features, such as purpose, were applied on the basis of post hoc judgements” [18]. Attention was also paid to the dichotomy of monologue (40% of each social-context group) and dialogue (60% of each social-context group).

3.1.3 *Corpus Gesproken Nederlands (CGN)*

The CGN initially had a carefully-structured design which however had to be revised for pragmatic reasons⁴ [19]. In the overall design, the principal criterion was taken to be the socio-situational setting in which language is used, whereas communicative goal, medium, number of speakers participating, the relationship between speaker(s) and hearer(s), and the sociolinguistic characteristics of speakers (i.e. age, gender, region, socio-economic class) were seen as supplementary criteria.

The released version⁵ has 8.916m words and more than half of this material (4.7m) comes from informal, spontaneous, face-to-face and telephone dialogues.

⁴ These reasons are described in [20, p. 341] as follows: “... because of the time, financial, and legal constraints under which the project must operate, but also for practical reasons, it is impossible to include all possible types of speech and compromises are inevitable.”

⁵ For more details, see http://lands.let.ru.nl/cgn/doc_English/topics/version_1.0/overview.htm

The remaining part consists of the following SE categories: interviews with teachers, simulated business negotiations, broadcast interviews/discussions/debates, non-broadcast political discussions/debates/meetings, classroom lessons, lectures/seminars, broadcast live commentaries (e.g. in sports), broadcast news reports/reportage, broadcast news, broadcast commentaries/columns/reviews, ceremonious speeches/sermons, and read speech. The list shows that attention was paid to the distinction between monologues vs. dialogues, and broadcast vs. non-broadcast.

3.1.4 *Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK)*

The FOLK corpus was designed according to several principles inspired by previous spoken corpora, in particular the BNC1994. The aim was to gather “a maximally diverse range of verbal communication in private, institutional, and public settings” [21, p. 383]. The released version has 1.95m tokens and contains recordings collected within the FOLK project, as well as within other projects of the same institute, i.e. map tasks, biographical interviews. Data collection was relatively free as far as more detailed distinctions within the social-setting categories are concerned, although there was an effort to apply speaker-related sociolinguistic criteria (i.e. age, gender, region). For the complete list of SE categories, see [21].

3.1.5 *Göteborg Spoken Language Corpus (GSLC)*

The GSLC was created more opportunistically than the other corpora described here, which means without any prior corpus design. The main goal was to ensure the broadest possible range of different SEs [22]. The corpus consists of 1.42m words collected within 27 SE categories. Most of them could be found in any other spoken corpus (and cover the range of four social-setting categories in the BNC1994), but some are fairly unique (for example bus driver/passenger conversation, physical therapy, or quarrel). These latter were either produced in the workplace, e.g. in a factory, travel agency, hotel, shop, at the doctor’s, or in a task-oriented experimental setting (the complete list is available in [22]). The categorization in the GSLC is mostly ad hoc and low-level, no effort was made to establish any higher-level categories according to e.g. number of speakers or situational context.

3.1.6 *Slovenský hovorený korpus (SHK)*

The SHK is a long-running project which regularly releases new and improved versions of spoken corpora of Slovak.⁶ The newest version – s-hovor-6.0 – has 6.593m words. The SHK is a collection of recordings from various SEs within all sorts of social settings [23]. There are both dialogues and monologues with varying degrees of spontaneity and formality, e.g. spontaneous dialogues, lectures, sermons, broadcast discussions, oral-history narratives.

⁶ For more details, see <https://korpus.sk/shk.html>

3.2 Publicly available spoken data of Czech

A disclaimer is in order first: the following discussion applies only to publicly available corpora mainly used for linguistic research, though we are well aware that there are other corpora (or more broadly, data sets) which may not be generally accessible and/or which serve other than linguistic purposes (e.g. speech recognition). With that in mind, most spoken Czech corpora to date have focused primarily on prototypical spoken language, which is defined as dialogic SEs within an informal private setting, among family members and friends [13]. Over the years, this niche has spawned parts of the PMK [24] and BMK corpora [25] the ORAL series corpora [26], and the ORTOFON corpus [27]. Unlike many other types of SEs, where pre-existing recordings can be harvested, this type of SE generally requires collecting data from scratch, i.e. fieldwork. The earliest available recordings of this sort date back to 1988; today, the ICNC continues collecting data in this tradition and aims to keep doing so for the foreseeable future.

Other types of SEs represented within public spoken Czech corpora include:

- the controlled interview, which is mainly used in dialect-oriented research and which usually takes place in a somewhat more formal setting; parts of the PMK and BMK corpora would fall into this category, as well as the DIALEKT corpus [28], [27]
- classroom interactions, in which dialogue is also prevalent, as gathered within the SCHOLA corpus of school communication, capturing entire lessons [29]
- broadcast TV programs, mainly debates and talk shows, collected in the DIALOG corpus [30].

This short overview suggests some areas that are not covered by the currently available roster of corpora, for instance monologues and non-broadcast SEs in a public setting.

3.3 Summary

This overview has shown different approaches to creating spoken corpora, from carefully planned to completely unplanned data collection. We emphasized those corpora that include the most diverse spoken data in terms of various characteristics of SEs. The overview of SEs included within these corpora serves as a source of inspiration in terms of the possible directions of expansion of spoken data collection in Czech.

4 CLASSIFICATION CRITERIA RELEVANT FOR SPOKEN CORPUS DESIGN

In order to inform current and future spoken data collection, we have chosen some aspects from the available theoretical descriptions and practical implementations of spoken data classification. In what follows, we foreground criteria which

can be derived directly from the situation in which the SE occurred, without having to ask the participants. Accordingly, some of the aspects mentioned above will tend to be downplayed, others may be amalgamated into a single criterion. Not all criteria apply to all SEs. One clear exception to the focus on situation-derivable criteria are socio-demographic characteristics, which participants typically need to state explicitly. However, their relevance to spoken corpus design is clear, so we mention them in a separate section for completeness' sake.

The criteria are divided into two broad categories: setting-related and participant-related. Each category consists of multiple subcategories. In order to describe them, we offer a simplified, often dichotomous account (cf. also [6]), but it goes without saying that the characteristics of real-life SEs are often far from black and white. When selecting and organizing the criteria, an effort was made to map them as closely as possible to the existing categories mentioned above, and to avoid ambiguous terms (e.g. 'spontaneous', which can mean either 'informal' or 'unprepared').

4.1 Setting-related criteria

4.1.1 Degree of officiality

This aspect takes into account the social role of the speaker. The term social role refers to the set of behaviors, rights, obligations, beliefs, and norms as conceptualized by people in a social situation [31]. This category distinguishes whether the speaker represents an institution⁷ (e.g. the headmaster's opening speech at the beginning of the school year) or is entrusted with a ceremonial task to perform (e.g. a birthday toast). On the other side of the scale are situations where everybody can join the conversation at their own discretion.

4.1.2 Degree of publicness

This aspect is closely connected to the size of the audience and the relationship between them. It indicates whether the SE is public, accessible to everyone (e.g. a speech on a public square, a political debate on TV), or restricted to the members of the community within which the SE takes place (e.g. preaching, work training), or whether the SE only has one addressee and thus is private (e.g. conversation with a doctor, lawyer, friend).

4.1.3 Mediation of communication

There are situations where both participants are physically present in the same place (face-to-face), and situations where they are not and their communication is transmitted via a mediated channel (e.g. phone, Skype, live TV debate).

⁷ The term *institution* is broadly understood as any generally practised pattern of behavior, regulated by a given culture, often associated with specific SEs [32].

4.1.4 Synchronicity

This criterion focuses on whether the communication takes place at the same time for both participants (e.g. face-to-face, via telephone) or if the time of speech production is distinct from the time of speech perception (e.g. pre-recorded material on TV or on the web).

4.2 Participant-related criteria

4.2.1 Number of (active) speakers

A prototypical monologue is a speech by one speaker who is informed in advance that the time for his/her speech will be reserved. Of course, in practice, verbal interaction with (one of) the addressee(s) can also be initiated, and monologues can also arise spontaneously from the situation, without being explicitly sanctioned. But the basic condition is that one speaker speaks and does not expect to be interrupted until s/he yields, whereas in a dialogue, the speaking role is shared by two or more participants and turn-taking is managed dynamically.

4.2.2 Degree of preparedness

Either the speaker knows about the purpose and topic of the SE and can therefore, at his/her discretion, make preparations (e.g. with prior research, presentation slides, written notes), or s/he does not know in advance that s/he will be speaking at a given moment and thus has to respond on-the-fly⁸ (e.g. an opinion poll on the street, a private chat with friends).

4.2.3 Number of addressees

Instead of capturing this as a continuous numerical variable, it makes sense to partition the continuum into broad classes which capture some qualitative shifts. Sociologists distinguish the following levels: speech directed towards one person, a small group of up to 19 people, a large group up to 39 people or a large group of addressees (the public) [33, p. 68].

4.2.4 Degree of addressee activeness

We distinguish whether the addressee can claim speaking initiative and thus become an active speaker in his/her own right (as is typically the case in a dialogue), whether s/he is allowed to ask questions, or if s/he does not even speak or influence the other speaker at all (e.g. broadcast SEs).

4.2.5 Relationship between participants

4.2.5.1 Amount of shared background

This aspect captures the amount of common understanding of the wider context of the SE, which may be high e.g. for family members, long-time friends, but also

⁸ This type of production is sometimes referred to as spontaneous, but as mentioned above, we refrain from using this term in this paper.

for professionals from the same field who are working on the same task. On the other side of the scale, there are SEs where participants share little background (private, professional or other) with respect to what is being spoken about.

4.2.5.2 Degree of familiarity

In this aspect, the closeness of participants is taken into account, i.e. how intimately acquainted they are with each other. To suggest the range possible here, consider e.g. the dynamic between family members vs. between applicant and recruiter in a job interview. Both types of situations can yield SEs which are private in the sense of 4.1.2 above, but they differ vastly in terms of their degree of familiarity.

4.2.5.3 Symmetry of social roles

Each SE is also influenced by the mutual social status of the participants in communication. This could be related to various factors, e.g. age or profession. The relationship between the social roles of both participants could be symmetric (e.g. conversation among friends of the same age), or asymmetric (e.g. conversation between boss and subordinate). When making social role symmetry judgments, it is important to realize that each participant plays many different social roles and to focus on the role(s) which is/are most saliently activated in the context of the given SE.

4.2.6 Socio-demographic characteristics

This category comprises the following kinds of items: gender, age, highest achieved level of education, region, profession, place of residence, size of settlement etc. It is a good idea to collect socio-demographic data which is as detailed as possible, and only later possibly bin the values into larger groups (e.g. age groups instead of looking at exact age). Without such binning, balancing or any kind of demographic representativeness (possibly even reflecting the demographic distribution in the population) may be an impossible goal to achieve.

5 CONCLUSION

We addressed the topic of classification of SEs, trying to summarize theoretical approaches and confront them with practical implementations in existing spoken corpora. After reviewing the theoretical literature and the composition of several spoken corpora, both Czech and non-Czech, we proceeded to sketch our own categorization, inspired by these sources. To make the criteria more specific, we attempted to exemplify them by suggesting extremes (for dichotomous categories) or points along the continuum. The resulting classification system can serve as an aid in identifying the types of SEs that are still missing from corpora of spoken Czech.

To wit, there is so far no Czech corpus that contains monological SEs from official settings like most of the spoken corpora described under 3.1 do. This is one of

the shortcomings that we aim to remedy in further data collection, focusing on this type of SE in various contexts, especially public, including professional lectures. Since for the time being, we would like to avoid material that has been extensively scripted and/or edited prior to broadcasting or publishing, we are not currently planning to collect monological SEs from mass media such as radio, television and web shows.

So far, spoken Czech corpora at the ICNC have been mainly focusing on only one type of SE, prototypical spoken language. While we would like to broaden our scope by including new SE types, many of the lessons learned in the past directly carry over: we can take advantage of previous experience and existing infrastructure. For instance, even with the new SE types, we can adhere to the same transcription process, using the same software and battle-tested transcription guidelines (possibly with minor adjustments where necessary or practical). This will streamline the entire process, as well as hopefully make it easier to search across different spoken corpora, and possibly even allow us to combine them into one super-representative corpus in the future.

ACKNOWLEDGMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

- [1] Svartvik, J. (ed.) (1990). *The London-Lund Corpus of Spoken English: Description and Research*. Lund Studies in English 82.
- [2] Deppermann, A., and Hartung, M. (2012). Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des “Forschungs- und Lehrkorpus Gesprochenes Deutsch” (FOLK) am Institut für Deutsche Sprache (Mannheim). In Felder, E., Müller, M., and Vogel, F. (eds). *Korpuspragmatik*, pages 414–450, Berlin, de Gruyter.
- [3] Kopřivová, M. (2017). Mluvený korpus. In P. Karlík, M. Nekula, and J. Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*.
- [4] Gajdošová, K., and Šimková, M. (2018). *Frekvenčný slovník hovorenej slovenčiny na báze Slovenského hovoreného korpusu*. Bratislava, VEDA.
- [5] Hirschová, M. (2017). Komunikační situace. In Karlík, P., Nekula, M., and Pleskalová, J. (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. Accessible at: https://www.czechency.org/slovník/KOMUNIKAČNÍ_SITUACE.
- [6] Chloupek, J. (1986). *Dichotomie spisovnosti a nespisovnosti*. Brno, Filozofická fakulta. Spisy univerzity J. E. Purkyně v Brně.
- [7] Daneš, F. et al. (1997). *Český jazyk na přelomu tisíciletí*. Praha, Academia.

- [8] Hoffmannová, J. et al. (2016). *Stylistika mluvené a psané češtiny*. Praha, Academia.
- [9] Ervin-Tripp, S. M. (1964). An Analysis of the Interaction of Language, Topic and Listener. *American Anthropologist* 66, pages 86–102.
- [10] Vachek, J. (1942). *Psaný jazyk a pravopis*. In *Čtení o jazyce a poesii*, pages 231–306.
- [11] Hoffmannová, J. and Zeman, J. (2017). Výzkum syntaxe mluvené češtiny: inventarizace problémů, *Slovo a slovesnost* 78(1), pages 45–66.
- [12] Clancy, B. (2015). *Investigating Intimate Discourse: Exploring the spoken interaction of families, couples and friends*. Routledge.
- [13] Čermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics* 14(1), pages 113–123.
- [14] Joos, M. (1967). *The five clocks*. New York, Harcourt Brace & World.
- [15] Chloupek, J. (1995). Sjednocující a rozrůžňující faktory v mluvené komunikaci. In *K diferenciaci současného mluveného jazyka*, pages 33–39, Ostrava, Repronis.
- [16] Knowles, G., Taylor, L., and Williams, B. (1996). *A Corpus of Formal British English Speech: The Lancaster/IBM Spoken English*. Routledge, London & NY.
- [17] Love, R., Demby, C., Hardie A., Brezina, V., and McEnery, T. (2017). The Spoken BNC2014. Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, pages 319–344.
- [18] Burnard, L. (ed.) (2000). *The British National Corpus Users Reference Guide*. Accessible at: <http://www.natcorp.ox.ac.uk/docs/userManual/>
- [19] Oostdijk, N. (2002). The Design of the Spoken Dutch Corpus. In Peters, P., Collins, P., and Smith, A. (eds.), *New Frontiers of Corpus Research*. Amsterdam, pages 105–112.
- [20] Oostdijk, N. et al. (2002). Experiences from the Spoken Dutch Corpus Project. *Proceedings of the LREC 2002*, pages 340–347.
- [21] Schmidt, T. (2014). The Research and Teaching Corpus of Spoken German – FOLK. In *Proceedings of the Ninth International conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA).
- [22] Allwood, J. et al. (2003). Annotations and Tools for an Activity Based Spoken Language Corpus. In van Kuppevelt, Jan C.J., and Smith, R.W. (eds.), *Current and New Directions in Discourse and Dialogue*, pages 1–18, Springer.
- [23] Šimková, M., Garabík, R., Karčová, A., and Gajdošová, K. (2008). Hovorený korpus slovenčiny. In M. Kopřivová, and M. Waclawičová: *Čeština v mluveném korpusu*, pages 227–233, Praha, NLN – ÚČNK.
- [24] Čermák, F. et al. (2007). *Frekvenční slovník mluvené češtiny*. Praha, Karolinum.
- [25] Hladká, Z. (2005). Zkušenosti s tvorbou korpusů češtiny v ÚČJ FF MU v Brně. In *SPFFBU A 53*, pages 115–124. Brno, Masarykova univerzita. Accessible at: <http://hdl.handle.net/11222.digilib/101736>
- [26] Kopřivová, M., Lukeš, D., Komrsková, Z., and Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. In *Korpus – Gramatika – Axiologie* 15, pages 47–67.
- [27] Komrsková, Z., Kopřivová, M., Lukeš, D., Poukarová, P., and Goláňová, H. (2017). New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Jazykovedný časopis*, 68(2), pages 219–228.
- [28] Goláňová, H. (2015): A new dialect corpus: DIALEKT. In Gajdošová, K., and Žáková, A. (eds.): *Proceedings of the Eight International Conference Slovko 2015 (Natural Language Processing, Corpus Linguistics, Lexicography)*, pages 36–44. Lüdenscheid, RAM-Verlag.

- [29] Šebesta, K. (2010): Korpusy češtiny a osvojování jazyka. *Studie z aplikované lingvistiky*, 2, pages 11–33. Accessible at: https://studiezaplikovanelingvistiky.ff.cuni.cz/wp-content/uploads/sites/19/2016/03/karel_sebesta_11-33.pdf
- [30] Čmejrková, S., Jílková, L., and Kaderka, P. (2004). Mluvená čeština v televizních debatách: korpus DIALOG. *Slovo a slovesnost*, 65, pages 243–269.
- [31] Vláčil, J. (2017). Role. In Z. R. Nešpor, editor, *Sociologická encyklopedie*. Praha, Sociologický ústav AV ČR, v.v.i. Accessible at: <https://encyklopedie.soc.cas.cz/w/Role>
- [32] Keller, J. – Vláčil, J. (2017). Instituce. In Z. R. Nešpor (ed.), *Sociologická encyklopedie*. Praha, Sociologický ústav AV ČR, v.v.i. Accessible at: <https://encyklopedie.soc.cas.cz/w/Instituce>
- [33] Novotná, E. (2010). *Sociologie sociálních skupin*. Praha, Grada.

THE DIALEKT CORPUS AND ITS POSSIBILITIES

HANA GOLÁŇNOVÁ¹ – MARTINA WACLAWIČOVÁ²

^{1,2} Institute of the Czech National Corpus, Faculty of Arts, Charles University
in Prague, Czech Republic

GOLÁŇNOVÁ, Hana – WACLAWIČOVÁ, Martina: The DIALEKT corpus and its possibilities. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 336 – 344.

Abstract: DIALEKT, a corpus of Czech dialects, has been continuously curated and expanded by the Spoken Corpora section of the Institute of the Czech National Corpus. The following paper aims first to give a concise characteristic of the corpus, addressing its sociolinguistic parameters and possible subcorpora derivable thereof, its two-layer approach to the transcription of dialect recordings, and lemmatization & morphological tagging of the corpus. Subsequently, we move on to examples of how linguists can use the corpus and discuss two related projects which expand upon currently available possibilities: an archive of dialect-specific differential phones of the Czech language (completed) and an interactive web environment for spatial map-based visualization of data from all kinds of spoken corpora (in preparation). Thanks in part also to these additional tools, the DIALEKT corpus should serve both experts in the field as well as the general public.

Keywords: spoken corpus, dialect corpus, dialectology, corpus design, transcription

1 INTRODUCTION

In 2017, the Institute of the Czech National Corpus published a new specialized corpus of spoken Czech, prepared by its Spoken Corpora section: DIALEKT, a corpus of Czech dialects [1] [2]. The corpus captures and presents the traditional regional dialects present on the territory of the Czech Republic. In its first public version, the size of the corpus is approx. 100,000 words, but more data has been collected continuously and expanded versions will be published in the future. However, previous versions will also remain available in their original form so as to enable reproducible research. The corpus is accessible via the KonText web interface developed by the ICNC.

In terms of audience, the DIALEKT corpus aims to reach both language experts (dialectologists, other linguists, and researchers from related fields) as well as amateurs from the general public. It is also expected to serve as a teaching resource at all education levels [3].

The goal of this paper is to introduce the first public version of the DIALEKT corpus and to showcase its possible uses, including resources and tools derived from the corpus which either already exist or are being currently worked on.

2 COMPOSITION AND AIMS OF THE DIALEKT CORPUS

2.1 Composititon of the DIALEKT corpus

The DIALEKT corpus consists of speech recordings made in all traditional dialect regions of the Czech Republic and their transcripts. The current version contains 324 recordings of a total length of 13 hours, comprising 178 unique speakers. The number of speakers is influenced by the fact that dialectological recordings are relatively short (approx. 1–6 minutes, typically 2 minutes), because they mostly focus on one person's account of a single topic. In terms of territorial coverage, the corpus features recordings from all dialect regions of the Czech Republic, including Czech language islands in Poland. For the time being, it does not include recordings from the Bohemian, Moravian and Silesian borderlands, which, however, do not belong among traditional dialect regions. Because of massive population relocation after World War II, these formerly mostly German-speaking areas lack a traditional dialect substrate. As far as word counts are concerned, the individual regional categories are currently not represented evenly. Plans for the immediate future do not include attempts at balancing this out, our main goal is to collect and publish as much data as possible.

2.2 Data collection

The recordings collected in the corpus come from a variety of sources, thanks to which they cover a fairly long time span. They are divided into two time strata, older and newer, and this information can be used to constrain searches or other operations involving the corpus. The older stratum consists of recordings made from the late 1950s up to the 1980s. Part of this material was collected by the Department of Dialectology of the Czech Language Institute of the Czech Academy of Sciences and published in the Addenda to the Czech Linguistic Atlas [4]; the rest comes from private collections, which for the most part have also been previously published. The new stratum then spans recordings from the 1990s up to the present day. This new stratum encompasses recordings made in the course of research activities at various universities, by private individuals, and last but not least, also by the Institute of the Czech National Corpus itself.

The data collection methodology follows the principles usually applied in the field of Czech dialectology. The recordings are mostly informal in nature, even though many of them were obtained within the structured interview research paradigm, i.e. with researchers interviewing subjects. The result is unprepared discourse, predominantly monologues, captured in the private setting of the subject's home. Topics focus on the traditional rural way of life, covering agriculture, arts & crafts, local customs and traditions, contemporary events etc. More specifically, we encounter e.g. recipes for regional dishes, descriptions of crafts like weaving, accounts of Christmas and Easter, memories of the beginning of World War II, or local legends.

For the moment being, our goal is to capture the most archaic state of the traditional regional dialects and we are not concerned with generational differences. Correspondingly, the recordings only feature members of the oldest generation still alive, so as to capture as many of the original dialect features as possible. Informants were selected from among local natives in rural areas who belonged to the settled stratum of the population, mostly spent their entire lives in that one place, and were tied to an agricultural way of life or a particular craft. They all fall into the 60+ age group, in other words, they were born between the late 19th century and the middle of the 20th century.

2.3 Sociolinguistic characteristics and creation of subcorpora

The DIALEKT corpus contains detailed sociolinguistic characteristics of the speakers and the communication situation captured in the recording, thanks to which the dialectological material can be sorted into various groups. One of the user-friendly ways of achieving this is by defining subcorpora in KonText. Recording-related metadata comprise detailed information about the place of recording, e.g. type of locale (urban, rural) and its size, geographical localization – country, region (Bohemia, Moravia, Silesia), dialect group (*skupina*), subgroup (*podskupina*), division (*úsek*) and type (*typ*). Additional characteristics of the communication situation include source of the recording (e.g. an institution), recording date, time stratum membership (older vs. newer), main topic, type of discourse (monologue, dialogue, and combinations thereof), number of speakers, and presence of a researcher. Speaker-wise, we also track a range of metadata, from gender, age, education, place of residence in childhood and longest place of residence (including detailed dialectological classification), to the speaker's longest professional occupation. All of this information can be displayed in the KonText corpus interface and used to retrieve frequency statistics. Most of these sociolinguistic characteristics can also be used to define subcorpora. For instance, if we are interested in phone-level phenomena in north-east Bohemia, we can either restrict our search on-the-fly using KonText's *Restrict search* functionality, or we can create our own permanent subcorpus based on this dialect region, which will always be available after logging in and can be used for repeated searches.

3 PROCESSING DATA

3.1 DIALEKT's two-layer approach to the transcription of dialect recordings

Dialect recordings intended for the DIALEKT corpus are transcribed using the ELAN annotation software (much like those for ORTOFON [5], a corpus of informal spoken Czech conversations). There are two parallel transcription layers, a dialectological one and an orthographic one (i.e. a base transcript). These are time-aligned to the sound recording, along with auxiliary layers capturing additional

paralinguistic information concerning the individual speakers' utterances or the entire communication situation. In the KonText corpus interface, it is possible to search either the dialectological or the orthographic layer separately, or both at the same time, similarly to a parallel corpus with the same documents in multiple languages. Results can also be displayed in this parallel corpus mode, with both layers standing side by side.

The transcription on the dialectological layer follows the approach traditionally used in Czech dialectology, as outlined in the Rules for the Scientific Transcription of Dialectological Records of Czech and Slovak [6] and applied e.g. in the compendium of Czech Dialect Texts [7] or in the Addenda to the Czech Linguistic Atlas [4]. The goal of the transcription is to faithfully capture what the speaker said, in the context of the framework of systematic description of Czech dialects. Differential dialect phones are encoded with special purpose symbols, as traditionally used in dialectological transcripts (e.g. *ǎ* stands for a fronted *a*, *e̞* for an open *e*, *ɫ* for a dark *l*, *w* for a bilabial *v* etc.). Sentence punctuation follows the standard rules of written Czech, but sentences do not start with a capital letter.

jennou vo Vánocich tatínek pouďal, že... mňel něaki známi v Helkovicich, tag že se tam pújdem pod'wat', mňe ůzal taki na lížích.

'once at Christmas Daddy said, that... he had some friends in Helkovice, so we would go to see them, he took me on skis too'

The transcription on the orthographic layer is very similar to the general rules currently applied to regular spoken corpora of the Czech National Corpus [5], [8]. It is fairly close to usual orthography, but diverges from it in trying to capture some characteristic features of spoken language and some regional phenomena. In order to make lemmatization and morphological tagging possible, phone-level differences in word roots are disregarded and overridden by the standard form. Vowel length is also standardized (even in Silesian varieties with systematic shortening). However, morphological variation is kept as is, e.g. endings of all types of declension (*synoj* vs. standard *synovi* 'son' (dative)) and conjugation (*nosijó* vs. standard *nosí* (pl.) 'they wear'). Dialectal or regional lexis (*calta* 'Christmas cake', *zemák* 'potato', *ostat* 'to stay') is preserved, and if no parallel exists in the standard language, no artificial standardization of the root is attempted (*kútky* 'open fireplace'). As for punctuation, unlike the dialectological layer, the orthographic layer has pausal punctuation.

jednou vo Vánocích tatínek povídal že .. měl .. nějaký známý v Helkovicích .. tak že se tam pújdem podívat .. mě vzal taky na lyžích

'once at Christmas Daddy said, that... he had some friends in Helkovice, so we would go to see them, he took me on skis too'

Having two different transcript layers and being able to search them both – even at the same time, as parallel corpora – has many advantages. The dialectological layer allows for highly specific searches targeting various possible phone-level realizations of words (e.g. *chcel*, *cht'il*, *cht'el* ‘he wanted’). On the other hand, if we start from the orthographic layer, we can search for a standardized word form (e.g. *byli* ‘they were’) and retrieve all of its pronunciation variants recorded on the dialectological layer (*bili*, *buli*, *boli*, *beli*, *byli* ‘they were’).

The pausal punctuation of the orthographic layer can also be used to elucidate some phenomena present on the dialectological layer. For instance, the occurrence of a pause can explain why an expected voicing assimilation across word boundaries failed to materialize. The usual case, where voicing assimilation across word boundaries does happen, is illustrated by the following example:

dialectological layer: *tam šlo přez brambori takovi dlouhi strašidlo velkí*
‘such a long, big specter went across the potato field’
orthographic layer: *tam šlo přes brambory takový dlouhý strašidlo velký*
(the absence of punctuation indicates that this stretch of speech was uttered without pauses)

In contrast to this, the missing voicing assimilation across word boundaries on the dialectological layer in the next example can be explained by the presence of a pause, which is recorded on the orthographic layer:

dialectological layer: *voňi zas bili rádi, že se s nima bawíme*
‘they were glad again we were talking to them’
(contrast with expected realization *zas bili* ‘again were’)
ortografická rovina: *voni zas . byli rádi . že se s nima bavíme*
(the period indicates a very short pause)

3.2 Lemmatization and morphological tagging of the corpus

The DIALEKT corpus offers fairly rich linguistic annotation of the transcripts – it is lemmatized and morphologically tagged. Lemmas are always taken from standard language and subsume all non-standard forms of a word. In other words, they encompass not only all inflected forms of a word, but also all regional pronunciations of those inflected forms (e.g. even *chcu* ‘I want’, *cht'el* ‘he wanted’, *sceli* ‘they wanted’, etc. will all be lemmatized as *chtít* ‘to want’). A lemma-based search therefore makes it possible to recover all inflected forms of a given word in all their regional variants. The lemmatization is the same for both transcription layers and yields corresponding results when searching on either one.

Each token is also assigned a morphological tag consisting of 16 characters (mostly letters) encoding various morphological categories (e.g. *aňi ti slova* NNNP4-

----A---- *neumím používat* ‘I can’t even use those words’). When searching by morphological tags, the result set consists of all matching word forms in all their recorded pronunciation variants.

The process of lemmatization and tagging was made difficult by the high variability of the dialect material and a lack of specialized training data. In spite of this, the resulting annotation is relatively accurate and yields fairly reliable search results. Users of the corpus can thus rely on lemma and tag searches to their advantage, making it easier to explore the Czech language system in all its regional diversity.

4 POSSIBLE USES, RESOURCES AND TOOLS

4.1 Example use cases of the corpus

Dialect variants from all levels of linguistic description (phonetics & phonology, morphology, syntax and lexis) can be found in the DIALEKT corpus. Even though it ranks among smaller corpora in terms of size, it faithfully captures a range of phone-level dialect specificities. The corpus allows us to track their conditions of use – not only with respect to their territorial spread, but also in relation to speaker or context characteristics. This makes it possible to track the lexicalization of phone-level phenomena. It turns out this type of lexicalization does not only affect frequent words, but also specific groups of words. It mostly occurs with words related to a given region, its climate, typical way of life and arts & crafts; also affected are proper nouns, in particular toponyms and anthroponyms, and occasionally also expressive words (the recordings offer a somewhat limited range of expressivity, being monological in nature). Considering for instance the rounded pronunciation *w*, a relict of the original bilabial pronunciation, we can confirm it occurs in the north-east Bohemian dialect region, predominantly with male speakers. It is particularly well-documented in frequent words like pronouns (*won* ‘he’, *takowej* ‘such’), but also with regional expressions (*tkalcowat* ‘to weave’), toponyms (*chod’il na wístawi do Trutnowa* ‘he went to Trutnov for exhibitions’), and anthroponyms (*menoval se Schowánek* ‘his name was Schovánek’). In the north-east Bohemian dialect region, we can also find cases where the *l* phone and the *y* variant are lexicalized. Dark *l* is encountered predominantly in the past participle of verbs (*był* ‘he was’, *vzal* ‘he took’, *šel* ‘he went’, *voděřel* ‘he opened’, *dal* ‘he gave’), *y* is used in forms of the word *být* ‘to be’, and systematically in conjunction with dark *l* (before or after it), which indicates lexicalization (*była perleťej wíkládaná* ‘it was inlaid with nacre’).

Among other things, the companion website to the DIALEKT corpus lists many tips for working with the corpus, which can be useful e.g. for lexicographical work. A succinct overview is given of the basic techniques for working with the corpus, including search, display of various types of information, and sorting of search results. The process of creating both simple and more complex search queries,

convenient especially for lexicographers, is then documented using examples (searching e.g. for all words beginning with the letter *b*, for a specific word form, for a substring, for multi-word sequences, etc.). When searching for words beginning with a given letter, the lexicographer can also generate a frequency list of matching lemmas or word forms which can be used to determine whether a target lexical unit is present in the corpus or not, and if present, then in what precise form. In addition to these lexicographical tips, detailed instructions are provided on how to create subcorpora, which can also be very helpful in working with the corpus data.

4.2 Archive of dialect-specific differential phones of the Czech language

The companion website of the DIALEKT corpus also features an overview of the dialect phones [9] occurring on the territory of the Czech Republic and the corresponding symbols used in dialectological transcripts. For each dialect phone, an accessible description is given, complemented with information about territorial spread within the Czech Republic and examples of use in context in the form of a short recording excerpt and its transcript (*bělé mladi* ‘they were young’, *košyg je hotovy* ‘basket is finished’, *že byl* ‘that he was’, *p’ekně* ‘pretty’, *ośm’elil* ‘he mustered the courage’). The year of each recording is also indicated. These examples of differential phones specific to various Czech dialects were selected from the material of the corpus, i.e. from authentic dialect data. We strive to represent dialect vowels and consonants by the most typical examples available, avoiding transitional or otherwise unusual realizations. This archive of dialect-specific differential phones of the Czech language is not only a very useful tool for our own internal needs, but it can also serve other dialectologists, both professional and amateur, working on transcribing dialectological material. Last but not least, it constitutes an interesting teaching resource at all education levels.

4.3 Interactive map-based web environment

We are gradually putting together an interactive web environment which integrates language data from CNC corpora with a map-based interface. We are currently revising the classification of individual municipalities with respect to the system of dialect areas [11], which is a fairly demanding work, done in collaboration with cartographer K. Kupka and dialectologists from the Czech Language Institute of the Czech Academy of Sciences. One of the results of this effort should be an interactive map visually anchoring the recording locales of the DIALEKT corpus within the system of dialect areas. At selected representative dialect points of interest, it will be possible to display additional information about the characteristic features of the corresponding dialect region, division (*úsek*) and type (*typ*), as well as a short transcription sample together with an analysis and/or a recording of dialect speech. Users will also be able to cross-reference speakers and recordings from the same or neighboring locations across the DIALEKT corpus and other spoken corpora

of the CNC, making it easier to contrast and compare different sources of information about the territorial spread of a given phenomenon [12]. The interactive web app format offers advantages over the traditional visualization of language phenomena on static maps (e.g. in language atlases [13], [14]). It can be attractive not only for professional and classroom use, but also for the general public, both for educational and entertainment purposes, e.g. in the form of a quiz.

5 CONCLUSION

At the SLOVKO 2015 conference, plans for the DIALEKT corpus and related future outlooks were presented [10], which can now be confronted with what has actually been achieved so far. The intended final size of the corpus was cca 200,000 words. The first stage of the corpus, which is currently publicly available, totals 100,000 words. We expect that the second stage, which will be completed and published in 2020, will add another 100,000 words, thus reaching the initial goal. As initially planned, two transcription layers are available for searching – dialectological and orthographic transcripts. They are aligned to sound recordings, which are divided into short segments. For the moment being, it is not possible to play back entire recordings at once, but this remains a planned feature for one of the future versions of the corpus. It is encouraging that even in its first version, the DIALEKT corpus has already become a valuable resource both for linguistic research and teaching. For instance, it is one of the resources used in the process of compiling the territorially comprehensive Dictionary of Czech dialects. The data in the corpus were also used to create the Archive of dialect phones of the Czech language, and an interactive web environment which is currently in preparation and planned to be launched by the end of 2019.

ACKNOWLEDGMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

- [1] Goláňová, H., Waclawičová, M., Komrsková, Z., Lukeš, D., Kopřivová, M., and Poukarová, P. (2017). DIALEKT: nářeční korpus, verze 1 z 2. 6. 2017. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>
- [2] Goláňová, H. – Waclawičová, M. (2018). Co je v ČNK nového IX (Zprávy z českého národního korpusu). *Korpus – gramatika – axiologie*, 2018 (17), pages 78–82.

- [3] Waclawičová, M., and Goláňová, H. (2019). Nářeční korpus DIALEKT a jeho použití ve výuce češtiny. *Český jazyk a literatura*, 69(3), pages 127–133.
- [4] Balhar, J. et al. (2011). *Český jazykový atlas Dodatky*. Praha, Academia.
- [5] Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2017). ORTOFON: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>
- [6] Dialektologická komise České akademie věd a umění (1951). *Pravidla pro vědecký přepis dialektických zápisů českých a slovenských*. Praha, Česká akademie věd a umění.
- [7] Lamprecht, A., and Michálková, V. et al. (1976). *České nářeční texty*. Praha, Státní pedagogické nakladatelství.
- [8] Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L. and Křen, M. (2017). ORAL: korpus neformální mluvené češtiny, version 1 as of 2 June 2017. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>
- [9] Goláňová, H., and Waclawičová, M. (2019). Archiv diferenčních hlásek nářečí českého jazyka. Version as of 26 February 2019. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>.
- [10] Goláňová, H. (2015). A new dialect corpus: DIALEKT. In Gajdošová, K., and Žáková, A. (eds.). *Proceedings of the Eight International Conference Slovko 2015 (Natural Language Processing, Corpus Linguistics, Lexicography)*. Lüdenscheid, RAM-Verlag, pages 36–44.
- [11] Goláňová, H., and Kupka, K. (2019). *Mapa nářečí českého jazyka*. Version as of 9 January 2019. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>
- [12] Goláňová, H., Kopřivová, M., Lukeš, D., and Štěpán, M. (2015). Kartografické a geografické zpracování dat z mluvených korpusů. *Korpus – gramatika – axiologie*, 2015 (11), pages 42–54.
- [13] Balhar, J. et al. (1999, 2002, 2005). *Český jazykový atlas 3, 4, 5*. Praha, Academia.
- [14] Balhar, J., and Jančák, P. et al. (1992, 1997). *Český jazykový atlas 1, 2*. Praha, Academia.

ANNOTATIONS IN THE CORPUS OF TEXTS OF STUDENTS LEARNING SLOVAK AS A FOREIGN LANGUAGE (ERRKORP)

MICHAELA MOŠAŤOVÁ¹ – KATARÍNA GAJDOŠOVÁ²

¹ Faculty of Arts, Comenius University, Bratislava, Slovakia

² Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences,
Bratislava, Slovakia

MOŠAŤOVÁ, Michaela – GAJDOŠOVÁ, Katarína: Annotations in the corpus of texts of students learning Slovak as a foreign language (ERRKORP). *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 345 – 357.

Abstract: The article presents the upcoming acquisition corpus of written texts of students learning Slovak as a Foreign Language and focuses on the annotation of texts, which includes information about the text as well as social and linguistic details about the student. The article also discusses the tags that identify individual errors in the texts and concept of creating the tagset itself.

Keywords: language error, learner corpus, slovak, tagging, annotation

1 INTRODUCTION

Language errors are immanently present in the process of learning any foreign language. The student as well as the teacher are constantly confronted with not quite successful or even unsuccessful written and spoken communication. Which is why the identification, interpretation and didactic reflection of language errors are an inherent part of teaching any foreign language – especially with the intention of reduction and prevention.

For some time, we have been aware of the need for a complex analysis of language errors in the field of Slovak as a Foreign Language (SFL) which, in 2018, led to the foundation of the collaborative project of the Studia Academica Slovaca Center at the Comenius University Faculty of Arts in Bratislava (SAS) and the Department of the Slovak National Corpus at the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences in Bratislava (SNC) with the aim of creating an acquisition corpus of written texts authored by foreigners learning SFL. Acquisition collections are specialized text collections with the primary use of studying processes related to the acquisition of a specific language and to its teaching. When building our corpus, we have been inspired by those of related Slavic languages, mainly Czech [1], but also the similar collections of English [2], German [3] and Russian languages [4].

The article presents the preparation of the first publicly accessible acquisition corpus of texts by foreigners learning SFL, entitled ERRKORP, including data collection and their metadata. The article also discusses the set of tags designed for manual annotation of collected texts. We have tested the proposed tagset on a collection of 65 texts written by students learning Slovak as a foreign language from most of the language proficiency levels (A1–C1) and a number of proveniences (Ukraine, Serbia, Italy, China, Belarus, USA, Australia). The tagset has grown more stable and precise in the course of the annotation process. We present its final version in Chapter 4. In the near future, we are planning to tag manually texts selected from the digital storage of the project and thus create a pilot version of a corpus of students learning Slovak as a foreign language. The pilot corpus version will not include a collection of testing texts.

2 DATA COLLECTION

There are several parallel methods and several places for the collection of data for creating the corpus. Primarily, the SAS Center is the main provider of texts for the corpus. The Center is in possession of written work by students learning SFL at the Summer School of Slovak language and culture – an event held regularly for the past 55 years, and there are also other courses taught to foreigners at the SAS center throughout the year. Texts are also shared by our visiting lecturers of Slovak Language and Culture affiliated with universities across Europe and elsewhere. The collection of texts follows the regulations of the GDPR.

Handwritten texts are being digitized and converted into text form. We understand text as a text written independently by a student who already speaks and/ or is studying SFL at a certain language proficiency level (A1–C2).

At the SNC, there is now a digital hub collecting texts written by foreigners learning SFL on the territory of Slovakia and beyond. The texts are entered online at <https://errkorp.juls.savba.sk>, and the provider of the texts enters information about the text as well as the Slovak speaker. At present, the corpus archive consists of 898 texts from 212 students coming from 34 countries worldwide.

3 DOCUMENT METADATA

For exact explication in terms of further research it is important to have access not only to the text itself but also to detailed input about the conditions in which the text was written. Also important is some information about the author of the given text, because as it is well known, a number of extra-linguistic factors contribute to the making of mistakes [5].

Metadata about the student is itemized in Table 1. A combination of entries will enable the creation of individual subcorpora with concrete specifications, these will serve for the research of a specific linguistic phenomenon or of a specific language area which has proven problematic with a given linguistic group.

Item description	Item	Constant values
student's name and surname ¹	spk_name	
sex	spk_sex	<ul style="list-style-type: none"> ● male ● female
age	spk_age	
age group / decade	spk_decade	
highest degree earned	spk_edu	<ul style="list-style-type: none"> ● elementary ● secondary ● Bachelor's degree or equivalent ● Master's degree or equivalent ● Doctor's degree or equivalent
information about the student's professional involvement with the language	spk_jobling	<ul style="list-style-type: none"> ● yes ● no
student's country of origin	spk_country1	
student's current long-term residence / place of study	spk_country2	
mother tongue / first language	spk_L1	
other languages the student speaks, based on proficiency level e.g. ru-B2, en-B1, de-A1	spk_languages	
language proficiency in the Slovak language	spk_level_ERR	<ul style="list-style-type: none"> ● A1.1 ● A1.2 ● A2.1 ● A2.2 ● B1 ● B2 ● C1 ● C2
contact with Slovak speakers outside of school	spk_contact	<p>More than one option applicable:</p> <ul style="list-style-type: none"> ● none ● parent ● partner, other family member ● friend or colleague

¹ This information is for internal purposes only. In the corpus, a student's name and surname will not be revealed.

Item description	Item	Constant values
kind of stay in Slovakia	spk_stay_SR	More than one option applicable: <ul style="list-style-type: none"> • none • study • work • other (visitor, family, vacation) • asylum or protected status • permanent residence
length of stay in Slovakia	spk_years_SR	<ul style="list-style-type: none"> • less than 6 months • 6 – 12 months • 1 – 2 years • 2 years or longer
method of studying Slovak	spk_learning	<ul style="list-style-type: none"> • individual study with a teacher • corporate/private lessons • self-study • college, university – in Slovak • secondary education in Slovak • primary education in Slovak
other method of studying Slovak	spk_learning_other	
duration of Slovak study	spk_learning_years	<ul style="list-style-type: none"> • less than 6 months • 6 – 12 months • 1 – 2 years • 2 years or longer
frequency of contact with Slovak in hours per week	spk_learning_hours_week	<ul style="list-style-type: none"> • less than 5 hours • 5 – 15 hours • over 15 hours
textbook used	spk_textbook	<ul style="list-style-type: none"> • selection from a list, e.g. Križom-krážom A1

Tab. 1. Overview of metadata about the speaker in the upcoming ERRKORP corpus

Metadata about the text is itemized in Table 2. Their combination enables us to create suitable subcorpora with relevant values for text-oriented research, e.g. monitoring certain error types in texts written during institutional testing and their comparison with the same errors in texts from other environments, such as during self-study.

Naturally, the combination options of texts allow a great variety, especially when involving specific metadata about the student.

Item description	Item	Constant values
date text was written	doc_date	YYYY-MM-DD
place where text was written	doc_place	
origin of text	doc_origin	<ul style="list-style-type: none"> • at learning institution (school, course) • outside of learning institution
type of text	doc_type	<ul style="list-style-type: none"> • student's own (creative) writing/text • translation from another language
original language in case of a translation	doc_type_lang	if translation – from what language
text form	doc_medium	<ul style="list-style-type: none"> • hand-written • electronic
text as part of testing	doc_text_test	<ul style="list-style-type: none"> • yes • no
time limit for writing the text	doc_time_limit	<ul style="list-style-type: none"> • yes • no
materials allowed	doc_materials	<ul style="list-style-type: none"> • yes • no
type of material	doc_material_type	<ul style="list-style-type: none"> • dictionary • other
material other than dictionary	doc_material_type_other	
text title based on assignment	doc_topic	
word count	doc_word_limit	<ul style="list-style-type: none"> • less than 50 words • 50 – 150 words • over 150 words
keywords	doc_keywords	
genre of text	doc_genre	<ul style="list-style-type: none"> • description / informational text • journalism – reporting • journalism – analyses • journalism – other • artistic genres / fiction • argumentation • non-fiction • administrative text • e-mail correspondence • other / unspecifiable

Tab. 2. Overview of metadata about the text in the upcoming ERRKORP corpus

Since with some of the texts (especially archive ones) it is not possible to identify all the items of the metadata about the text and the student, it is important to define the elementary items, which include metadata about the student – sex, country of origin, mother tongue and language proficiency level. If the given items cannot be found in archived files about a certain text, this text cannot be entered in the database.

4 ANNOTATION SCHEME

When creating our annotation tagset, we were inspired by the work of Stephen Pit Corder [6], Carl James [7], academic articles by Czech authors K. Šebesta and S. Škodová ([8], [9]), R. Kotková [10] and the CzeSL-SGT-cs corpus annotation tagset [11].

When annotating linguistic material, we usually apply a combination of two annotation methods.

1. First, we based our annotations on the concept of functional linguistic typology, which differentiates whether the error in the written text is:

a) on the level of a single segment – a matter of a single grapheme, diacritic, or punctuation mark and usually the error comes from the orthographic or phonetic and phonological levels of the language;

b) on the level of a word segment – thus identified errors most commonly occur in the case of a grammatical relational morpheme or derivational morpheme (prefix or suffix) or in the case of the stem, and so these are errors related to the morphological or derivational level of the language;

c) on the level of a word – these are errors on the semantic level of the language, and style. Sometimes the errors occur on the morpho-syntactic level (omitted auxiliary verbs, reflexive pronouns with reflexiva tantum),

d) on the level of a phrase within a single sentence – these are usually morpho-syntactic errors: wrong congruence, wrong word order of enclitics, and lexical and stylistic errors (incorrect usage of phrases and idioms);

e) on the level of text – such errors exceed the sentence structure, these are errors of style or of pragmalinguistical character (terms unfitting a particular style of text, incorrect usage of text connectors, linking words, etc.).

2. Simultaneously, we evaluate almost every error also in terms of surface typology, according to which every error occurs as one of three options:

a) omission, or absence,

b) addition,

c) substitution (of a segment – grapheme, morpheme, word, etc.).

Annotation tags usually consist of two types of information: 1) which part of the annotated written text the error is related to (a single grapheme, diacritic, word, phrase, etc.) and 2) in what form the error occurs (omission, redundancy, or substitution). For all possible text segments that are affected by errors and which are

also relevant for further linguo-didactic research, we chose specific tags such as comma, char for a grapheme, quant for accent (see the Tables below). In terms of surface typology, we use one of the following three tags: 0, 1 or subst, e.g. if the text contains a redundant word, e.g. **budem napísať* (*napíšem* – ‘I will write down’), we use the tag word1. If a word is omitted, e.g. **opýtal ma* (*opýtal sa ma* – ‘he asked me’), we use the tag word0 and if the student used an unfitting word, which can be replaced by a different lexeme in Slovak, e.g. **idem do lekára* (*idem k lekárovi* – ‘I am going to the doctor’), we annotate with the tag substword.

However, there are some possible combinations that we decided not to include in the annotation tagset due to their zero or minimal frequency in texts: for example substcomma (although the tagset includes the tags comma0 and comma1), because this kind of error did not occur during a test of annotating.

4.1 Errors on the level of a single segment

Table 3 presents the tags on the level of a single segment.

Tag	Description	Examples (correct word/ collocation – translation)
char0	missing grapheme	<i>všeci</i> (<i>všetci</i> – ‘all’), <i>moe</i> (<i>moje</i> – ‘my’)
char1	addition of a grapheme	<i>zamrzlina</i> (<i>zmrzlina</i> – ‘ice-cream’)
charmata	exchange of subsequent graphemes	<i>Sbrsko</i> (<i>Srbsko</i> – ‘Serbia’), <i>ked</i> (<i>kde</i> – ‘where’)
substvow	vowel substitution	<i>krieslo</i> (<i>kreslo</i> – ‘chair’), <i>hidiny</i> (<i>hodiny</i> – ‘clock’)
substcons	consonant substitution	<i>úloga</i> (<i>úloha</i> – ‘task/homework’), <i>tažka</i> (<i>taška</i> – ‘bag’)
substdiph	diphthong substitution	<i>stol</i> (<i>stól</i> – ‘table’), <i>možem</i> (<i>môžem</i> – ‘I can’)
cap0	omission of capital letter	<i>minčania</i> (<i>Minčania</i> , <i>obyvatelia Minska</i> – ‘inhabitants of Minsk’)
cap1	addition of a capital letter	<i>Európska Únia</i> (<i>Európska únia</i> – ‘European Union’)
alt	error in alternation	<i>časniki</i> (<i>časnici</i> – ‘waiters’) <i>listoky</i> (<i>listky</i> – ‘tickets’), <i>mám pesa</i> (<i>mám psa</i> – ‘I have a dog’), <i>plakám</i> (<i>plačem</i> – ‘I am crying’)
quantbase0	omission of quantity in the root	<i>kamarat</i> (<i>kamarát</i> – ‘friend’), <i>kulturny</i> (<i>kultúrny</i> – ‘cultural’)
quantbase1	addition of quantity in the root	<i>téplo</i> (<i>teplo</i> – ‘warm/hot’), <i>próblem</i> (<i>problém</i> – ‘problem’)

Tag	Description	Examples (correct word/collocation – translation)
quantpref0	omission of quantity in the prefix	<i>vyber</i> (<i>výber</i> – ‘choice’)
quantpref1	addition of quantity in the prefix	<i>výdanie</i> (<i>vydanie</i> – ‘edition’)
quantsuf0	omission of quantity in the suffix	<i>novy</i> (<i>nový</i> – ‘new’), <i>tuto osobu</i> (<i>túto osobu</i> – ‘this person’)
quantsuf1	addition of quantity in the suffix	<i>môžu</i> (<i>môžu</i> – ‘they can’), <i>krátké</i> (<i>krátke</i> – ‘short’)
y1	substitution of i – y	<i>sir</i> (<i>syr</i> – ‘cheese’), <i>umit’</i> (<i>umyt’</i> – ‘to wash’)
y0	substitution of y – i	<i>knižnyca</i> (<i>knižnica</i> – ‘library’), <i>boly</i> (<i>boli</i> – ‘they were’)
substchar	substitution of characters	<i>jedlo. Ktoré sa volá</i> (<i>jedlo, ktoré sa volá</i> – ‘a dish called’)
caron0	omission of the softness mark (ˇ)	<i>den</i> (<i>deň</i> – ‘day’), <i>cervený</i> (<i>červený</i> – ‘red’)
caron1	addition of the softness mark (ˇ)	<i>vo Viední</i> (<i>vo Viedni</i> – ‘in Vienna’), <i>útulný</i> (<i>útulný</i> – ‘cozy’)
comma0	omission of a comma	<i>Ludia si myslia že</i> (<i>Ludia si myslia, že</i> – ‘People think that’)
comma1	addition of a comma	<i>V Číne, máme dve možnosti</i> (<i>V Číne máme dve možnosti</i> – ‘In China, we have two options’)
dot0	omission of a period	<i>13 januára 2016</i> (<i>13. januára 2016</i> – ‘January 13, 2016’)
dot1	addition of a period	<i>v 2014. roku</i> (<i>v roku 2014</i> – ‘in 2014’)
hyph0	omission of a hyphen	<i>križomkražom</i> (<i>križom-kražom</i> – ‘criss-cross’)
hyph1	addition of a hyphen	<i>rímsko-katolícky</i> (<i>rímskokatolícky</i> – ‘Roman Catholic’)
defword	three or more substitutions of characters in a word (in the root) simultaneously – with the exception of incorrect diacritics	<i>samuslina</i> (<i>zmrzlina</i> – ‘ice-cream’), <i>popreč</i> (<i>cez</i> – ‘through’)
defdiacr	three or more errors in diacritics	<i>zaujimáva</i> (<i>zaujímavá</i> – ‘interesting’), <i>mozú</i> (<i>môžu</i> – ‘they can’)

Tab. 3. Annotation tags for errors on the level of a single segment

Separately, we analyzed a set of error that are quite frequent in the Slovak language: the substitution of vowels, consonants and diphthongs. Substitution of a vowel means that a different vowel or a diphthong is used to replace the correct vowel, e. g. *diesat'* (*desat'* 'ten'). For the quantity of vowels and the syllabic "r" and "l" (accent) – which is linguodidactically one of the most complicated and most time consuming issues for foreigner learning SFL – we used three different tags: quantity in the prefix, quantity in the suffix (= relational morpheme) and quantity in the stem. However here, by stem we do not understand the strictly linguistic terms of stem but the part of word left after cutting off the grammatical relational morpheme and also possibly the derivational prefix, e.g. *zá-hraničn-ý* (*zahraničný* 'foreign'). This way will allow us to eventually analyze efficiently the cases where errors occur in derivational prefixes, grammatical suffixes and the stem.

The defword tag covers cases when several grapheme substitutions in a word means lack of understanding of the word, which is defective. In annotation testing, the defdiacr tag proved meaningful in cases when a word contains at least three orthographic errors but it is still comprehensible, even without diacritics (after all, present day written communication – chats, text messages, status posts on social networks, and even emails prove the usage of such texts even by native speakers).

4.2 Errors on the level of a morpheme or word

Table 4 introduces tags on the level of a single morpheme or word.

Tag	Description	Examples (correct word/collocation – translation)
word0	omission of a word	<i>opýtal ma</i> (<i>opýtal sa ma</i> – 'he asked me')
word1	addition of a word	<i>píšem s perom</i> (<i>píšem perom</i> – 'I write with a pen'); <i>ja sa volám Eva</i> (<i>volám sa Eva</i> – 'my name is Eva')
substword	substitution of a word	<i>tajná vôňa</i> (<i>tajomná vôňa</i> – 'mysterious fragrance'), <i>prezentovať parfum</i> (<i>darovať parfum</i> – 'to give perfume as a gift')
morph	substitution of a grammatical morpheme	<i>z galérii</i> (<i>z galérie</i> (sg.) / <i>z galérií</i> (pl.) – 'from a gallery (sg.) / from galleries' (pl.)), <i>dám svojim kamarátkam</i> (<i>dám svojim kamarátkam</i> – 'I will give ... to my friends')
substderiv	substitution of a derivational morpheme	<i>historitické</i> (<i>historické</i> – 'historical'), <i>sezónové akcie</i> (<i>sezónne akcie</i> – 'seasonal promotions')
gend	substitution of gender	<i>ten centrum</i> (<i>to centrum</i> – 'the center'), <i>ten esej</i> (<i>tá esej</i> – 'the essay')

Tag	Description	Examples (correct word/collocation – translation)
num	substitution of number	<i>koláč s ovociami</i> (<i>koláč s ovocím</i> – ‘fruit cake’), <i>ryža sú na tanieri</i> (<i>ryža je na tanieri</i> – ‘rice is on the plate’)
asp	error in aspect	<i>budem prísť</i> (<i>prídem</i> – ‘I’ll come’)
temp	error of tense	<i>zavolať, že o dva dni musel odísť</i> (<i>zavolať, že o dva dni musí odísť</i> – ‘he called to say that he had to leave in two days’)
defmorph	error in grammatical morpheme	<i>v mojej meste</i> (<i>v mojom meste</i> – ‘in my town’), <i>plno ľudej</i> (<i>plno ľudí</i> – ‘a lot of people’)

Tab. 4. Annotation tags for errors on the level of a morpheme or a word

The tag substword indicates the types of mistakes where a word is substituted based on formal similarity of words or semantics. Often, these can be seen as cases of interlinguistic homonymy (“false friends”). Similar are cases on the level of the derivational morpheme, these are errors tagged as substderiv. The tag defmorph is justified when a student used a grammatical morpheme that is absent in the corresponding paradigm of the given part of speech (it is more a case of interference from the student’s L1 or from another language).

4.3 Errors on the level of word phrases within one sentence and errors on the level of text

All tags in Table 5, especially order, congr, neg, phrase, arise from the linguodidactic need of SFL teaching to explore, on a statistically relevant sample of linguistic data, the real scope of their occurrence on the various language proficiency levels (from beginners to advanced language users) as well as in relation to language L1. The annotation of these errors shifts from verbal specification to larger sequences exceeding word sequence.

Tag	Description	Examples (correct word/collocation – translation)
order	word order error	<i>Tu nachádza sa veľa parkov</i> (<i>Tu sa nachádza veľa parkov</i> – ‘There are many parks here’)
congr	congruence error	<i>veľa ľudí boli</i> (<i>veľa ľudí bolo</i> – ‘there were many people’), <i>päť študentov majú</i> (<i>päť študentov má</i> – ‘five students have’)

Tag	Description	Examples (correct word/collocation – translation)
neg	negation error	<i>nie budem</i> (<i>nebudem</i> – ‘I will not’), <i>nikto prišiel</i> (<i>nikto neprišiel</i> – ‘nobody came’)
space	error in splitting of words	<i>nie len</i> (<i>nielen</i> – ‘not only’), <i>preto že</i> (<i>pretože</i> – ‘because’)
styl	unfitting choice of words in terms of style	<i>na konferencii papkáme obed</i> (<i>na konferencii jeme obed</i> – ‘we eat lunch at the conference’)
phrase	literary translation of a phrase from another language	<i>nemá rozprávania</i> (<i>niet o čom</i> – ‘nothing to talk about’), <i>ich je dve</i> (<i>sú dve</i> – ‘there are two’)
pragm	incorrect usage within the text	<i>Ahojte, pani profesorka!</i> (<i>Dobrý deň, pani profesorka!</i> – ‘Good morning, Miss X.’)
theme	error in the information structure of a sentence (theme/rheme)	<i>Moja mama rada varí. Veľmi dobre varí.</i> (<i>Moja mama rada varí. Varí veľmi dobre.</i> – ‘My mom enjoys cooking. She’s a great cook.’)
connect	connector error (hypersyntax)	<i>V texte sa zaoberáme slovenčinou. Jeho používanie v rôznych situáciách je ovplyvnené... (V texte sa zaoberáme slovenčinou. Jej používanie v rôznych situáciách je ovplyvnené... – ‘In the text, we deal with the Slovak language. In various situations, its usage is influenced by...’)</i>

Tab. 5. Annotation tags for errors on the level of a phrase within a sentence and on the level of text

5 ANNOTATION PROCESS

In the section about data collection we talked about converting digitized texts of students into texts. This is step zero in the process of setting up material for the acquisition corpus of texts. The following step is the manual annotation of errors based on the tagset described above. Since the project has been running on a low budget, with limited personnel capacity and high load for the annotation team, we opted for using the simplest possible annotation methods and tools. After tokenization, the annotator marks the text in a cvs file with three columns (see Table 6). The first column contains the student’s original text vertically. In the second column, the annotator enters the corrected version of the errors from the relevant rows in the first column. The third column shows the relevant tag for the errors, based on the annotation tagset. Several errors in a single word are marked by the annotator next to each other, separated by commas.

<s>		
Ráno		
@@@		
vstála	vstala	quantbase1
som		
@@@*		
som		
vstala		
*@@@order		
,		
umyvala	umývala	quantbase0
NONE	som	word0
NONE	sa	word0
,		
obliekala		
son	som	substcons
NONE	si	word0
cerveni	červené	caron0,quantsuf0,morph
tricko	tričko	caron0
.		
</s>		

Tab. 6. Sample of a manual text annotated

The presented tagset of errors covers the most frequently occurring errors in written texts by foreigners learning SFL. We believe that the combination of above-mentioned two annotation methods creates the conditions for the most precise classification of errors with the objective of appropriate tagging of data needed for further analytical and explanatory part of research in the field of Slovak as a Foreign Language.

6 CONCLUSION

In the article, we introduced the concept of annotations of texts in the project of the ERRKORP acquisition corpus. In terms of the project's schedule [13], in the following months we will choose from the so far collected data in the data hub specific texts in order to include them in the pilot version of the corpus. The selected texts will be tagged manually, following the presented annotation tagset, which is also accessible free of charge as a corpus within the Slovak National Corpus.

It will also serve for other kinds of research in the field of applied linguistics and the teaching of Slovak as a Foreign Language focusing on the study of learning SFL from various aspects (such as research of the types of errors on the respective proficiency levels, longitudinal research of errors made by individuals, or research of errors against the background of the chosen starting language, etc.).

References

- [1] Šebesta, K., Bedřichová, Z., Šormová, K., Štindlová, B., Hrdlička, M., Hrdličková, T., Hana, J., Petkevič, V., Jelínek, T., Škodová, S., Poláčková, M., Janeš, P., Lundáková, K., Skoumalová, H., Sládek, Š., Pierscieniak, P., Toufarová, D., Richter, M., Straka, M., and Rosen, A. (2014). CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací, version 2 as of 28 July 2014. Ústav Českého národního korpusu FF UK, Praha. Accessible at: <http://www.korpus.cz>.
- [2] Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). International Corpus of Learner English v2 (Handbook + CD-Rom). Louvain-la-Neuve, Presses universitaires de Louvain, Louvain-la-Neuve.
- [3] Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., and Torsten, A. (2012). Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Accessible at: <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/view>.
- [4] RLC. Russian Learner Corpus. Accessible at: <http://web-corpora.net/RLC/>.
- [5] Pekarovičová, J. (2004). Slovenčina ako cudzí jazyk. Predmet aplikovanej lingvistiky. Bratislava, Stimul. 220 p.
- [6] Corder, S. P. (1981). Error Analysis and Interlanguage. Oxford, Oxford University Press, 120 p.
- [7] James, C. (1998). Errors in Language Learning and Use. London – New York, Longman, 304 p.
- [8] Šebesta, K., Škodová, S. et al. (2012). Čeština – cílový jazyk a korpusy. Liberec, Technická univerzita v Liberci. 168 p. Accessible at: <http://akces.ff.cuni.cz/system/files/ce%20-%20c%3ADlov%3BD%20jazyk.pdf>.
- [9] Škodová, S., Štindlová, B., Rosen, A., Jelínek, T., and Vidová Hladká, B. (2019). Příručka k morfologické anotaci textů nerodilých mluvčích češtiny. Version 1.0 as of 17 January 2019. Accessible at: http://utkl.ff.cuni.cz/~rosen/public/2018_pri-rucka_morfologicke_anotace.pdf.
- [10] Kotková, R. (2017). Čeština nerodilých mluvčích s mateřským jazykem neslovanským. Praha: Univerzita Karlova, 154 p.
- [11] CzeSL-SGT – korpus češtiny nerodilých mluvčích s automaticky provedenou anotací. Manuál. Accessible at: <http://utkl.ff.cuni.cz/~rosen/public/2014-czesl-sgt-cs.pdf>.
- [12] Kilgariff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1, pages 7–36. Accessible at: <http://www.sketchengine.eu>.
- [13] Project of the corpus of texts written by students learning Slovak as a foreign language – ERRKORP. Accessible at: <https://korpus.sk/errkorp.html>.

PARTS OF SPEECH IN NOVAMORF, A NEW MORPHOLOGICAL ANNOTATION OF CZECH

VLADIMÍR PETKEVIČ¹ – JAROSLAVA HLAVÁČOVÁ² – KLÁRA OSOLSOBĚ³ –
MARTIN SVÁŠEK – JOSEF ŠIMANDL

¹ Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University, Czech Republic

² Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic

³ Institute of the Czech Language, Faculty of Arts, Masaryk University, Czech Republic

PETKEVIČ, Vladimír – HLAVÁČOVÁ, Jaroslava – OSOLSOBĚ, Klára – SVÁŠEK, Martin – ŠIMANDL, Josef: Parts of speech in NovaMorf, a new morphological annotation of Czech. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 358 – 369.

Abstract: A detailed morphological description of word forms in any language is a necessary condition for a successful automatic processing of linguistic data. The paper focuses on a new description of morphological categories, mainly on the subcategorization of parts of speech in Czech within the NovaMorf project. NovaMorf focuses on the description of morphological properties of Czech word forms in a more compact and consistent way and with a higher explicative power than approaches used so far. It also aims at the unification of diverse approaches to morphological annotation of Czech. NovaMorf approach will be reflected in a new morphological dictionary to be exploited for a new automatic morphological analysis (and disambiguation) of corpora of contemporary Czech.

Keywords: NovaMorf, morphological annotation, parts of speech, morphological categories, subcategorization

1 INTRODUCTION

We present a repertoire of morphological categories and, mainly, the parts of speech (POS) and their subcategorization distinguished in NovaMorf, the project of an innovated description of Czech morphology as a linguistic base for a new morphological analysis and subsequent disambiguation of Czech texts. For over 25 years, morphological dictionary and analysis of Czech were based on (almost) unchanged annotation systems. After years of experience with the use of language corpora, it turned out that these systems, designed many years ago, have become somewhat obsolete and need to be amended, concerning the system itself, the tagset, and morphological dictionary used by morphological analysis.

The NovaMorf starting point is [3]. Other suggestions for solving partial problems are based on it ([5], [6], [7], [10], [11]). NovaMorf is also based on the Prague morphological annotation system ([1], [2]) and Ajka/Majka system developed in Brno ([4], [8], [9], [12]) on recent grammars of Czech and corpus data. NovaMorf critically evaluates these systems and creates a new one so that the resulting description is linguistically adequate, economical and consistent (e.g., unlike current systems, it consequently applies the so-called “golden rule of morphology”¹). It also takes into account the Universal Dependencies approach, trying to be very close to it.² This new annotation system, as a national standard, is to be used for morphological annotation of corpora of contemporary Czech and in various NLP applications (e.g., parsing, spell and grammar checking) dealing with Czech.

We only briefly introduce:

- (i) the repertoire of suggested morphological categories (Chap. 2),
- (ii) the repertoire of parts of speech and their subcategorization (Chap. 3).

2 MORPHOLOGICAL CATEGORIES

We design the following categories with each category being assigned a set of values. Each morphological interpretation of a word form is assigned just one value for a given category (except for global and inflectional mutations).

1. Part of speech – POS (cf. 3.1)
2. POS subcategorization – SUB (cf. 3.2.1)
3. Deixis – DEI (cf. 3.2.2)
4. Aspect – ASP
5. Abbreviation – ABR
6. Gender – GEN
7. Number – NUM
8. Case – CAS
9. Person – PER
10. Degree of Comparison – DEG
11. Negation – NEG
12. Verb form – VRB
13. Adjectival Short Form – NOM
14. Aggregate/Compound Type – AGR
15. Global mutation – GMU
16. Inflectional mutation – FMU

Let us now characterize individual parts of speech.

¹ The rule states that two different word forms cannot be assigned an identical annotation (= [lemma, tag] pair).

² The detailed description of the NovaMorf approach also contains a detailed comparison with the UD approach. For scope reasons, no details can be presented here.

3 PARTS OF SPEECH

3.1 POS category

The part of speech category (POS) is the basic one since each morphological interpretation of a word form is assigned a POS value.

Values of the POS category:

- Noun (code N)
- Adjective (A)
- Pronoun (P)
- Numeral (C)
- Verb (V)
- Adverb (D)
- Preposition (R)
- Conjunction (J)
- Interjection (I)
- Particle (T)
- Foreign word (F)
- Affix segment (S)
- Symbol (Z)
- Unknown word (X)
- Aggregate (G³)

In addition to traditional POS (noun ... particle, cf. 3.2.1ff.), we introduce the following “parts of speech” in a broader sense.

Foreign word – code F

Foreign word is a string that does not have its own meaning in Czech; typically a word of a foreign language occurring in a Czech sentence, usually within a quotation or saying (*we shall overcome; per se*) or as part of a proper name (*New York City*). Typical examples: *the, you, du, to*. This does not concern loan words which are part of the vocabulary of Czech (*image, khaki*).

Its lemma = the form itself. Foreign words have no subcategories.

Affix segment – code S

Affix segment is a string that is typically marked with a separator: a hyphen, space, slash. We distinguish:

- numeral prefix segment (C)
- postfix segment (p)
- other prefix segments (e)

³ The code of this POS is not used since aggregates are described by their components only, see below.

(i) **Numeral prefix segment:** an initial numeral segment of a word standing alone and being added to the full word further on in the text:

- *tří až čtyřprocentní* ‘three to four percent’ (prefix segment: *tří*)

(ii) **Postfix segment:** a final segment of a word that becomes a full word by adding a part of a previous string in front of it. Its lemma = the form itself.

- *řekl/a* ‘he said / she (said)’ (postfix segment: *a*, lemma(*a*) = *a*)

Unlike a prefix segment, it can be assigned values of some morphological categories being derived from the entire word form they abbreviate (thus *a* is a postfix segment that is assigned past participle values; this may sound paradoxical, but a relation to the full verb *řekl* is established).

(iii) **Other prefix segments:** an initial non-numeral segment of a word standing alone, but forming a full word with a substring of another string further on in the text:

Examples:

- *česko-* a *rusko-německý* ‘Czech- and Russian-German’ (prefix segments: *česko*, *rusko*)

Symbol – code Z

Linguistically, symbol is not one of the traditional parts of speech, but it is very useful to assign – along the lines of existing annotation systems – each symbol the same code (Z), and to understand the set of symbols as a special word class. The lemma of a symbol is typically a symbol itself, but inverted commas, apostrophes and various kinds of brackets are normalized. Symbols are divided into:

- (i) punctuation marks (z) (e.g. comma: “,”)
- (ii) other signs (J) (e.g. % or math symbols)

Unknown word – code X

An unknown word is a string whose POS cannot be recognized, typically a typo or a foreign language word not being classified as a foreign word (see above); it is not contained in the morphological dictionary. Its lemma = its form.

This value is already present in existing annotation systems.

Aggregate

An aggregate is a special word class reserved for describing a word form, consisting of a combination of 2 or 3 word forms – aggregate components that may

belong to different classical parts of speech. Therefore, none of them can be assigned to the aggregate, which is described via its components only, each component being assigned its own tag including POS. The lemma of an aggregate is a *multiple lemma* – an ordered set of component lemmas.⁴ For instance:

- *připravilas* ‘you prepared’ consists of two verbal components: *připravila* ‘prepared’ and the 2nd pers. sg. enclitic *-s* (= *jsi* ‘you_are’), with each component being assigned its own tag; the lemma is a multiple one: lemma(*připravilas*) = {*připravil*, *být*} ‘{prepare, be}’
- *abyste* ‘so that you’ consists of the conjunction *aby* ‘so_that’ and the present tense 2nd pers. pl. form of the verb *být* ‘be’: each component is assigned its respective tag, the multiple lemma(*abyste*) = {*aby*, *být*} ‘{so_that, be}’

The aggregate is not considered a compound word made up via a usual word formation process; a compound word is assigned a classic POS (adjective *černobílý* ‘black and white’) and has other morphological properties as well.

3.2 POS categories

The majority of parts of speech is further subcategorized, some of them (pronouns, numerals, adverbs, some nouns and adjectives of the “numeral type”) being subcategorized into two categories:

- subcategory SUB (3.2.1)
- subcategory, called Deixis (DEI, 3.2.2).

3.2.1 SUB category

The SUB category is relevant to all parts of speech, except for prepositions, interjections, foreign words, aggregates, and unknown words. For each relevant POS, we specify what SUB category values are distinguished and how they are encoded. The corresponding one-letter code is usually the same as the second code in the Prague system ([1], [2]), which is also called Subcategory. In NovaMorf, however, this category is interpreted quite differently. Subcategory in the current Prague system is a mixture of values describing various characteristics of word forms. Some values relate to individual forms, other ones to whole paradigms. On the contrary, the SUB category in NovaMorf is **strictly global**, i.e. it is always relevant to the whole paradigm of a given word form. The same applies to the second subcategory: Deixis (3.2.2).

The SUB category values are different for different parts of speech since they describe different properties. However, some properties (some SUB and DEI

⁴ Multiple lemmas remain assigned to specific word forms (aggregates, passive participles and forms described by global or inflectional mutations) even after disambiguation in POS tagging: they are not to be further disambiguated. A user can search via each component of a multiple lemma in a corpus.

category values) are shared by more parts of speech, hence they are assigned the same value (code). Unshared features are encoded with different, unambiguous codes. Below is an overview of the shared values of the SUB category.

- **Other** (code 0 – zero). It concerns nouns, adjectives, verbs, adverbs and indicates that a word form does not have any of the other properties distinguished in the category, thus it is not necessary to express this property by different codes depending on POS.
- **Possessive** (code U) is shared by adjectives and pronouns.
- **Deverbal** (code V) is shared by nouns, adjectives and adverbs. It specifies that a given word is derived from a verb (code V is proposed being also the code for Verb as POS). The value is assigned to a specific group of words, rather than to all words derived from verbs.
- **Numeral** (code C) is relevant to nouns, adjectives, adverbs, and numeral prefix segments. It specifies that a given word contains an element common to numerals, i.e. expressing a number, or possibly the word is used instead of a numeral (e.g. *tisícovka* návštěvníků ‘a thousand visitors’). Thus, the code C is proposed being also the code for Numeral as POS. In some grammars, such words are referred to as belonging to specific subcategories of numerals.

3.2.1.1 Noun subcategorization

For nouns, the following values of the SUB category are distinguished only:

- Deverbal (V): *pokrytí* ‘covering’...
- Numeral (C): words expressing an association with numerals, e.g. *pětka* ‘five’...
- Other (0 – zero): *město* ‘town’...

The V value is assigned to deverbal nouns ending with the *-ní* / *-tí* suffix. These nouns have specific (morpho)syntactic properties and can behave differently than other nouns in a sentence: unlike the other nouns they can have a reflexive particle associated with them (*štitění se práce* ‘loathing work’) or they can be modified by adverbs regularly derived from adjectives (*zpívání falešně* ‘singing out of the tune’).

We do not distinguish between deverbal nouns and lexicalized deverbal nouns that do not express a verbal action: *vázání ječmene* ‘tying barley’ vs. *lyžařské vázání* ‘ski bindings’. In the morphological dictionary, they will constitute a single entry: *vázání*, with SUB=V.

The nouns irregularly derived from verbs are not considered deverbal: e.g. *utrpení* ‘suffering’ (other is *utrpení* ‘suffering’), *uvědomění* ‘awareness’... Nor the compound words such as *krupobití* ‘hailstorm’ are considered deverbal, although their second component is an action noun (= *bití* ‘beating’), since corresponding compound verbs usually do not exist (**krupobít* ‘to hailstorm’).

3.2.1.2 Adjective subcategorization

For adjectives, we distinguish the following SUB category values:

- Possessive (U): *matčín* ‘mother’s’...
- Derived from present transgressives (G): *sedící* ‘sitting’...
- Derived from past transgressives (M): *přeživší* ‘having survived’...
- Other deverbal (V): passive participle forms (*namazán* ‘lubricated’⁵; adjectives ending with *-ný* and *-tý* (*namazaný* ‘lubricated’) and adjectives ending in *-telný* (*rozpoznatelný* ‘recognizable’)...
- Numeral (C): lexemes expressing association with numerals, e.g. *dvojkový* ‘binary’...
- Other (0 – zero): *starý* ‘old’...

A nominal (short) adjectival form is not a SUB category value since SUB is a **global category**, specifying the **whole** paradigm of some lemma. The short form is assigned the same lemma as the long one, e.g. lemma(*stár*) = *starý* ‘old’.

3.2.1.3 Pronoun subcategorization

For pronouns, two subcategories are distinguished: SUB and Deixis (DEI). The usual subcategorization of pronouns does not reflect the dual classification view: e.g., *něčí* ‘someone’s’ is both possessive and indefinite. The second property is the value of the DEI category. The double subcategorization of pronouns was first used in the Ajka system in Brno (cf. [4], [8], [9]).

Since DEI is a category that is common to several parts, we discuss it as a special category (cf. 3.2.2).

The following SUB values are distinguished for pronouns:

- Personal (o): *já* ‘I’, *oni* ‘they’, *se* (reflexive)...
- Nominal (N): *kdo* ‘who’, *nikdo* ‘nobody’...
- Possessive (U): *můj* ‘my’, *čí* ‘whose’...
- Other, mainly delimitative (v): *každý* ‘every’, *týž* ‘same’...

3.2.1.4 Numeral subcategorization

Also for numerals, we distinguish two subcategories: SUB and DEI. The following SUB values are distinguished:

⁵ Compared to previous systems, we consider the forms of the passive participle to be short adjectives derived from verbs, rather than verbal forms. A passive participle is assigned a (non-disambiguated) multiple lemma whose components are (i) the long form of the corresponding adjective and (ii) the infinitive of the underlying verb. In corpora, a user can search for passive participles via both the adjectival or verbal component; the verbal component can be used, i. a., for solving diathesis problems in deep syntax (the connection between active forms of a verb and its passive participle forms is not lost, since they are represented by the verbal component).

- Cardinal (z): *pět* ‘five’ ... *kolik* ‘how many’...
- Ordinal (r): *pátý* ‘fifth’, *poprvé* ‘first’...
- Multiple (n): *dvakrát* ‘twice’...
- Fractional (h): *půl* ‘half’, *třetina* ‘third’, *čtvrt* ‘quarter’...
- Relative to the whole (u), including the following numerals:
 - aggregate: *dvě* ‘two’, *patero* ‘five’...
 - ensemble: *dvoje* ‘two’, *paterý* ‘five’...
 - group: *dvojice* ‘pair’, *pětice* ‘group of five’...
 - generic: *dvojí* ‘two kinds’, *paterý* ‘five kinds’...
- Number written in Arabic or Roman digits (=): 586...

3.2.1.5 Verb subcategorization

For verbs, only the following SUB values are distinguished:

- Auxiliary (b): only *být* ‘be’, *bývat* ‘used to be’
- Other (0 – zero): *navštívit* ‘visit’, *koupat* ‘bathe’...

The verbs *být* and *bývat* are considered auxiliary in case they participate in past tense constructions, pluperfect, present and past conditional mood constructions and future of imperfective verbs; and also conditional forms *by* ‘would’... The forms of the verb *být* a *bývat* used

(a) to coform periphrastic passive constructions: *Jsem/Bývám často překvapen*. ‘I am / I am (usually) often surprised.’

(b) as autosemantic words to express existence: *Bůh je*. ‘God is.’, and

(c) as a copula: *Dělník je doma / v lese / k dispozici / překvapený*. ‘The worker is at home / in the forest / available / surprised.’

are not considered auxiliary for annotation purposes, i.e. they are assigned the SUB=0 value.

3.2.1.6 Adverb subcategorization

For adverbs, we distinguish the following values for SUB:

- Pronominal (P), i.e.
 - local: *kudy* ‘where’, *tudy* ‘this way’, *odkud* ‘from where’, *nikam* ‘nowhere’...
 - temporal: *kdy* ‘when’, *kdykoli* ‘whenever’, *nikdy* ‘never’, *pokaždé* ‘each time’...
 - modal: *jak* ‘how’, *všelijak* ‘in various ways, anyhow’, *nijak* ‘in no way’...⁶
- Compound (s): *dopředu* ‘forward’, *namodro* ‘blue’...
- Numerical (C): *napůl* ‘half’, *vedví* ‘asunder’...

⁶ Local, temporal and modal adverbs are not distinguished in the SUB category since it is often very difficult to disambiguate between them.

- Regularly derived from verbal adjectives (V): *zamýšleně* ‘thoughtfully’...
- Other (0 – zero): *dobře* ‘well’

For adverbs, we also specify the Deixis category (DEI, cf. 3.2.2).

3.2.1.7 Conjunction subcategorization

For conjunctions, the following values for SUB are distinguished:

- Coordinate (^ – circumflex): *a* ‘and’, *ale* ‘but’...
- Subordinate (, – comma): *protože* ‘because’...
- Mathematical operations (* – asterisk): *krát* ‘times’...

Note. We interpret the word forms *abych* ‘so_that_I’, *abys* ‘so_that_you’..., *kdybych* ‘if_I’, *kdybys* ‘if_you’... as aggregates (cf. 3.1).

3.2.1.8 Particle subcategorization

For particles, the following SUB values are distinguished:

- Desiderative (p): *at’* ‘let’...
- Responsive (o): *ano* ‘yes’...
- Discursive marker (d): *bohužel* ‘unfortunately’...

3.2.1.9 Affix segment subcategorization

For affix segments, the following SUB values are distinguished (cf. 3.1):

- numeral prefix segment (C)
- postfix segment (p)
- other prefix segments (e)

3.2.2 DEI (deixis) category

In the DEI category, as the second POS subcategory, we distinguish the following values:

- Definite (U): personal pronouns, definite numerals: *dva* ‘two’
- Indefinite (N): *někdo* ‘someone’, *několik* ‘several’, *někdy* ‘sometime’...
- Negative (Z): *nikdo* ‘nobody’, *nijak* ‘in no way’...
- Interrogative (T): *kdo(ž)* ‘who’, *jaký* ‘which’, *kolik* ‘how many’, *kde* ‘where’...
- Relative (V): *jehož* ‘whose’...
- Reflexive (S): *se, si* ‘-self’...
- Demonstrative (D): *ten* ‘that’, *takový* ‘such’...; numerals *tolik* ‘so many’...; pronominal adverbs *tady* ‘here’...

The DEI category is primarily relevant to pronouns, numerals, pronominal adverbs, although not all values are used for all these parts of speech. For other parts of speech, this property is undefined (-). Every unambiguous pronoun, numeral, and pronominal adverb receives exactly one of these values.

The values of both SUB and DEI categories are combined, but not arbitrarily. Possible combinations of their values are shown in Table 1 (3.2.2.1), Table 2 (3.2.2.2) and Table 3 (3.2.2.3).

3.2.2.1 Pronouns

Table 1 lists combinations of possible values of SUB and DEI categories for pronouns.

PRONOUNS	definite	indefinite	negative	interrog.	relative	reflex.	demonstr.
Personal	<i>ty</i> 'you'	- ⁷	-	-	-	<i>si, se</i>	-
Nominal	-	<i>leccos</i> 'anything'	<i>nic</i> 'nothing'	<i>kdo</i> 'who'	<i>jenž</i> 'who'	-	-
Possessive	<i>jejich</i> 'their'	<i>něčí</i> 'someone's'	<i>ničí</i> 'nobody's'	<i>čí</i> 'whose'	<i>jejíž</i> 'whose'	<i>svůj</i>	-
delimitative, other	<i>každý</i> 'each'	<i>všelijaký</i> 'sundry'	<i>žádný</i> 'no'	<i>jaký</i> 'which'	-	-	<i>ten</i> 'this'

Tab. 1. Pronouns. Combination of SUB (column headings) and DEI (row headers) values and representative lexemes

Similarly as the Prague system, we do not distinguish ambiguous (interrogative / relative) pronouns: *jaký* 'which', *kteřý* 'which', *kdo* 'who'...

Some lexemes are relative only: *jenž* 'who', *jehož* 'whose'...

3.2.2.2 Numerals

Table 2 lists combinations of possible values of SUB and DEI categories for numerals.

NUMERALS	Definite	Indefinite	interrogative	Demonstrative
Cardinal	<i>sto</i> 'hundred'	<i>několik</i> 'several'	<i>kolik</i> 'how many'	<i>tolik</i> 'so many'
Ordinal	<i>pátý</i> 'fifth'	<i>několikátý</i> 'ord. numb. for several'	<i>kolikátý</i> 'ord. numb. for how many'	<i>tolikátý</i> 'ord. numb. for so many'
Multiple	<i>dvakrát</i> 'twice'	<i>několikrát</i> 'several times'	<i>kolikrát</i> 'how many times'	<i>tolikrát</i> 'so many times'
Fractional	<i>půl</i> 'half'	-	-	-
Relative to the whole	<i>dvoji</i> 'two kinds'	<i>několikero</i> 'several kinds'	<i>kolikery</i> 'how many kinds'	<i>tolikery</i> 'so many kinds'

Tab. 2. Numerals. Combination of SUB (column headings) and DEI (row headers) values and representative lexemes

⁷ "-." means *undefined*. Similarly in the other tables.

3.2.2.3 Adverbs

Table 3 lists combinations of possible values of SUB and DEI categories for adverbs.

PRONOMINAL ADVERBS	definite	indefinite	negat.	interrog.	relat.	demonstr.
	<i>všude</i> 'everywhere', <i>vždy</i> 'always'	<i>někde</i> 'somewhere', <i>poněkud</i> 'somewhat'	<i>nikdy</i> 'never'	<i>kdy</i> 'when'	<i>kdež</i> 'where'	<i>tady</i> 'here'

Tab. 3. Adverbs. Combination of SUB (column headings) and DEI (row headers) values and representative lexemes

4 CONCLUSION

We have presented the repertoire of categories and parts of speech and their subcategorization proposed in the NovaMorf project, which focuses on the innovation of the morphological description of Czech. We have shown two POS subcategorizations, SUB and DEI, exemplifying appropriate combinations of their values for pronouns, numerals and adverbs. Only marginally, we have dealt with a complex issue of lemmatization. We consider the outlined concept of the annotation system more systematic and consistent than existing Prague and Brno annotation systems; moreover, our system does not contain less information than they do. The proposed description of POS subcategorizations, together with a detailed description (not presented here) of all morphological categories, will be reflected in a new morphological dictionary, which will be used, i. a., for annotating new corpora of Czech (and possibly also for reannotating existing ones).

References

- [1] Hajič, J. (2000). Přehled morfologických značek. Available at: https://ucnk.ff.cuni.cz/doc/popis_znacek.pdf.
- [2] Hajič, J. (2004). Disambiguation of Rich Inflection (Computational Morphology of Czech). Praha, Karolinum.
- [3] Hlaváčová, J. (2009). Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Dissertation thesis. Praha: Univerzita Karlova. Available at: <http://utkl.ff.cuni.cz/phpBB3/viewtopic.php?f=11&t=1>.
- [4] Osolobě, K. (1996). Algoritmický popis české formální morfologie a strojový slovník češtiny. Dissertation thesis. Brno, Filosofická fakulta MU.
- [5] Osolobě, K. (2015). Korpusy jako zdroje dat pro úpravy nástrojů automatické morfologické analýzy (Slovotvorné varianty adjektiv na [(ou)[i]cí z hlediska morfologického značkování). Časopis pro moderní filologii, 97(2), pages 136–145.

- [6] Osolobě, K., and Žižková, H. (2016). Automatická morfologická analýza z hlediska pokrytí a nepokrytí morfologických variant. Available at: <http://ucnk.ff.cuni.cz/kl2016/abstract-detail.php?id=151>.
- [7] Osolobě, K., Hlaváčová, J., Petkevič, V., Svášek, M., and Šimandl, J. (2017). Nová automatická morfologická analýza češtiny. *Naše řeč* 100(4), pages 225–234.
- [8] Rychlý, P., Šmerk, P., Pala, K., and Sedláček, R. (2008). Morphological Analyzer Ajka. Available at: <https://nlp.fi.muni.cz/projects/ajka/>.
- [9] Sedláček, R. (2010). Morphematic analyser for Czech. Dissertation thesis. Brno, Fakulta informatiky MU.
- [10] Šimandl, J. (2015). Slovotvorný přehled slov s číselným významem I: číslovky určité. *KGA* 12, pages 54–74.
- [11] Šimandl, J. (2016). Slovotvorný přehled slov s číselným významem II: číslovky neurčité. *KGA* 13, pages 48–60.
- [12] Šmerk, P. (2010). K počítačové morfologické analýze češtiny. Dissertation thesis. Brno, Fakulta informatiky MU.

IMPROVING NOMINALIZED ADJECTIVES TAGGING

KLÁRA OSOLSOBĚ¹ – HANA ŽIŽKOVÁ¹

¹Faculty of Arts, Masaryk University, Brno, Czech Republic

OSOLSOBĚ, Klára – ŽIŽKOVÁ, Hana: Improving nominalized adjectives tagging. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 370 – 379.

Abstract: Part of speech transitions represent an interesting issue in terms of Automatic Morphological Analysis (AMA). In these cases, two parts of speech have to be considered: initial and final. However, their automatic recognition is complicated by the same form. This article presents the results of a corpus study aimed at mapping nominalized adjectives tagging with a focus on detecting candidates for nominalization among frequent adjectives. Analysis of the data obtained from the ČNK SYN v5 corpus shows different reasons for incorrect tagging. Taking into account these reasons, we propose three solutions for the improvement nominalized adjectives tagging.

Keywords: nominalized adjectives, automatic morphological analysis, disambiguation, corpus, tagging

1 INTRODUCTION

The division of vocabulary units into parts of speech is crucial for a systematic description of the language. Traditionally, in the classification of the part of speech, the synthesis of three criteria is based on formal, syntactic and semantic. However, in the case of natural language processing, we can only proceed from the form of the analyzed unit. This is because the automatic morphological analysis, which assigns units to part of speech, works mostly with the formal criterion of determining the part of speech. The syntactic and semantic criterion is sometimes used in disambiguation, but the rules are often difficult to formalize. In the case of part of speech transitions, the form is identical, and for this reason, the tagging is challenging.

There are three types of part of speech transitions [1]:

1. The initial and final part of speech is non-flexible. For example conjunctions → particle: *Prší, **ale** svítí slunce.* (conjunction) ‘It’s raining but the sun is shining.’ vs. *To **ale** prší!* (particle) ‘But it rains!’.

2. The initial part of speech is flexible, the final part of speech is non-flexible. For example noun → adverb: *Zadíval se na **modro** vod.* (noun) ‘He looked at the blue of waters.’ vs. *Obarvil látku na **modro**.* (adverb) ‘He dyed the fabric in blue.’.

3. Both the initial and the final part of speech are flexible. For example adjective → noun: *Petr je **nemocný**.* (adjective) ‘Peter is sick.’ vs. ***Nemocný** se uzdravil.* (noun) ‘The sick recovered.’.

This paper focuses on the third type of part of speech transition: nominalized adjectives.

Nominalized adjectives, sometimes also called syntactic nouns, have the same form and inflexion as adjectives, but syntactically they behave as a noun [2]. In this article, however, we do not distinguish nouns with adjectival inflexion (e.g. *mluvčí* ‘speaker’) and nominalized adjectives (e.g. *popravčí* ‘executioner’). We refer to all analyzed units as nominalized adjectives because both groups have the same adjectival inflexion and the same syntactic distribution of nouns.

2 APPROACH

We carried out a corpus study with the intention of mapping how the nominalized adjectives are tagged and which units can be included in the group of nominalised adjectives. We chose to use the largest available corpus at the time, SYN v5 ČNK (3,836 billion words)[3].

We proceeded in several steps. First, all possible endings of the nominalised adjectives were defined using *Slovník afixů užívaných v češtině* [4] and available Czech grammar books ([1], [2], [9], [10], [11]). CQL queries were then formulated to obtain the lists of nouns and adjectives with defined endings. These were compared, and the accuracy of the tagging was evaluated. The first 600 most frequent adjectives were checked for nominalized adjectives. Subsequently, we tried to find a key, how to classify analysed data so that the classification is relevant for automatic morphological analysis. The frequency of use, the context, and the occurrence in dictionaries were taken into account. Also, the assignment to a semantic group was taken into account.

After applying the listed steps, 319 nominalized adjectives were selected and subjected to a detailed analysis. The tagging of all selected units was observed in context. If the unit had been tagged incorrectly, we were curious about why this error occurred and whether it was possible to set a rule that could be used to tag part of speech correctly/properly. We have focused on the most frequent collocation of analysed units.

In the study we did not intentionally include zoological and botanical terms (*vrubozobí* – Anseriformes, *blanokřídli* – Hymenoptera etc.). We did not follow up the proper nouns, we only focused on the common nouns. Only the positive forms of adjectives were taken into consideration. Also a relatively large and open group of nouns type *Kladenští* ‘Kladno inhabitants’ was left aside [2].

3 FINDINGS

Analysis of the data shows that errors in the tagging of nominalized adjectives are due to two reasons in particular: inaccuracies in the morphological dictionary and erroneous disambiguation.

3.1 Inaccuracies in the morphological dictionary

There seem to be four types of inaccuracies in the morphological dictionary. There are some nouns (e.g. *šipkovaná*; ‘treasure hunt’) and adjectives (e.g. *hokejbalová* ‘hockeyball’, *jatečné* ‘slaughter’) which have been entered incorrectly as both POS = N and POS = A. Some adjectives (e.g. *basiliánský* ‘basilian’, *stehová* ‘stitched’) have been incorrectly entered as POS = N. We also found that many units which have only one interpretation, POS=A, are actually nominalized adjectives and can be used as a noun or an adjective (e.g. *vyučující* ‘teacher’, *popravčí* ‘executioner’) depending on the context. Similarly, other units only have the POS = N interpretation, but they can also be an adjective, POS=A (e.g. *košíková* ‘basketball’).

3.2 Erroneous disambiguation

Erroneous disambiguation leads to incorrect tagging as a noun instead of an agreeing postnominal or prenominal adjective. In the case of an agreeing postnominal adjective, such as *švihák lázeňský*; ‘spa dude’ (215 occurrences), we recorded 140 cases tagged incorrectly. Table 1 shows similar examples with incorrect disambiguation of other agreeing postnominal adjectives.

Problémem ale může být nedostatečné pojistné /pojistné/N krytí nebo nepřizpůsobitelnost parametrů pojištění (...) (SYN v5) ‘However, the problem may be insufficient insurance coverage or non-adaptability of insurance parameters (...)’
(...) stal se ze mě švihák lázeňský /lázeňský/N. (SYN v5) (...) ‘I became a spa dude.’
Dalším jídlem, které porotě předložily, bylo kuřecí /kuřecí/N prsičko se špenátovou fáší (...) (SYN v5) ‘Another meal presented to the jury was a chicken breast with spinach (...)’
Mám hovorné /hovorné/N prodavače rád. (SYN v5) ‘I like talkative salespeople.’
V listopadu jsem pozvána do poroty další taneční /taneční/N soutěže Miss Belly dance, už se moc těším. (SYN v5) ‘In November I was invited to the jury of another Miss Belly dance competition, I am looking forward to it.’

Tab. 1. Examples of erroneous disambiguation

We recorded erroneous disambiguation in cases where the unit precedes a proper noun:

Se závěrečným hvizdem rozhodčího /rozhodčí/A Samka tak vypukla na novopackém stadionu obrovská radost (...) (SYN v5) ‘With the final whistle of the referee Samko so broke out at the stadium in Nová Paka great joy (...)’
--

Sousedé se jednou sešli v hospodě U Švejka, hostinský /hostinský/A Petr Spittank rozdal noty a 14. ročník dětských radovánek byl na světě. (SYN v5) The neighbors once met at the U Svejka pub, the innkeeper Petr Spittank gave out notes and the <u>14th year of children's</u> fun was born.
Vzpomínky na natáčení má i jeho příbuzná /příbuzný/A Hana Ševčíková . (SYN v5) 'Also his relative Hana Ševčíková has memories of the <u>shooting</u> .'
(...) ve spolupráci s naší redakcí připravily výherní akci o půlroční předplatné /předplatné/A MF DNES . (SYN v5) '(...) in cooperation with our editorial team, they prepared the <u>winning event</u> for a six-month subscription to MF DNES.'
Že se děti nemůžou dočkat prázdnin konstatovala i její třídní /třídní/A Eva Oherová . (SYN v5) 'Even her class teacher Eva Oherová stated that children could not wait for the holidays.'

Tab. 2. Examples of units preceded by a proper noun

We also noticed the erroneous disambiguation if the unit was preceded by the lemmas *pán* 'mister' and *paní* 'missis':

Po hodině hledání ve skladu jim pan vedoucí /vedoucí/A přišel říci, že jejich pohovku nemohou najít (...) (SYN v5) 'After an hour of searching in the warehouse, (Mr.) supervisor came to tell them they couldn't find their sofa (...)'
Ani na to jim obezřetná paní domácí /domáci/A neskočila. (SYN v5) 'Even the prudent (Mrs.) landlady did not get to it.'
Po jeho odjezdu mně paní představená /představený/A citovala jeden z jejich rozhovorů. (SYN v5) 'After his departure, (Mrs.) Lady Superior quoted me one of their interviews.'
Pan vrátný /vrátný/A zakryl rukou sluchátko a řek mi, že to volá divadlo Šumperk. (SYN v5) '(Mr.) porter covered the handset with his hand and told me <u>it calls</u> the Šumperk theater.'
Tentokrát je nebohý pan účetní /účetní/A po smrti a stojí frontu před nebeskou bránou. (SYN v5) 'This time, the poor (Mr.) accountant is dead and faces the front of the heavenly gate.'

Tab. 3. Examples of units preceded by lemmas *pan* 'mister' and *paní* 'missis'

In rare cases, it seemed that the lemma and tag were incorrect:

Třídní /Třídeň/NNFS7-----A----- se zatvářila jako jeptiška, sepjala ruce a spustila (...) (SYN v5) 'The class teacher looked like a nun, clasped her hands and started (...)'

<p>Nemám ani tušení, jaká nemocenská/nemocenské/N by mne čekala v případě onemocnění. (SYN v5) 'I have no idea what kind of sickness benefit awaits me in case of illness.'</p>

Tab. 4. Examples of units with incorrect lemma and tag

4 SOLUTIONS

We propose three solutions for improving nominalized adjectives tagging: remove the inaccuracies from the morphological dictionary; add the obtained data described below to the Multiword Expressions Lexical Database (LEMUR) (see below); and apply our findings for disambiguation.

4.1 Removing the inaccuracies from the morphological dictionary

We believe that refinement of the data in the morphological dictionary [5] used for the ČNK corpora will lead to a more precise automatic morphological analysis. Below are proposals for adding analysed data to a morphological dictionary or for clarifying the interpretation of existing data.

We believe that refinement of the data in the morphological dictionary [5] used for the ČNK corpora will lead to a more precise automatic morphological analysis. Below are proposals for adding analysed data to a morphological dictionary or for clarifying the interpretation of existing data.

1) Only nouns, POS=N

The analysis showed that five units are nouns, even though they are listed in the morphological dictionary as both noun and adjective: *bytná* 'landlady', *bytný* 'landlord', *číhaná* 'lurking', *přisedící* 'associate', *šipkovaná* 'treasure hunt'.

2) Only adjectives, POS=A

The analysis showed that 35 units are adjectives (Appendix 1), even though they are listed in the morphological dictionary as both, noun and adjective.

We propose that units that represent school grades, *výborná* 'excellent', *výtečná* 'very good', *chvalitebná* 'good', *dobrá* 'satisfactory', *dostatečná* 'poor', *nedosta-tečná* 'failure' should be considered as adjectives. We think that from the contexts one can see the ellipsis of a noun. Within this semantic group, tagging will be unified and improved. We are aware of the problematic nature of this proposal. Ultimately, however, a pragmatic view of improved automatic tagging prevailed along with the most consistent tagging. By removing six units from a group of nouns, the consistency of the tagging within one semantic group will be preserved. The automatic part of speech tagging will be greatly improved, because it will not have to deal with the disambiguation, which is quite complicated especially in the case of frequent expressions as *výborná* 'excellent' and *dobrá* 'good'.

3) Nouns and adjectives POS=N, POS=A

The analysis showed that 50 units currently have only one interpretation (POS=N or POS=A). However, they can be both adjective and/or a noun (Appendix 2). In addition to the POS=N interpretation, it is also necessary to add a grammatical gender. [8]

4.2 Adding data to the Multiword Expressions Lexical Database

The analysis showed just how diverse the group of nominalized adjectives are. Although we tried to find different ways of characterization that could be generalized, it turned out to be almost impossible. The analysis confirmed that nominalized adjectives occur predominantly as one part of speech in certain contexts. Whatever this seems to be trivial, knowing the relevant collocations can greatly improve automatic morphological tagging.

The Multiword Expressions Lexical Database, LEMUR, ([6], [7]) was created by the Institute of Theoretical and Computational Linguistics, Charles University and the Institute of the Czech National Corpus FF UK, and is used in the disambiguation of corpora of the Czech National Corpus. Larger the database is, better result in tagging can be reached.

We will demonstrate our approach on lemmas, which can be both a noun or an adjective and belong to the semantic group of agentive nouns.

1) Units preceded by lemmas *pan* ‘mister’ and *paní* ‘missis’ and followed by a proper noun

- lemma *pan* ‘mister’

hostinský, kantýnský, lázeňský, nadřízený, obžalovaný, odsouzený, podřízený, představený, vrátný

innkeeper, canteenman, spamaster, superior, defendant, convicted, subordinate, superior, porter

- lemma *paní* ‘missis’

hostinská, kantýnská, lázeňská, nadřízená, obžalovaná, odsouzená, podřízená, představená, vrátná, zubatá

innkeeper, canteenlady, spamaster, superior, defendant, convicted, subordinate, superior, porter, Death

- lemmata *pan* i *paní*

The lemma *pan* or *paní* can help the gender disambiguation of units listed below because they have very often homonymous form.

cestující, domácí, dozorčí, duchovní, pokladní, produkční, provozní, radní, recepční, rozhodčí, spolubydlíci, třídní, účetní, vedoucí, vrchní, výčepní

passenger, landlord, landlady, supervisor, clergyman, cashier, production manager, operating, councilor, receptionist, referee, roommate, class teacher, accountant, leader, waiter, bartender, barmaid

2) Collocations

The collocation overview does not aim to list all collocations, but to list those that can help with automatic tagging.

We are aware of the fact that some of the collocations below, e.g. *rozhodčí smlouva*; ‘arbitration agreement’ may also occur in the opposite part of speech classification than we have stated: *Rada města schválila rozhodčí smlouvu*. (SYN v5) ‘The City Council approved the arbitration agreement.’ vs. *Fotbalový rozhodčí smlouvu nepodepsal*. (our own example) ‘The football referee did not sign the contract.’ However, we believe that similar contexts are really rare as we did not find examples in the corpus.

We believe that by applying relatively frequent phrases and collocations, the results of the automatic tagging will improve rather than deteriorate. Specifically, the collocation *rozhodčí smlouvu* ‘arbitration agreement / referee contract’ occurs in the SYN v5 corpus 108 times, all occurrences are erroneously marked as *rozhodčí / rozhodčí / N*; ‘referee’ *smlouvu / smlouva / N*; ‘agreement’. In this case, there would be a 100% improvement. If we take into account the lemmas *rozhodčí* and *smlouva*, it is found in SYN corpus v5 614 times, of which the *rozhodčí* is only 34 times correctly tagged as an adjective. Even in this case, there would be a significant improvement in labelling.

The specific improvement of the tagging will differ from the unit to unit. Complete list of collocations states Žižková [8].

Unit	Noun	Adjective
dozorčí ‘supervisor / supervisory’	operační dozorčí ‘operational supervisor’	dozorčí rada , dozorčí orgán , dozorčí služba , dozorčí komise , dozorčí útvár , dozorčí důstojník , dozorčí úřad ; ‘supervisory board, supervisory body, supervisory service, supervisory commission, supervisory unit, supervisory officer, supervisory authority’
lázeňský ‘spa’		švihák lázeňský, lázeňský dům ‘spa dude, spa house’
předsedající ‘chairman / presiding’	předsedající schůze , předsedající zasedání ; ‘chairman of the meeting, chairman of the session’	předsedající země , předsedající soudce , předsedající stát ‘presiding country, presiding judge, presiding state’
Představená ‘Lady Superior / presented’	matka představená, představená kláštera , představená řádu ; ‘Mother Superior, Superior of the Monastery, Superior of the Order’	
radní ‘councilor / town hall’	radní kraje , radní města ‘district councilor, city councilor’	radní věž ‘town hall tower’

recepční; 'receptionist / reception'	recepční hotelu , recepční kempu , recepční autokempu , recepční penzionu ; 'hotel receptionist, camp receptionist, campsite receptionist, guesthouse receptionist'	recepční služba , recepční pult , recepční estetika 'reception service, reception desk, reception aesthetics'
rozhodčí 'referee / arbitration'	hlavní rozhodčí, pomezí rozhodčí, čárový rozhodčí 'chief referee, sideline referee, line referee'	rozhodčí soud , rozhodčí nález , rozhodčí senát , rozhodčí výbor , rozhodčí tribunál , rozhodčí orgán , rozhodčí panel , rozhodčí řád , rozhodčí sbor , rozhodčí spis , rozhodčí výrok , rozhodčí institut , rozhodčí proces , rozhodčí soudce 'arbitration tribunal, arbitration report, arbitration senat, arbitration committee, arbitration tribunal, arbitration body, arbitration board, arbitration code, arbitration board, arbitration records, arbitration statement, arbitration institutes, arbitration process, arbitration judge'

Tab. 5. Examples of collocations

4.3 Taking into account when disambiguating

In order to neutralise gender differences, masculine in plural is more frequent in the semantic group of agentive nouns (*dozorčí* 'supervisor', *přisedící* 'associate', *lázeňský* 'spamaster' etc.) and designation of persons having a certain quality (*dospělý* 'adult', *trpící* 'suffering' etc.). Only the unit *pokojská* 'maid' is more frequent in plural in feminine. We propose to take into account the neutralization of gender differences in disambiguation of units listed in Appendix 3.

5 CONCLUSION

The results of this investigation show that there is a way how to improve the nominalized adjectives tagging.

Thanks to the selected CQL queries and subsequent manual searches, we compiled a list of 319 terms that we considered to be a possible nominalized adjective.

The detailed analysis of nominalized adjectives showed that the part of speech is not always tagged properly. There are two reasons for the erroneous tagging: inaccuracies in the morphological dictionary used in the ČNK corpora, and the disambiguation errors. So three solutions for improving nominalized adjectives tagging were proposed.

The first proposal involves removing inaccuracies from the morphological dictionary. We proposed a change of interpretation to noun for 5 units to POS=N, then for 37 units change to adjective, POS=A, for 51 units we recommended a change to noun, POS=N, and adjective, POS=A, interpretation. The second proposal foresees the extension of the LEMUR database to the proposed collocations collected for 147 units. Thirdly, findings on disambiguation were formulated for 89 units.

The analysis shows how diverse and hard it is to tag properly a group of expressions that are subject to the part of speech transition. Nevertheless, we believe that the proposed solutions will at least partially improve the automatic part of speech tagging.

References

- [1] Dokulil, M. et al. (1986). *Mluvnice češtiny 1. Fonetika. Fonologie. Morfonologie a morfe-mika. Tvoření slov.* Praha, Academia.
- [2] Štícha, F. (2013). *Akademická gramatika spisovné češtiny.* Praha, Academia.
- [3] Křen, M. et al. (2017). *Korpus SYN, version 5 as of 24 April 2017.* Praha, Ústav Českého národního korpusu FF UK. Accessible at <http://www.korpus.cz>.
- [4] Šimandl, J. (ed.). (2018). *Slovník afixů užívaných v češtině* [online]. Praha, Karolinum [cit. 2018-08-24]. Accessible at <https://www.slovnikafixu.cz>.
- [5] Hajič, J., and Hlaváčová, J. (2013). *MorFFlex Praha: LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*
- [6] Jelínek, T., M. Kopřivová, V. Petkevič, and Skoumalová, H. (2018). Variabilita českých fra-zémů v úzu. *Časopis pro moderní filologii, Karlova univerzita*, 100(2), pages 151–175.
- [7] Hnátková, M., T. Jelínek, M. Kopřivová, V. Petkevič, A. Rosen, H. Skoumalová, and Vondříčka, P. (2018). Lepší vrabec v hrsti nežli holub na střeše. *Víceslovné lexikální jednotky v češtině: typologie a slovník. Korpus – gramatika – axiologie, Univerzita Hradec Králové a ÚJČ AV ČR*, 17, pages 3–22.
- [8] Žižková, H. (2019). *Slovnědruhové přechody jako problém automatické morfologické ana-lýzy. Disertační práce. FF MU.*
- [9] Komárek, M. (1986). *Mluvnice češtiny 2. Tvarosloví.* Praha, Academia.
- [10] Karlík, P., Nekula, M., and Rusínová, Z. (1995). *Příruční mluvnice češtiny.* Praha, Lidové noviny.
- [11] Štícha, F. (2018). *Velká akademická gramatika spisovné češtiny.* Praha, Academia.

Appendix 1

barská, basiliánský, černá, divoká, dobrá, dostatečná, hokejbalová, chvalitebná, inst-rinsický, jatečné, kopulové, lutrový, maltézský, novellovský, oscilátorové, paname-rická, panský, pětimiliardová, pětimiliardový, podrostové, poloninský, safesová, sa-fesový, samodruhá, skopová, stehová, tajná, tvůrčí, umpirová, umpirový, verbovní, výborná, výtečná, vyvolený, zákolanská

‘bar, basilian, black, wild, good, sufficient, hockeyball, very good, intrinsic, slaughter, dome, low-wines, maltese, novel, oscillator, panamerican, manor, five

billion, undergrowth, polonin, safe, pregnant, mutton, stitched, secret, creative, umpire, recruiter, excellent, exquisite, chosen, from Zákolany’

Appendix 2

bioepřové, dančí, demonstrující, dojíždějící, dospívající, dostřelná, handicapovaný, hendikepovaný, košíková, kupující, místní, mrtvý, nakupující, nastávající, obviněný, oddávající, podezřelý, pohřešovaný, pokojská, pokojský, popravčí, postižený poškozený, prodávající, protestující, prvotrestaný, přednášející, předsedající, přespolní, přihlížející, příchozí, rezný, sázející, sloužící, soutěžící, stávkující, startující, studující, tonoucí, trpící, trvalá, účinkující, umírající, volající, vystupující, vyšetřující, vyučující, zavražděný, zraněný, zúčastněný

‘biopork, (of) fallow deer, demonstrating, commuter, teenage, firing, handicapped, basketball, buyer, local, dead, shopper, wife-to-be, accused, wedding registrar, suspect, missing, maid, chambermaid, executioner, handicapped, injured, seller, protesting, first punished, lecturer, presiding, cross-country, onlooker, incoming, rye, betting, serving, contestant, striking, starting, studying, drowning, suffering, permanent, acting, dying, calling, performer, investigating, teacher, murdered, injured, involved’

Appendix 3

cestující, demonstrující, dojíždějící, domácí, dospělá, dospívající, dozorčí, duchovní, handicapovaná, hendikepovaná, hostinská, kantýnská, kolemdoucí, kupující, lázeňská, místní, mrtvá, nadřizená, nakupující, nastávající, obviněná, obžalovaná, oddávající, odsouzená, okolojedoucí, pocestná, poddaná, podezřelá, podřízená, pohřešovaná, pokladní, postižená, postupující, poškozená, pracující, prodávající, produkční, protestující, protijdoucí, provozní, přednášející, předsedající, představená, přespolní, příbuzná, přihlížející, příchozí, přisedící, radní, recepční, rozhodčí, sázející, sloužící, služebná, soutěžící, spolubydlící, spoucestující, stávkující, strážná, studující, tonoucí, trpící, třídní, účetní, účinkující, umírající, vedoucí, věřící, volající, vrátná, vrchní, výčepní, vystupující, vyšetřující, vyučující, zavražděná, zraněná, zúčastněná

‘passenger, demonstrating, commuter, landlord, landlady, adult, teen, supervisor, clergyman, handicapped, handicapped, innkeeper, canteenlady, passerby, buyer, spamaster, local, dead, superior, shopper, wife-to-be, accused, indicted, wedding registrar, convicted, bystanders, wayfarer, subject, suspect, subordinate, missing, cashier, disabled, advancing, injured, working, seller, production manager, protesting, oncoming, operating, lecturer, chairman, superior, non-resident, relative, onlooker, incoming, associate, councilor, receptionist, referee, betting, serving, maid, contestant, roommate, fellow-traveller, striking, guard, studying, drowning, suffering, class teacher, accountant, performer, dying, leader, believer, calling, porter, waiter, bartender, performer, investigating, teacher, murdered, injured, involved’

MODIFICATIONS OF THE CZECH MORPHOLOGICAL DICTIONARY FOR CONSISTENT CORPUS ANNOTATION

JAROSLAVA HLAVÁČOVÁ – MARIE MIKULOVÁ –
BARBORA ŠTĚPÁNKOVÁ – JAN HAJIČ
Charles University, Prague, Czech Republic

HLAVÁČOVÁ, Jaroslava – MIKULOVÁ, Marie – ŠTĚPÁNKOVÁ, Barbora – HAJIČ, Jan: Modifications of the Czech morphological dictionary for consistent corpus annotation. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 380 – 389.

Abstract: We describe systematic changes that have been made to the Czech morphological dictionary related to annotating new data within the project of Prague Dependency Treebank (PDT). We bring new solutions to several complicated morphological features that occur in Czech texts. We introduced two new parts of speech, namely foreign word and segment. We adopted new principles for morphological analysis of global and inflectional variants, homonymous lemmas, abbreviations and aggregates. The changes were initiated by the need of consistency between the data and the dictionary and of the dictionary itself.

Keywords: morphological dictionary, Czech part of speech, corpus annotation, Golden rule of morphology

1 MOTIVATION

Despite recent advances in part of speech (POS) and morphological tagging using Deep Learning, the old truth that more data always gives better results ([1], [9]) still holds. At the same time, consistency in data annotation is a very important factor. For morphological annotation, especially for morphologically rich languages with thousands of possible combinations of morphological values, consistency can only be achieved when a dictionary lists all plausible morphological interpretations of all wordforms [3]. Naturally, such a dictionary must also be consistent with all the annotated data, which is an issue when legacy data are taken into account as annotated with previous – possibly not fully compatible – versions of the dictionary. Therefore, when extending the available set of manually annotated data for POS and morphological tagging, we have to follow the following principles:

(i) use different genre, register, style and/or domain to add diversity to the dataset;

(ii) develop the morphological dictionary in parallel with the annotation process, to ensure consistency among all the annotated data and also between the data and the dictionary.

To meet the requirement (i), we are manually extending the annotated data. We enlarge the morphological annotation of Czech written texts in the Prague Dependency Treebank 3.5 [6] by adding annotation of spoken data (from the Prague Dependency Treebank of Spoken Czech [10]), translation data (Czech part of the Prague Czech-English Dependency Treebank [4]) as well as a small amount of “user-generated” data from the internet translation services (corpus PDT-Faust¹). This will increase the amount of data available for NLP applications (such as MorphoDiTa [13] or DeriNet [15]) more than twice, genre-diversified (see Tab. 1).

It is important to pursue a manual morphological annotation of large data in parallel with the development of the dictionary (requirement (ii)). Therefore, while annotating, we are enriching the dictionary called MorfFlex [5], used in the original annotation, with words and wordforms stemming from new texts. Moreover, we are making systematic changes in capturing some phenomena in the dictionary. The long-time experience with the usage of the dictionary and the current annotation of real data has shown that several phenomena would be better to capture differently in order to achieve better consistency in the whole dictionary. The changes in the dictionary are being projected back into the data by repeated re-annotation to guarantee full consistency between the dictionary and the data.

Data type	written	spoken	translated	internet	Total
Tokens	1,725,242	742,257	1,162,072	33,772	3,663,343

Tab. 1. Morphological annotation in the future, consolidated edition of PDT

When formulating the principles of the dictionary and guidelines for annotation, as well as when making changes in the structure of lemmas and tags, it is necessary to find an optimal compromise between linguistic theory (often especially the traditional interpretation) and the needs of practical annotation, for which it is important to have simple and clear rules offering a solution for each token in any real text. We do not want to change the existing structure of MorfFlex, so we are capturing all the changes within the existing dictionary structure. Thus, at this time, we do not include the concept of multiple lemma nor extend the positional tag for marking variants as proposed in [7] and [8].

There are also other approaches to Czech morphology, most notably the NovaMorf project [12] and Universal Dependencies (UD) [11]. However, NovaMorf is still in its specification phase, while in MorfFlex we are bound by the already annotated corpus (PDT), and it is not yet clear if a conversion (both ways) can be lossless. In UD, the morphological features are adapted to the use in multilingual setting, and there is some loss if language-specific features are not used. On the other hand, there is an almost lossless conversion from MorfFlex-based annotation to the

¹ <https://ufal.mff.cuni.cz/grants/faust>

UD morphological features system, as described in [14]; future conversion to the UD system should thus be unproblematic.

In this paper, we describe changes that have been made to MorfFlex related to annotating new data within the project of the consolidated version of PDT.

2 GOLDEN RULE OF MORPHOLOGY

The MorfFlex dictionary lists more than 100,000,000 lemma-tag-wordform triples. For each wordform, full inflectional information is coded in a positional tag. Wordforms are organized into paradigms according to their formal morphological behavior. The paradigm (set of wordforms) is identified by a unique lemma. Apart from traditional morphological categories, the description also contains some derivational, semantic and stylistic information. The formal specification of the dictionary is in [2].

The so called “golden rule of morphology” (cf. [7], [8]) is applied to the dictionary. The rule says that any pair <lemma, morphological tag> is represented by at most one wordform.² The principle was, however, often violated in the previous version of the dictionary, mainly due to

- homonymy of lemmas;³
- different types of wordform variants.

Each of these problematic issues is addressed differently. The former one is solved by adding a numerical index to homonymous lemmas (see Sect. 3), the latter one by distinguishing two types of variants – global and inflectional ones (see Sect. 4). Until recently, both types of variants were marked uniformly at the 15th position of the tag. This did not allow to fully describe the complex variations that can occur for a single wordform.

3 LEMMA NUMBERING (INDEXING)

The problem of homonymy of lemmas is solved by giving numbers to the lemmas with the same spelling. We do not strive to make any distinction between meanings of homonymous words. The only differences we want to capture are those of formal morphological nature. Therefore, we add numbers only to lemmas that differ from the formal point of view. It means that we distinguish lemmas that have either

- different POS, e.g. *růst-1* as noun (‘a growth’) and *růst-2* as verb (‘to grow’), or
- different gender in case of nouns, e.g. *kredenc-1* as masculine and *kredenc-2* as feminine; they have the same meaning (‘a cupboard’), but different paradigms, or

² If the pair is meaningful, there is exactly one form, if it is not, there is none of them. There must not exist more than one wordform with the same lemma and tag.

³ The homonymy of wordforms has been resolved sufficiently in the previous versions of the dictionary.

- different aspect in case of verbs, e.g. *stát-1* with perfective aspect ('to happen') and *stát-2* with imperfective aspect ('to stand').

Thus, we have, e.g., lemma *jeřáb-1* for crane as a bird (animate masculine) and *jeřáb-2* for both a tree and crane as a device for lifting heavy objects (inanimate masculine). We do not distinguish the latter two meanings (tree vs. device), because they do not differ from the inflectional point of view. There might be a difference in derivation. In this case, the word *jeřábník* (a man who works with a crane-device) is derived from *jeřáb* as a device. It is not possible to derive *jeřábník* from *jeřáb* as a tree.

Due to a large number of complicated cases, we have decided not to take into account such derivational, stylistic and semantic differences. Thus we do not distinguish lemmas (if they inflect identically) that have:

- different meaning, e.g. *kohoutek* ('tap') and *kohoutek* ('flower');
- different derivational model: *matka* ('nut') and *matka* ('mother' with possessive adjective derivation);
- different style value: *ekonomka* ('female economist') and *ekonomka* ('school of economics', non-standard).

4 VARIANTS

Orthographic and stylistic variants of a word (hereinafter referred to as variants; e.g. archaic variant *these*, standard variant *teze*, and non-standard variant *téze* 'thesis') are the candidates for breaking the golden rule of morphology. We distinguish two types of the variants (see [7]):

- **Inflectional variants** are those variants that relate only to some wordforms of a paradigm defined by a special combination of morphological values, e.g. both *orli* and *orlové* ('eagles') are the wordforms of the noun *orel* ('eagle') and express plural masculine nominative.
- **Global variants** are those variants that relate to all wordforms of a paradigm, and always in the same way, e.g. *vyhýbat* and *vyhejbat* ('to avoid') – the whole paradigms of each verb differ in the distinction *-ý-* vs *-ej-* in the root.

There are two types of information that are used for the description of wordforms: lemma and tag. It is natural to express information about global variants within the lemma, because it is common for all its wordforms, and information about inflectional variants by means of a tag that applies only to specific wordforms.

4.1 Global variants

Global variants were not tackled uniformly in MorFFlex. Some global variants had different paradigms with different lemmas, others were grouped into one paradigm with one common lemma. In the former case there was no connection between the two variant lemmas. The latter case led to the most massive violations

of the golden rule because there were different wordforms with the same tags belonging to the same lemma.

Wordform	Lemma
<i>teze</i>	teze
<i>these</i>	these_a ^(^DD**teze)
<i>téze</i>	téze_h ^(^GC**teze)

Tab. 2. Global variants – example

We have decided to select one of the variants as “basic” and interconnect other variants via links to it. We use a notation that was originally designed for marking derivational relations. To distinguish variants from derivations, we introduce new codes for variants. We also simplify and reduce the set of style flags. We are now using only three types of global variants:

- DD – standard variant, including archaic ones,
- GC – non-standard (general Czech) variant, including dialectical, expressive, slang and vulgarisms,
- DS – distortion (a frequent typo, or otherwise distorted spelling).

Every variant, except for the basic one, has to be assigned a single indication of style. See examples in Tab. 2.

There are two main differences when compared to the previous treatment of variants; the global variants are really global – there cannot be a wordform belonging to the same lemma having different (or none) sign of style, and there is at most one indication of style for each paradigm.

4.2 Inflectional variants

For marking inflectional variants, we use the 15th position of the tag, as has been done before. The main difference lies in the fact that now we use this position strictly for inflectional variants. Another change is the simplification of the set of possible values. Numbers 1 to 4 mark standard variants, while numbers 5 to 9 relate to substandard ones. See examples in Tab. 3.

Wordform	Lemma	Positional Morph. Tag
<i>přijdeme</i>	přijít	VB-P---1P-AAP--
<i>přijdem</i>	přijít	VB-P---1P-AAP-6
<i>přídeme</i>	přijít	VB-P---1P-AAP-5
<i>přídem</i>	přijít	VB-P---1P-AAP-7
<i>přijdeme</i>	přijít	VB-P---1P-AAP-8
<i>přijdem</i>	přijít	VB-P---1P-AAP-9

Tab. 3. Inflectional variants – example

5 NEW FEATURES IN THE TAGSET

Czech texts contain not only “normal” words that fit well into traditional categories but also various sorts of strings (e.g. foreign words, abbreviations, etc.) that must be processed as well, and thus they need to be defined more precisely.

5.1 New part of speech: Foreign word

The POS of most foreign words were taken from their original languages. Thus, the wordform *in* was a preposition, *European* was an adjective, etc. However, in Czech texts, these words do not behave as their original POS might suggest. They are usually part of a longer foreign phrase, which may be a citation, a foreign name, etc. It seems inappropriate to assign usual morphological values to foreign wordforms within foreign phrases, since their role in Czech texts differs from their role in foreign texts. Therefore we have adopted a special POS concept of “foreign word” (presented for the first time in [8]).

Foreign word is such word that is not subject to Czech inflectional system and has no meaning of its own in Czech. Lemma of a foreign word is the same as the word itself. The tag contains special values at the POS and SUBPOS positions, namely F%. There are no other morphological values involved in the tag (see Tab. 4).

Foreign words should not be confused with indeclinable words that are of foreign origin, have already become part of the Czech vocabulary and have their meaning within the Czech language, e.g. the noun *kupé* (‘compartment in a train’) or an adjective *lila* (‘lilac colour’).

Wordform	Lemma	Tag
<i>European</i>	European	F%-----
<i>market</i>	market	F%-----

Tab. 4. Foreign word – examples

5.2 New part of speech: Segment

Segments are incomplete words. They are parts of words; in order to understand them, they must be joined with another string or word to create a complete word. As they are quite common in Czech texts and they were not previously captured consistently in the dictionary, we have created a new POS with the code S for them. According to their position in the complete word, we distinguish prefixal and suffixal segments.

Wordform	Lemma	Tag	Example
<i>česko</i>	česko	S2-----A----	<i>česko-ruská kniha</i> ‘Czech-Russian book’
<i>tří</i>	tří	S2-----A----	<i>tří až pětiletý</i> ‘three to five year old’
<i>nepoliticko</i>	politicko	S2-----N----	<i>nepoliticko-politické</i> ‘nonpolitical-political’

Tab. 5. Prefixal segment – examples

Wordform	Lemma	Tag	Example
<i>kou</i>	ka	SNFS7-----A----	<i>s manželem/kou</i> ‘with husband/wife’
<i>tice</i>	tice	SNFS1-----A----	<i>n-tice</i> ‘n-tuple’
<i>a</i>	a	SpQW----R-AA---	<i>řekl(a)</i> ‘he or she said’

Tab. 6. Suffixal segment – examples

Prefixal segments are strings that appear at the beginning of words. They are followed with a space or another separator, most often with a hyphen.

Lemma of prefixal segment is the string itself, unless it is in negative form. In that case, the positive form (without the prefix *ne-*) is considered to be the lemma. The tag of all prefixal segments has the code 2 at the 2nd position. Moreover, we specify for them also the 11th position concerning negation (see Tab. 5).

Suffixal segments are strings that may appear at the end of a wordform. They are usually attached directly to the word they combine with. The separator is most often a hyphen, parenthesis or a slash (/).

The suffixal segments express an affiliation to a specific POS. Thus, all the inflectional categories that describe the whole wordform, except for the first one (= the code for POS, which is S), are filled in the tag (with the exception of the aspect for verbs). The lemma is the closest “basic wordform” (see Tab. 6).

Wordform	Decomposed	Lemma	Tag
<i>zač</i>	<i>za co</i>	co	PQ--4-----z-
<i>začs</i>	<i>za co jsi</i>	co	PQ--4-----Z-
<i>doň</i>	<i>do něj</i>	on	P5ZS2--3-----d-
<i>dobřes</i>	<i>dobře jsi</i>	dobře	Dg-----lA--s-
<i>promluvil</i>	<i>promluvil jsi</i>	promluvit	VpYS----R-AAPs-
<i>kdyžs</i>	<i>když jsi</i>	když	J,-----s-

Tab. 7. Aggregate – examples

5.3 Aggregates

An aggregate is a wordform that is created by joining two or more wordforms (components of the aggregate) into one and cannot be simply assigned any POS. Aggregates are common especially in agglutinative languages, but there are two aggregate types in Czech, too:

- pronominal aggregates consisting of a preposition and the pronoun *on* (‘he’) or *co*, *copak* (‘what’);
- verbal aggregates consisting of a wordform of almost any POS with the string *s* added to the end. It stands for the wordform *jsi* (‘you are’).

The lemma of pronominal aggregates is the lemma of the pronoun. The lemma of a verbal aggregate is the lemma of its first component. The fact that a wordform is an aggregate is coded at the 14th position of the tag. The code of pronominal

aggregates corresponds to the initial letter of the preposition that forms their first component, verbal aggregates are coded with the letter *s* (see Tab. 7). Verbal and pronominal aggregates can combine; such aggregates are marked with the initial letter of the preposition, but in an uppercase letter (see the example *začs* in Tab. 7).

In the original MorfFlex, the pronominal aggregates were signaled by means of the second position in the tag. The lemma of pronominal aggregates was always the aggregate itself, the lemma of verbal aggregates with a verb at the beginning was the infinitive of the leading verb. Verbal aggregates composed of other POS (e.g. *kdyžs* ‘when you are’) were not treated as aggregates at all.

5.4 Abbreviations

An abbreviation that abbreviates a single word (e.g. *str* – *strana* ‘p – page’) is captured as a special wordform of the paradigm of that word. Only those categories that are valid for each use of the abbreviation are coded in the tag. The fact that it is an abbreviation is expressed at the 15th position by the letters *b* or *a* (see examples of the lemma *strana* in Tab. 8).

Wordform	Lemma	Tag	Example
<i>s</i>	<i>strana</i>	NNFXX-----A---a	<i>na s. 12</i> ‘at page 12’
<i>str</i>	<i>strana</i>	NNFXX-----A---b	<i>na str. 12</i> ‘at page 12’
<i>l</i>	<i>letopočet</i>	NNIS2-----A---b	<i>n. l.</i> ‘of AD’
<i>V</i>	V-88 ;B	NNXXX-----A----	<i>V. Havel</i>
<i>ČR</i>	ČR ;B ^ (Česká republika)	NNXXX-----A----	<i>ČR</i> ‘Czech Republic’

Tab. 8. Abbreviation – examples

Lemmas of other abbreviations, especially those that are composed of uppercase letters only (e.g. *USA*), is the abbreviation itself with a special flag B. They are assigned the tag of a maximally subspecified noun (with the value X for any gender, case, number at the positions 3 – 5). The same holds for one-letter abbreviations that stand for a single word but it is not clear for which of the many alternatives. This is, e.g., the case of initials of proper names (e.g. *V. Havel*, *V. Mrštík*). The abbreviations of this type have usually added the number 88 to their lemma as a human-readable indication of their status. There are some exceptions – very common abbreviations with only one meaning. Lemma of such abbreviations does not have the indexing number 88, as they cannot be mistaken for anything else. They have a semantic explanation as a note attached to the lemma (see Tab. 8).

6 CONCLUSION

We have described a project of manual morphological annotation on new text types within the new version of PDT. The need for consistency between the treebank(s) and within the dictionary has triggered deep and extensive changes in the

Czech morphological dictionary MorfFlex. The release of the new version of MorfFlex together with the new dataset is planned for the end of 2019. Thanks to the newly achieved higher consistency, we believe that the resulting larger, high-quality dataset and dictionary will contribute to better usability of the treebanks for linguistic inquiries, for new annotation projects using Czech, and also an increased accuracy of the NLP tools that learn from them.

ACKNOWLEDGMENTS

The research has been supported by the Czech Science Foundation under the project GA17-12624S. The research has also been supported by the LINDAT/CLARIN and LINDAT/CLARIAH-CZ projects of Ministry of Education, Youth and Sports of the Czech Republic (LM2015071 and LM2018101).

REFERENCES

- [1] Banko, M., and Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. In Proceedings of the 39th annual meeting on ACL. Association for Computational Linguistics, pages 26–33.
- [2] Hajič, J. (2004). Disambiguation of Rich Inflection. (Computational Morphology of Czech). Karolinum, Prague.
- [3] Hajič, J. (2000). Morphological Tagging: Data vs. Dictionaries. In Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference, Seattle, pages 94–101.
- [4] Hajič J., Hajičová E., Panevová J., Sgall P., Bojar O., Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký J., Šindlerová J., Štěpánek J., Toman J., Urešová Z., and Žabokrtský Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In Proceedings of the 8th International Conference on LREC 2012, European Language Resources Association, Istanbul, pages 3153–3160.
- [5] Hajič, J., and Hlaváčová, J. (2013). MorfFlex CZ. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University. Accessible at: <http://hdl.handle.net/11858/00-097C-0000-0015-A780-9>.
- [6] Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, D. J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š., and Žabokrtský, Z. 2018. Prague Dependency Treebank 3.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University. Accessible at <http://hdl.handle.net/11234/1-2621>
- [7] Hlaváčová, J. (2017). Golden Rule of Morphology and Variants of Wordforms. *Jazykovedný časopis / Journal of Linguistics*, 68(2), pages 136–144.

- [8] Hlaváčová, J. (2009). Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Disertační práce. Univerzita Karlova.
- [9] Church, K., and Mercer, R. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1), pages 1–24.
- [10] Mikulová M., Mirovský J., Nedoluzhko A., Pajas P., Štěpánek J., and Hajič J. (2017). PDTSC 2.0 – Spoken Corpus with Rich Multi-layer Structural Annotation. In *Lecture Notes in Computer Science*, No. 20th International Conference TSD 2017, Prague, pages 129–137. Cham, Switzerland: Springer International Publishing.
- [11] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on LREC 2016*, pages 1659–1666. Paris.
- [12] Petkevič, V., Hlaváčová, J., Osolsobě, K., Šimandl, J., and Svášek, M. (2019). Microsyntactic Parts of Speech in NovaMorf, a New Morphological Annotation of Czech. In *Proceedings of SLOVKO 2019* (this volume).
- [13] Straková J., Straka M., and Hajič J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations*, Association for Computational Linguistics, pages 13–18. Baltimore.
- [14] Zeman, D. (2018). *The World of Tokens, Tags and Trees*. Studies in Computational and Theoretical Linguistics, Charles University, Prague.
- [15] Žabokrtský Z., Ševčíková M., Straka M., Vidra J., and Limburská A. (2016). Merging Data Resources for Inflectional and Derivational Morphology in Czech. In *Proceedings of the 10th International Conference on LREC 2016*, pages 1307–1314, Paris, European Language Resources Association.

LEVELS OF ANNOTATION IN THE SLOVENE TRAINING CORPUS ssj500k 2.2

MIJA BON – POLONA GANTAR
Faculty of Arts, University of Ljubljana, Slovenia

BON, Mija – GANTAR, Polona: Levels of annotation in the Slovene training corpus ssj 500k 2.2. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 390 – 399.

Abstract: This paper presents the Slovene Training Corpus ssj500k 2.2, which has been annotated on the levels of tokenization, sentence segmentation, part-of-speech tagging, lemmatization, syntactic dependencies, named entities, verbal multi-word expressions, and semantic role labeling. It describes the individual layers of annotation and shows the scope of using the training corpus in the production of various lexicons, such as the lexicon of multi-word units and the valency lexicon of modern Slovene. It concludes by presenting our future work, i.e. the annotation of multi-word expressions based on the Slovene Lexical Database.

Keywords: corpus linguistics, training corpus, corpus annotation, Slovene language

1 INTRODUCTION

A training corpus is a linguistic source, which is generally manually annotated or corrected and used mainly for training statistical models for different purposes, such as part-of-speech tagging or parsing [1]. Training data can be used in supervised machine learning systems, which enable efficient automatic annotation of even very large corpora. The latest version of the Slovene Training Corpus 2.2 [17] consists of two training corpora, the whole of jos100k corpus V2.0 [7] and 400,000 words from training corpus jos1M 1.2. [8]. The training corpus ssj500k 2.2 [17] is freely available at Clarin.si repository under the Creative Commons (CC) license, Attribution-NonCommercial 3.0.¹ Compared to the previous, 2.1 version [16], this version corrects various errata in spacing and text metadata and, in cases where it was possible to do so automatically, adds UD morphological and dependency annotations to the corpus [17].

2 CORPUS DESCRIPTION

The ssj500k 2.2 training corpus contains 586,248 tokens, 27,829 sentences, and 500,295 words/lemmas, manually annotated on six levels: the whole corpus is

¹ <http://eng.slovenscina.eu/tehnologije/ucni-korpus>,
<https://www.clarin.si/repository/xmlui/handle/11356/1210>.

lemmatized and morphosyntactically (POS) annotated, about half of the corpus is annotated with syntactic dependencies and verbal multi-word expressions, a third of it is annotated with named entities, and approximately a quarter of it with semantic role labels. The whole of the corpus is thus morphosyntactically annotated, with specific parts of the corpus also containing other layers of annotation. Namely, the part of corpus labeled with SRL also includes syntactic annotation, MWEs, and named entities. This is particularly useful in linguistic analysis, where different levels of annotation can be combined to form a more comprehensive overview of a particular linguistic phenomenon.

<i>Level of annotation</i>	<i>Annotated sentences</i>
<i>Part-of-speech</i>	27,829
<i>Lemmatization</i>	27,829
<i>Verbal multi-word expressions</i>	13,511
<i>Syntactic dependencies</i>	11,411
<i>Named entities</i>	9,478
<i>Semantic role labeling</i>	5,491

Tab. 1. Number of annotated sentences on each level of ssj500k 2.2

2.1 Sentence segmentation

The ssj500k 2.2 is segmented into 27,829 sentences with 586,248 tokens. A statistical overview is given in Table 2. The data had already been annotated on the levels of segmentation and tokenization in preliminary corpora. The ssj500k 2.2 corpus was further manually validated and corrected.

<i>Element</i>	<i>n</i>
<i>Text</i>	1,677
<i>Paragraph</i>	8,137
<i>Sentence</i>	27,829
<i>Token</i>	586,248

Tab. 2. Statistical data on elements of ssj500k 2.2²

2.2 Part-of-speech tagging and lemmatization

The entirety of the ssj500k 2.2 was lemmatized and tagged on morphosyntactic (POS) and syntactic levels; it consists of 500,295 words. Part-of-speech tagging was done by using the tagset JOS system [6], which includes 12 POS categories with 1,903 possible attributes [14]. The part-of-speech tagging and lemmatization of

² <http://eng.slovenscina.eu/tehnologije/ucni-korpus>

preliminary corpora was performed by using the Obeliks tool, described in [14], with 91.34% accuracy for all tags and 98.30% for POS only. The lemmatizer had an approximately 98% accuracy [14, p. 4]. The whole corpus was manually corrected.³

Because of an increasing interest in the field of NLP in creating consistent annotation schemes that would enable the comparison of annotated data of individual languages, the Slovene JOS annotation scheme was adjusted to conform to the Universal Dependencies framework. The Universal Dependencies v2 standard includes 17 POS categories, and the Slovene corpus uses 16 of those UPOS tags suitable for Slovene: ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROP, PUNCT, SCONJ, VERB, X. The conversion of the *ssj500k* into a UD treebank was initially envisioned as an automated process. However, due to numerous differences between the two systems of annotation, especially on the level of syntactic description, a complex system of conversion rules was additionally created.⁴

2.3 Syntactic dependencies

The JOS dependency treebank model, used for surface-syntactic dependency annotation, was designed within the framework of the project Linguistic Annotation of Slovene [19]. The model which is based on syntactic dependencies but which also takes into account the syntactic characteristics of the Slovene language, consists of a robust three-level system. The labels are described in Table 3. A tag is attributed to each token and punctuation. The highest level in the system is occupied by the so-called meta element, which demonstrates either the interrelation of syntactically highest elements in sentences or syntactically less predictable structures, e.g. an ellipsis. [19, p. 51–52].

Groups of labels	Labels	Description
First level labels link elements in different types of phrases.	<i>dol</i>	Links heads and modifiers in phrases.
	<i>del</i>	Links parts of verbal phrases.
	<i>prir</i>	Links heads in coordinate structures within clauses.
	<i>vez</i>	Links words or commas in conjunctive function.
	<i>skup</i>	Links (function) words in frozen multi-word structures.
Second level labels link sentence elements.	<i>ena</i>	Clause subject.
	<i>dve</i>	Clause object.
	<i>tri</i>	Adverbial of manner.
	<i>štiri</i>	Other adverbials.

³ During the making of the last version of reference corpus Gigafida 2.0. (<https://viri.cjvt.si/gigafida/>), a new meta-tagger was created, with the accuracy of 94.34% for MSD and 98.66% for lemmatization.

⁴ For more about the process see [3].

Third level label links <u>all other</u> <u>structures</u> .	<i>modra</i>	Links to root, punctuation, syntactically less predictable structures, parentheses etc.
---	--------------	---

Tab. 3. Labels in the JOS dependency model (<http://eng.slovenscina.eu/tehnologije/razclenjevalnik>)

At the same time, the Sentence Markup (SMU) tool [2] was additionally developed for manual annotation, visualization, and data search. Surface-syntactic dependency annotation was performed in 11,411 sentences, with approximately half of them re-annotated following the evaluation of the annotation system, POS errors, and quality analysis of the annotation. At least two annotators manually annotated syntactic dependencies. All cases of discrepancy were further examined by a third annotator.

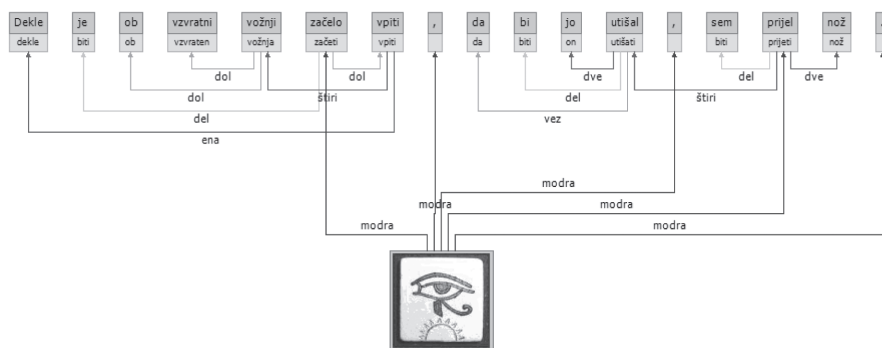


Fig. 1. Syntactic level in the SMU annotation tool

2.4 Named entities

Approximately a third of *ssj500k 2.2* (9,478 sentences) was manually annotated with named entity annotations in the WebAnno tool, with the aim of developing a named entity extractor for the Slovene language based on machine learning [20]. The annotation distinguished five types of NE: Person (*per*), Person Derivative (*deriv-per*), Location (*loc*), Organization (*org*), and Miscellaneous (*misc*). Apart from standard categories for named entities – i.e. names for people, pets, and groups of people; locations, including named buildings; organizations and institutions; and other proper nouns for things, for example book titles etc. – the category *deriv-per* was introduced, which annotates personal possessive adjectives, in order to improve anonymization of personal data [21].

7,015 named entities were marked in 9,478 sentences; this amounts to 1.35 named entities per sentence on average. The distribution of named entities by categories is given in Table 4.

Named entity	<i>n</i>	%
Loc	1,968	28%
Org	1,338	19%
Per	2,927	41.5%
Deriv-per	180	2.5%
Misc	602	9%
Total	7,015	100%

Tab. 4. Statistical data on named entities in the annotated part of *ssj500k 2.2*

2.5 Verbal multi-word expressions

The annotation scheme for verbal multi-word expressions (VMWEs) was based on categories developed within the international PARSEME COST Action Shared Task 1.1, adapted to the Slovene language [9, 10]. VMWE annotation includes the following four categories:

- inherently reflexive verbs (IRV),
- light-verb constructions, divided into full (LVC.full) and cause (LVC.cause),
- inherently adpositional verbs (IAV),
- verbal idioms (VID).

13,511 sentences were manually annotated following the Guidelines developed within the PARSEME shared task 1.1. In the first phase, 11,411 sentences were annotated by two annotators in accordance with the first version of the Guidelines. Discrepancies in annotations were discussed and adjusted accordingly. During the second phase, categories were automatically modified to comply with the second version of the Guidelines and manually checked. Additionally, 2,100 sentences were manually annotated by individual annotators in accordance with the modified Guidelines [12]. The first phase of annotation was performed in the SMU tool adjusted for VMWE labeling; the second phase employed the FLAT annotation platform, which enables labeling strings of text using previously defined categories ([10, p. 86], [12]).

The 13,511 annotated sentences contain 3,364 VMWEs in all forms (as they appear in sentences), with slightly fewer than 1,100 different expressions. 2,920 sentences (approximately 22%) contain at least one VMWE. Overall, the distribution of VMWEs in the annotated part of the *ssj500k 2.2* is 0.25 VMWE per sentence; in other words, there is one VMWE present, on average, in every fourth sentence ([10, p. 86-87], [12]).

<i>VMWE category</i>	<i>n</i>	<i>%</i>
<i>IRV</i>	1,627	48%
<i>IAV</i>	710	21%
<i>VID</i>	724	22%
<i>LVC-cause</i>	64	2%
<i>LVC.full</i>	239	7%
<i>together</i>	3,364	100%

Tab. 5. Distribution of VMWEs in ssj500k 2.2

2.6 Semantic Role Labelling

Semantic Role Labeling (SRL) refers to the process of detecting and assigning semantic roles to semantic arguments determined by the predicate or verb of a sentence. The framework for semantic role labeling was developed within the bilateral project Semantic Role Labeling in Slovene and Croatian ([15], [11]). It follows the path of previous SRL efforts (PDT, Vallex, FrameNet, Propbank etc.) while also considering the specifics of both target languages. The SRL tagset, based on the Prague Dependency Treebank, consists of 25 semantic labels: 5 arguments, 17 adjuncts, and 3 labels for multi-word predicates [11, p. 93–94], as seen in Table 6 and described in detail in [15].

<i>agent</i>	<i>ACT</i>
<i>patient</i>	PAT
<i>recipient</i>	REC
<i>origin</i>	ORIG
<i>result</i>	RESLT
<i>location</i>	LOC
<i>source (location)</i>	SOURCE
<i>goal (location)</i>	GOAL
<i>event</i>	EVENT
<i>time</i>	TIME
<i>duration</i>	DUR
<i>frequency</i>	FREQ
<i>aim</i>	AIM
<i>cause</i>	CAUSE
<i>contradiction</i>	CONTR
<i>condition</i>	COND
<i>regard</i>	REG

<i>accompaniment</i>	ACMP
<i>restriction</i>	RESTR
<i>manner</i>	MANN
<i>means</i>	MEANS
<i>quantification</i>	QUANT
<i>multi-word predicate</i>	MWPRED
<i>modal</i>	MODAL
<i>phraseological unit</i>	PHRAS

Tab. 6. SRL tagset for Slovene

A total of 5,491 sentences were annotated with semantic roles. The first 500 manually annotated sentences were used for automatic labeling, using mate-tools semantic role labeler with the German feature set [11, p. 94]. The second phase included automatic annotation of the remaining 4,991 sentences and their manual verification by five annotators [11, p. 94–95]. The annotation was performed in the SMU tool.

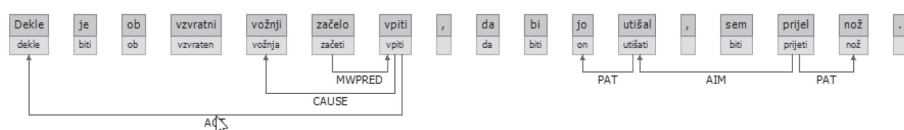


Fig. 2. SRL layer of annotation in the SMU tool

All 25 semantic labels were found in the corpus; predictably, however, the most frequent ones were argument roles of PAT and ACT, the former with a significantly higher frequency than the latter, and RESULT. These were followed by adjunct roles of TIME, MANN, and LOC [11, p. 96].

Slovene was found to have relatively stable patterns for its most frequent verbs, such as *biti* ‘to be’, *imeti* ‘to have’, *dobiti* ‘to get’, *morati* ‘must’, *moči* ‘can’, *hoteti* ‘will’, *želeli* ‘want’, *reči* ‘to say’, *povedati* ‘to tell’, e. g. [11, p. 95–97]:

‘to have’ *imeti*

- WHO (ACT) has WHAT (PAT) [for WHOM (REC), from whom (ORIG), where (LOC), when (TIME) ...]

‘must’ *morati*

- WHO (ACT) must INF (MODAL)

‘to go’ *iti*

- WHO (ACT) goes WHERE (GOAL) [how (MANN), when (TIME), under what conditions (COND) ...]
- to go (PHRAS)
- to go SUPINE (MWPRED).

3 CONCLUSION AND FUTURE WORK

The Slovene Training Corpus *ssj500k 2.2* was primarily intended for machine learning and linguistic analysis. So far, the training corpus has been used for automatic tagging of the Slovene reference corpus *Gigafida 2.0*, creating a lexicon of MWEs, machine learning for automatic annotation of corpora *Gigafida* and *Kres* on the level of MWEs and SRL, as well as the analysis of sentence patterns and building of valency lexicons.

In the future we plan to continue working on new layers of annotations based on the Slovene Lexical Database, firstly with a newly developed typology of MWEs. The new typology enables us to also identify non-verbal MWEs, such as noun MWEs (*rdeče številke* lit. red numbers, ‘deficit’, *kaplja v morje* lit. a drop in the see, ‘negligible amount’, fixed prepositional phrases (*med drugim* ‘among others’, *v skladu s/z* ‘in accordance with), and multi-word discourse markers (*v tem primeru* ‘in this case’). Our aim is to recognize and define MWEs as part of language with individual meaning and/or syntactic function. MWEs are categorized into three types: phraseological units – PU (*kaplja čez rob* ‘the last straw’, *leta tečejo* ‘years go by’), *za vraga* ‘heck’), fixed expressions – FE (*varnostni trikotnik* ‘warning triangle’), *črna luknja* ‘black hole’, *d. o. o.* (*družba z omejeno odgovornostjo* ‘limited company’), and syntactic combinations – SC (*na prostem* ‘in the open’, *za zdaj* ‘for now’, *v skladu s/z* ‘in accordance with’, *in tako naprej* ‘and so on’, *po eni strani – po drugi strani* ‘on the one hand – on the other hand’). Collocations and extended collocations, which can also be seen as MWEs, will be extracted from the corpus via Sketch Engine and other tools developed within the project *New Grammar of Contemporary Standard Slovene*. Currently, the first phase of the manual annotation of MWEs is in progress. Following this, IAA will be analyzed to prove or disprove the consistency of categories. In the final stage, a lexicon of MWEs will be completed and made part of the Dictionary of Modern Slovene [13].

ACKNOWLEDGMENTS

Annotation levels were defined within the framework of the national project *Nova slovnica sodobne standardne slovenščine: viri in metode* (New grammar of contemporary standard Slovene: sources and methods, ARRS J6-8256); corpus annotation and analyses were completed as part of the *ARRS P6-0215* research program (Slovene language – basic, contrastive, and applied studies). Corpus development has been a part of the infrastructural program of the Center for Language Resources and Technologies at the University of Ljubljana.

References

- [1] Arhar, Š. (2009). Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo*, 54, pages 3–4.
- [2] Dobrovoljc, K., Krek, S., and Rupnik, J. (2012). Skladijski razčlenjevalnik za slovenščino.
- [3] Dobrovoljc, K., Erjavec, T., and Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. *Konferenca Jezikovne tehnologije in digitalna humanistika*, pages 190–192. Ljubljana.
- [4] Dobrovoljc, K., Erjavec, T., and Krek, S. (2017). The Universal Dependencies Treebank for Slovenian.
- [5] Dobrovoljc, K., Erjavec, T., and Krek, S. UD Slovenian SSJ. Accessible at: https://universaldependencies.org/treebanks/sl_ssj/index.html.
- [6] Erjavec, T., Fišer, D., Krek, S., and Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1806–1809. Paris, ELRA.
- [7] Erjavec, T., Krek, S., and Fišer, D. (2010). jos100k corpus V2.0. Accessible at: <http://hdl.handle.net/11356/1213>.
- [8] Erjavec, T., Krek, S., and Dobrovoljc, K. (2019). Training corpus jos1M 1.2, Slovenian language resource repository CLARIN.SI. Accessible at: <http://hdl.handle.net/11356/1213>.
- [9] Gantar, P., Krek, S., and Kuzman, T. (2017). Verbal multiword expressions in Slovene. *Europhras 2017*, pages 247–259. Springer.
- [10] Gantar, P., Arhar Holdt, Š., Čibej, J., Kuzman, T., and Kavčič, T. (2018). Glagolske večbesedne enote v učnem korpusu ssj500k 2.1. In *Proceedings of the conference on Language Technologies & Digital Humanities*, pages 85–92.
- [11] Gantar, P., Štrkalj Despot, K., Krek, S., and Ljubešič, N. (2018). Towards Semantic Role Labeling in Slovene and Croatian. In *Proceedings of the conference on Language Technologies & Digital Humanities*, pages 92–98.
- [12] Gantar, P., Arhar Holdt, Š., and Čibej, J. (in print). Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene. *Contributions to Contemporary History*.
- [13] Gorjanc, V., Gantar, P., Kosem, I., and Krek, S. (2017). *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana, Založba FF.
- [14] Grčar, M., Krek, S., and Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana, Institut Jožef Stefan.
- [15] Krek, S., Gantar, P., Dobrovoljc, K., and Škrjanec, I. (2016). Označevanje udeleženskih vlog v učnem korpusu za slovenščino. In *Proceedings of the Conference on Language Technologies & Digital Humanities*, pages 106–110. Faculty of Arts, University of Ljubljana.
- [16] Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., and Zajc, A. (2018). Training corpus ssj500k 2.1, Slovenian language resource repository CLARIN.SI. Accessible at: <http://hdl.handle.net/11356/1181>.
- [17] Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., and Zajc, A. (2019). Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI. Accessible at: <http://hdl.handle.net/11356/1210>.

- [18] Ledinek, N., and Erjavec, T. (2009). Odvisnostno površinskoskladenjsko označevanje slovenščine: specifikacije in označeni korpusi. *Simpozij Obdobja* 28, pages 219–224.
- [19] Ledinek, N. (2014). *Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine*. Ljubljana, Založba ZRC, ZRC SAZU.
- [20] Štajner, T., Erjavec, T., and Krek, S. (2013). Razpoznavanje imenskih entitet v slovenskem besedilu. *Slovenščina 2.0*, 2, pages 58–82. Accessible at: http://slovenscina2.0.trojina.si/arhiv/2013/2/Slo2.0_2013_2_04.pdf.
- [21] Zupan, K., Ljubešič, N., and Erjavec, T. (2017). Annotation guidelines for Slovenian named entities Janes-NER.

MEANING AND SEMANTIC ROLES IN CzEngClass LEXICON

ZDEŇKA UREŠOVÁ – EVA FUČÍKOVÁ – EVA HAJIČOVÁ – JAN HAJIČ

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Czech Republic

UREŠOVÁ, Zdeňka – FUČÍKOVÁ, Eva – HAJIČOVÁ, Eva – HAJIČ, Jan: Meaning and semantic roles in CzEngClass lexicon. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 403 – 411.

Abstract: This paper focuses on Semantic Roles, an important component of studies in lexical semantics, as they are captured as part of a bilingual (Czech-English) synonym lexicon called CzEngClass. This lexicon builds upon the existing valency lexicons included within the framework of the annotation of the various Prague Dependency Treebanks. The present analysis of Semantic Roles is being approached from the Functional Generative Description point of view and supported by the textual evidence taken specifically from the Prague Czech-English Dependency Treebank.

Keywords: semantic roles, valency, parallel corpus, lexical semantics, lexical resource

1 INTRODUCTION

Since the Functional Generative Description (FGD) [10] has never systematically explored lexical semantics, it is not surprising that no description of lexical synonymy and semantic roles can be found in the pioneering works of this theory, neither there is a systematic theoretical description in the FGD follow-up works. However, some experiments with enhancing the valency lexicon of Czech verbs, starting with VALLEX 2.5 ([2], [18]), with semantic roles for the verbs of communication and the verbs of exchange ([3], [4]), building mainly on [5], where a lexicographic representation of lexical-semantic conversions is presented. The so far last version of VALLEX¹ [6] is divided into data and rule component, in order to present the representation of grammaticalized and lexicalized alternations [7]. Instead of semantic roles the term “situational participants” is used. We consider the approach of Kettnerová and Lopatková ([5], [8]) the fundamental starting point for our research and in various aspects we build upon it.

¹ <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2307> and <http://ufal.mff.cuni.cz/vallex/3.5/>

2 FGD APPROACH TO MEANING

The FGD's understanding of meaning is consistent with the concept of meaning in European structural linguistics as firstly formulated in de Saussure's works and his followers, specifically in the works of Prague scholars. As stated in [10], FGD considers the linguistic meaning distinct from (cognitive, ontological) content (or factual knowledge). FGD distinguishes two types of asymmetry: first "same content – different meaning" and second, "different content – same meaning" and makes it clear that for some correct interpretations (without ambiguity) of a sentence it is not enough to reach into the layer of linguistic meaning, but one needs to go into "the layer of cognitive content".

Apart from the distinction of linguistic meaning and ontological content, FGD took over and substantially precised the concept of language layers. Units of the "lower" layer serve as a form, while the units on the "higher" layer aligned with these forms serve as their functions. This stratificational approach, while considering several layers, emphasizes the (deep syntactic) "tectogrammatical" (TG) layer as the main one, representing linguistically structured meaning.

As implied from the above, it is not unexpected that the deep („underlying") syntactic constructions of which the TG layer consists of do not represent cognitive content. As noted in [16, p. 326], "for many semanticists this (= TG) layer would not belong to the domain of meaning at all".

3 FGD APPROACH TO LEXICAL SEMANTICS

No special attention was paid to the issues related to lexical semantics within the FGD until the development of FGD-related valency lexicons started (PDT-VALLEX [17], VALLEX [2], EngVallex [27] and CzEngVallex [26]. The approach to valency in all these lexicons is based on the valency theory developed within the FGD ([19], [20]). These lexicons mostly focus on verbs.

This FGD Valency Theory recognizes five "actants" of predicates (called also "inner participants," or by other theories called "arguments"): ACT (Actor), PAT (Patient), ADDR (Addressee), EFF (Effect) and ORIG (Origin). In addition, FGD distinguishes free modifiers (by other theories called "adjuncts") that capture circumstantial relations, such as manner-type, temporal, spatial, causal, etc. Valency characteristics of predicates are captured in their valency frames. Each valency frame consists of "(valency) slots" [9] corresponding to predicate-specific actants (obligatory or optional) and to obligatory free modifiers.²

Clearly, a given predicate verb may be ambiguous – it may have several different valency frames (verb senses), which may or may not differ in the number

² The valency frames were later enriched by quasi-valency and typical modifiers Error: Reference source not found.

and type of slots. For example, the verb “to stay” may have an ACTor (who stays) and LOC (where he stays, as an obligatory free modifier, as in “John stays at home”), but also two actants – ACTor and PATient (who stays what, as in “the governor stayed the execution”). In the valency lexicons, these are considered two different entries and correspond to two different senses of the verb. The opposite is not necessarily true: if two potential valency frames have absolutely identical slots (including the morphosyntactic form associated with each slot), they are split into two entries only if their senses are clearly distinct. This is one case where the original valency lexicons “reach” for content to help distinguish the two or more senses (cf. 4.1, examples for the verb “*založit*” ‘loan’, ‘shorten’ or ‘bookmark’).

There is another lexical-semantic issue, namely lexical-semantic conversions, which are also partly beyond the borders of the language system. There is “a content match conditioned by the sentence context (lexical occupancy), not coming out of the language-structured meaning” [15] but from the cognitive entities within extralinguistic reality. Lexical-semantic conversions are understood as relations linked to changes in valency frames of verbs resulting from the changes in the cross-referencing of the situational participants and valence modifiers. The reference to the cognitive entities within extralinguistic reality is the reason why this language phenomenon is considered as partly leaving the borders of the language system.

Lexical-semantic conversions were first addressed by Kováčová [21]. She attempts to extend the FGD by distinguishing two types of meaning: situational (cognitive, lexical) and structural (grammatical) and introduces the notion of cognitive role for a participant in a linguistically structured situation. Conversion is understood as a relationship based on the identity of the situational meaning of expressions, with a specific difference in their structural meaning.

Following (with some reservations) the work of Kováčová, Kettnerová [5] proposes a lexicographic representation of lexical-semantic conversions. The proposed representation, based on “lexical-conceptual structure,”³ captures the correspondence between situational participants and valency complementations. The author considers (similarly to Kováčová) the differentiation of two types of meaning (situational content and structural meaning) crucial for delimiting the lexical units [8], [2].

Despite these two endeavors outside of the “linguistically structured meaning” principle (i.e., distinguishing verb senses with identical valency frames and conversions), the area of lexical semantics has not yet been elaborated in a systematic way in the FGD, as Hajičová [22, p. 142] herself points out: “we are aware that lexical semantics is a domain to be investigated”.

³ http://www.glottopedia.org/index.php/Lexical_conceptual_structure

4 SEMANTIC ROLES

Although the notion of “semantic role” (SR) is generally accepted and largely used in linguistics, there is no consent about a unified definition of SRs nor real consensus about SRs’ inventory. SRs are however considered an appropriate basis for a lexical semantic representation [23]. According to the International Organization for Standardization (ISO),⁴ semantic role is defined as a mode of involvement of a participant (i.e., a conceptual semantic unit referred to by one or more lexical items in an utterance) in an eventuality (i.e., event, state, process, or action).

4.1 FGD and semantic roles

FGD mostly defines the tectogrammatic level of language description as the level of linguistically structured meaning. TG „functors,” assigned to every unit of the TG representation, are understood as functions of sentence members of the surface syntax layer. They are essentially defined on semantic basis, assuming regular (“standard”, “basic”) correspondence between the domain of cognitive (semantic) roles and the domain of functors at the level of linguistically structured meaning. The TG functors, however, are not the same as cognitive roles (it would of course contradict the FGD understanding of meaning because cognitive layer is considered already “beyond” the tectogrammatrics). For example, the TG actants are subject to shifting of cognitive roles (described in detail in [19], [20]) when the “standard” or “regular” mapping of actants to cognitive roles breaks. However, when no shifting is involved, then most TG functors (ADDR, ORIG and EFF and most free modifiers) can be well compared with cognitive, i.e., extralinguistic content. In other cases, however, the same free modifier is used for what is undoubtedly different from the cognitive point of view; e.g., LOC is used not only for place – “to be under the fence. LOC,” but also for “State”, since “State” is often expressed by similar morphosyntactic forms as locations (“be in a state of ...”, e.g., “to be under pressure.LOC”). These examples show that the TG actants and in some cases, even free modifiers are defined not only on semantic but also on (morpho)syntactic basis. This is most strongly displayed for ACTor and even more often for PATient, which are simply defined as the first and second actant (i.e., syntactically) regardless of a possible application of the shifting principle and their mapping to the cognitive role.

For illustration, the verbs used in the following examples have the same linguistically structured meaning as expressed on the TG layer but different meaning from the cognitive perspective (i.e., different sense). If we consider valency to share the same basic principles with the FGD, then there should only be one “meaning” of the verb “založit” in the valency lexicon (since the two actants involved, ACT and

⁴ <https://iss.isolutions.iso.org/obp/ui#iso:std:iso:24617:-4:ed-1:vl:en>

PAT, are the same, including their morphostyntactic realization as nominative and accusative, respectively), but in fact the lexicon has three entries, since the sense distinctions are obvious (“loan”, “shorten” and “bookmark”).

Anna.ACT *založila Marii*.PAT – ‘Ann loaned [money] to Mary.’

Anna.ACT *založila sukni*.PAT – ‘Ann shortened a skirt.’

Anna.ACT *založila stránku*.PAT *v knize*. – ‘Ann put a marker in the book.’

On the other hand, the verb „*pocházeť*” [*originate*] used in the following examples has different linguistically structured meaning as well as different cognitive content. It is thus natural and within the FGD’s [linguistic-meaning-only] principle that these verb senses have separate valency frames, with different functors at the individual slots.

Anna.ACT *pochází z Prahy*. DIR1 – ‘Ann comes from Prague.’

Dům.ACT *pochází z roku 1950*. TFRWH – ‘The house dates from 1950.’

As already discussed in Sect. 3, both Kováčová [21] as well as Kettnerová and Lopatková ([5], [8]), need to relate TG functors between two (or more) valency frames when studying lexical-semantic conversions. Kováčová’s definition of cognitive role does not take into account the mapping between deep syntax (i.e., valency in the FGD framework) and ontological meaning. By contrast, Kettnerová and Lopatková ([5], [8]) work with the term “situational participants” appearing in “situations” called “situational content”. Situational content of a verb is supposed to be an abstraction (generalization) of the event situation expressed by this verb.

To sum up, the view adopted by Kováčová, who works with the term “cognitive role” referring to content, goes beyond the TG layer, i.e., beyond the FGD framework, whereas the approach of Kettnerová and Lopatková, who refer to the “situational participant” and “situational content” and their abstraction and do use it for the description of lexical conversions, is not claimed to be part of or a direct extension of FGD. Such a reference is only used as a “guidance”, and thus it is balancing on the boundary of the FGD framework, but their approach still remains largely “within the language system”.

4.2 CzEngClass approach to semantic roles

Whereas the FGD in detail elaborated the representation of linguistically structured meaning of verbs, the representation of cognitive content of verbs within the FGD was (for principled reasons, as discussed in Sect. 2, 3 and 4) missing. Nevertheless, CzEngClass approach to semantic roles (SRs) is based on the FGD framework and is inspired mainly by the formal representation of lexical-semantic conversions as elaborated in ([5], [8]) and incorporated into the newer versions of VALLEX.⁵

⁵ <http://ufal.mff.cuni.cz/vallex/3.5/>

CzEngClass [25] strives to extend the concept of SRs to cover the whole lexicon. However, the use of SRs in CzEngClass is not the starting point: the goal is to build a bilingual Czech and English lexicon of synonym classes of verbs. SRs are an important, but not the only part of the description of the lexicon entries. The project aims primarily at delimitation of classes of synonymous verb senses by studying their semantic ‘equivalence’ in Czech-English translational context. Finding the appropriate set of SRs that characterizes each synonym class is considered to be an important tool for the specification of synonymous verb senses. The set of SRs is shared by every class member, both English and Czech. Class members are not verbs as “words” (or lemmas), but verb senses as represented by their distinct valency frames in the valency lexicons. Every SR from the given common set of SRs (Roleset) in a particular synonym class is mapped to a valency slot (represented by a TG functor) captured in the valency frame.⁶

Class: *soupeřit* – ‘compete’

Roleset (semantic roles)	Participant_1 – Participant_2 – Prize
<i>soupeřit</i> (PDT-Vallex-ID-v-w6280hsa_1181)	ACT – ADDR – PAT
<i>compete</i> (EngVallex-ID-ev-w616f1)	ACT – ADDR – PAT
<i>vie</i> (EngVallex-ID-ev-w3553f1)	ACT – ADDR – PAT
...‘fight, scrap, wrangle, wrestle’ ...	
<i>soutěžít</i> (PDT-Vallex-ID-v-w6295f1)	ACT – ADDR – PAT
<i>bojovat</i> (PDT-Vallex-ID-v-w178f1)	ACT – ADDR – PAT
... <i>utkat se, zádovít, ...</i>	

This setup, especially the introduction of SRs as the unifying element for each synonym class, is a necessary step, since otherwise it would be difficult to relate the valency (as represented in their valency frames) of the synonymous verb senses to each other, which in turn is necessary as a guidance to determine if two verb senses are synonymous or not. We believe that this is a similar reason that led to the introduction of “situational participants” in the representation of cognitive content of verbs for the purposes of describing lexical-semantic conversions. Just as Kettnerová [5] refers to the layer of “situational participants” (see Fig. 1),⁷ CzEngClass also links the layer of SRs to the layer of TG functors by an explicit mapping provided for the individual members of the synonymous class. This allows to relate possibly distinct valency slots (or even other complementations of the verb, i.e., free modifiers) among the class members, providing not only a (semi)formalized criterion for determining which verb sense should be part of the synonym class, but also to use this information in various language processing tasks.

⁶ Or outside of it, in cases when the valency frame does not list the counterpart of the SR.

⁷ <http://ufal.mff.cuni.cz/vallex/3.0/theory.html#sec-sect-valence-alternace>

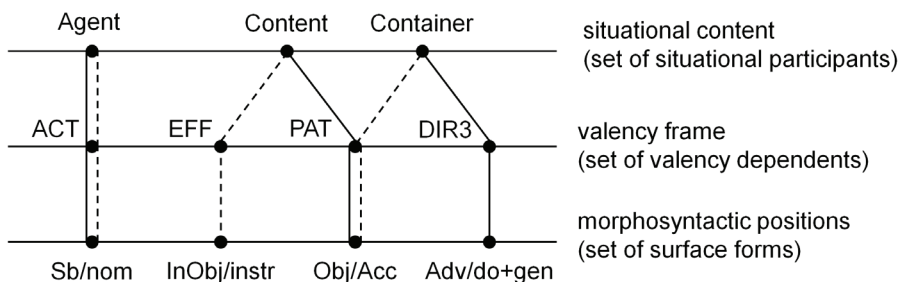


Fig. 1. Realization of locative conversion for “naplnit ‘fill’” (from [5])

Thanks to the provided mapping of SRs and TG functors there is enough information

to also explicitly relate SRs to the surface layer (which might be useful, e.g., in natural language generation). CzEngClass SRs reflect the cognitive (extralinguistic) characteristics of the verbal complementations as activated in the “standard” contexts that we imagine to generalize across (or “abstract from”) many possible situations which are described by the utterances that use or might use the verbs from a single class. Therefore, CzEngClass’ SRs are context-dependent semantic relations. In this respect, our concept is close to that of the Frame Semantics⁸, where the study of meaning is considered to be the study of cognitive scenes that are created or activated by utterances [24, p. 192].

Since the CzEngClass project is a work in progress, there are still unanswered questions, such as whether the term semantic role is appropriate, or should TG functors be defined as forms for cognitive functions (by introducing a separate semantic (cognitive) layer “above” the TG layer), what will be the exact relation of this layer to the FGD, and more.

5 CONCLUSIONS

In our paper, we have presented some considerations regarding the term “semantic roles” in relation to the Functional Generative Description theory. The introduction of “semantic roles” is, in our opinion, well motivated by the need to define lexical synonymy (as approached in the CzEngClass project), especially when different valency frames are to be related for different verbs. This has been already discussed, albeit in more or less different contexts, by Kováčová and Kettnerová and Lopatková for similar reasons while studying lexical conversions. We have concluded that such a notion is indeed important and necessary for the aforementioned goals.

⁸ <http://lingo.stanford.edu/sag/papers/Fillmore-Baker-2011.pdf>

In the future, we will primarily focus on the relation of Semantic Roles across the synonym classes. The question here is whether they can be shared across them, and under which circumstances. In doing so, we will continue to relate them to the FGD notions related to valency as well as to the way they are captured in the FGD-based valency lexicons.

ACKNOWLEDGMENTS

This work has been supported by the grant No. GA17-07313S of the Grant Agency of the Czech Republic. It uses resources hosted by the LINDAT/CLARIN (LINDAT/CLARIAH-CZ) Research Infrastructure, projects No. LM2015071 and LM2018101, supported by the Ministry of Education and Youth of the Czech Republic.

References

- [1] Lopatková M. (2003). Valency in the Prague Dependency Treebank: Building the Valency Lexicon. In *The Prague Bulletin of Mathematical Linguistics*, 79–80, pages 37–60, MFF UK.
- [2] Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., Žabokrtský, Z., and Barančíková, P. (2018). VALLEX 3.5 – Valenční slovník českých sloves. Charles University, Prague, Accessible at: <http://ufal.mff.cuni.cz/vallex/3.5/>
- [3] Kettnerová, V., Lopatková, M., and Hrstková, K. (2008). Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange. In LNCS 5246, *Proceedings of the 11th International Conference TSD 2008*, pages 109–116, Berlin/Heidelberg, Springer.
- [4] Kettnerová, V., Lopatková, M., and Bejček, E. (2012). Mapping Semantic Information from FrameNet onto VALLEX. In *The Prague Bulletin of Mathematical Linguistics*, 97, Praha, Univerzita Karlova.
- [5] Kettnerová, V. (2012). Lexikálně-sémantické konverze ve valenčním slovníku. Ph.D. thesis, Prague, Czech Republic: Charles University, 220 p.
- [6] Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). Valenční slovník českých sloves VALLEX. Praha, Czechia: Nakladatelství Karolinum, 698 p.
- [7] Kettnerová, V., Barančíková, P., and Lopatková, M. (2016). Lexicographic Description of Complex Predicates in Czech: Between Lexicon and Grammar. In *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, pages 893–904, Tbilisi, Georgia: Tbilisi University Press.
- [8] Kettnerová, V., Lopatková, M., and Bejček, E. (2012). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In *Proceedings of the 15th EURALEX International Congress*, Department of Linguistics and Scandinavian Studies, pages 434–443, Oslo, Norway: University of Oslo.
- [9] Hajičová, E., and Panevová, J. (1985). Valency (case) frames of verbs. In *Contributions to Functional Syntax, Semantics and Language Comprehension. Linguistic and Literary Studies in Eastern Europe* 16. pages 147–181.
- [10] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel, and Prague, Academia.

- [11] Dokulil, M., and Daneš, F. (1958). K tzv. významové a mluvnické stavbě věty [On the so-called semantic and grammatical structure of the sentence], *O vědeckém poznání soudobých jazyků*, pages 231–46, Prague.
- [12] Sgall, P. and Panevová, J. (1976). Obsah, význam a gramatika se sémantickou bází. *Slovo a slovesnost*, 37, pages 14–25.
- [13] Hajičová, E., and Sgall, P. (1980). Linguistic meaning and knowledge representation automatic understanding of natural language. In *PBML 34*, pages 5–21.
- [14] Panevová, J. (2010). Ke vztahu kognitivního obsahu a jazykového významu. *Korpus – gramatika – axiologie*, 1(1), pages 30–40, Hradec Králové, Czech Republic: Gaudeamus.
- [15] Panevová, J., Hajičová, E., and Sgall P. (2002). Úvod do teoretické a počítačové lingvistiky I. – Teoretická lingvistika. Praha, Karolinum.
- [16] Sgall, P. (2006). Language in its multifarious aspects. Praha, Univerzita Karlova.
- [17] Hajič, J., Panevová J., Uřešová Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Sweden: Vaxjo University Press.
- [18] Lopatková, M., Žabokrtský, Z., and Kettnerová, V. (2008). Valenční slovník českých sloves. Praha, Karolinum, 382 p.
- [19] Panevová, J. (1974). On verbal frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 22, pages 3–40.
- [20] Panevová, J. (1975). On verbal frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 23, pages 17–52.
- [21] Kováčová, K. (2005). Konverzivnost jako systémový vztah (thesis). Praha, Univerzita Karlova, Filozofická fakulta.
- [22] Hajičová E. (2017). *Syntax-Semantics Interface*. Praha, Czechia: Karolinum, 300 p.
- [23] Levin, B., and Rappaport H. M. (2005). *Argument Realization*. Cambridge, Cambridge University Press.
- [24] Fillmore, C. J. (1977). The Case for Case Reopened. In P. Cole, and J. M. Sadock (eds.), *Syntax and Semantics 8: Grammatical relations*, pages 59–81, New York, San Francisco, London, Academic Press.
- [25] Uřešová Z., Fučíková E., and Hajičová E. (2018). CzEngClass – Towards a Lexicon of verb Synonyms with Valency linked to Semantic Roles. *Jazykovedný časopis / Journal of Linguistics*, 68(2), pages 364–371.
- [26] Uřešová, Z., Fučíková, E., and Šindlerová, J. (2016). CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105, pages 17–50.
- [27] Cinková, S. (2006). From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings LREC 2006*, pages 2170–2175, Genova.

INTRODUCING SEMANTIC LABELS INTO THE DeriNet NETWORK

MAGDA ŠEVČÍKOVÁ – LUKÁŠ KYJÁNEK

Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Prague, Czech Republic

ŠEVČÍKOVÁ, Magda – KYJÁNEK, Lukáš: Introducing semantic labels into the DeriNet network. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 412 – 423.

Abstract: The paper describes a semi-automatic procedure introducing semantic labels into the DeriNet network, which is a large, freely available resource modeling derivational relations in the lexicon of Czech. The data were assigned labels corresponding to five semantic categories (diminutives, possessives, female nouns, iteratives, and aspectual meanings) by a machine learning model, which achieved excellent results in terms of both precision and recall.

Keywords: derivation, semantic category, comparative semantic concepts, suffix, machine learning

1 INTRODUCTION

Although word-formation in general and derivation in particular is defined as a process affecting both form and meaning of words, most language resources that focus on derivation lack explicit semantic information. The present paper describes a recent account of introducing semantic labels into the DeriNet network, which is a large, freely available resource modeling Czech derivation [25].

After a brief overview of how meaning is approached in selected theoretical treatments of derivation and how it is captured in existing language resources (Section 2), basic facts on the DeriNet network are summarized in Section 3. The pilot experiment on semantic labeling of derivational relations in DeriNet is described in Section 4. Adhering to basic principles of a cross-linguistic proposal of comparative semantic concepts in affixation [2], we have chosen five semantic categories to be assigned by a semi-automatic procedure using machine learning techniques. Results of the experiment and future steps are discussed in Section 5.

2 SEMANTICS OF DERIVATIONAL RELATIONS IN EXISTING DESCRIPTIONS AND LANGUAGE RESOURCES

2.1 Theoretical accounts of meaning in derivation

An elaborate description of word-formation in Czech was proposed by Dokulil [3] and, since then, broadly accepted and applied in all reference grammars of Czech,

incl. the representative volume by Dokulil et al. [4] and the latest reference grammars, e.g. [28]. As Dokulil's account proceeds primarily in the meaning-to-form direction, a sort of semantic classification is, in fact, an inherent part of descriptions of word-formation in Czech grammars. A closer look reveals, however, that the descriptions are organized, first, according to the part-of-speech category of the derivatives and, second, according to the part-of-speech category of the base words, and only then the meaning of affixes is taken into account. Since semantics is used as a third-level criterion, derivatives with the same derivational meaning are split into several subgroups if belonging to different part-of-speech categories. An even more detailed, though even more fragmented description of meaning, is provided by the recent dictionary of affixes used in Czech [26].

A theoretical approach to derivational meanings that we would like to adhere to in semantic labeling of the DeriNet network is anchored in linguistic discussion on comparative semantic concepts, which are argued to be more adequate for cross-linguistic studies than established grammatical categories rooted in descriptions of particular languages [6]. Applying this discussion to derivation, Bagasheva [2] proposes a set of 51 comparative semantic concepts (Table 1). The concepts, designed as language-independent, are not limited to a particular type of affixation (prefixation, infixation etc.) and are applied across part-of-speech categories.

<i>ABILITY</i>	<i>DESIDERATIVE</i>	<i>INCEPTIVE</i>	<i>PRIVATIVE</i>	<i>SIMILATIVE</i>
<i>ABSTRACTION</i>	<i>DIMINUTIVE / ATTENUATIVE</i>	<i>INSTRUMENT</i>	<i>PROCESS</i>	<i>SINGULATIVE</i>
<i>ACTION</i>	<i>DIRECTIONAL</i>	<i>ITERATIVE</i>	<i>PURPOSIVE</i>	<i>STATE</i>
<i>AGENT</i>	<i>DISTRIBUTIVE</i>	<i>LOCATION</i>	<i>QUALITY</i>	<i>SUBITIVE</i>
<i>ANTICAUSATIVE</i>	<i>DURATIVE</i>	<i>MANNER / VIEWPOINT</i>	<i>RECIPROCAL</i>	<i>TERMINATIVE</i>
<i>AUGMENTATIVE / AMELIORATIVE / INTENSIVE</i>	<i>DWELLER</i>	<i>ORNATIVE</i>	<i>REFLEXIVE</i>	<i>TEMPORAL</i>
<i>CAUSATIVE</i>	<i>ENTITY</i>	<i>PATIENT</i>	<i>RELATIONAL</i>	<i>UNDERGOER</i>
<i>COLLECTIVITY</i>	<i>EXPERIENCER</i>	<i>PEJORATIVE</i>	<i>RESULTATIVE</i>	
<i>COMITATIVE</i>	<i>FEMALE</i>	<i>PERCEPTIVE</i>	<i>REVERSATIVE</i>	
<i>COMPOSITION</i>	<i>HYPERONYMY</i>	<i>PLURIACTIONALITY</i>	<i>SATURATIVE / TOTAL</i>	
<i>CUMULATIVE</i>	<i>HYPONYMY</i>	<i>POSSESSIVE</i>	<i>SEMELFACTIVE</i>	

Tab. 1. Comparative semantic concepts proposed by Bagasheva [2]

Another inspiring approach to derivational meanings, though not applied in our work, was elaborated in the Meaning-Text Theory. Derivational relations between words are captured by a subset of Lexical Functions, which are defined as mathematical functions whose arguments and values are lexical units. Lexical Functions were applied in the Explanatory Combinatorial Dictionary ([15], [16]).

2.2 Meaning in language resources focusing on derivational morphology

Even if several resources focusing on derivational morphology have been made available for selected European languages in the last decade (see [13] for a detailed overview), semantic issues are, to the best of our knowledge, addressed in more or less explicit way only in some French resources (cf. Morphonette and Démonette; [7], [8]) and in the Czech resource Derivancze [20].

Derivancze is a tool that searches dictionary data for derivations. For an input word, this tool provides its base word and a word or words immediately derived from it, if available in the data. Each of 255 thousand derivational relations contained in this resource was assigned a semantic label. The set of a total of 17 semantic labels was extracted from existing resources, esp. from the morphological analyzer [18] and Czech WordNet [19].

Labels in Derivancze differ from our approach described below in that some of them are more fine-grained (e.g. female surnames are labeled differently from female counterparts of common nouns in Derivancze), they seem to be limited to a particular part of speech (e.g. the diminutive label is attested with nouns only) and, moreover, they cannot be used as a feature for searching the data.

Pieces of information that relate to semantics of Czech derivations can be found also in resources which do not have word-formation or derivation as their primary focus; cf. morphological analyzers and corresponding dictionaries and tools ([5], [22], [23]) and general and specialized lexicographic resources ([9], [14]). Several of them were used in compilation of the training and test data sets for our labeling experiment (cf. Section 4.2).

3 DERINET

The DeriNet network has been developed since 2013 as a database of Czech words connected with links corresponding to derivational relations [25]. In DeriNet, the relations between derived words and their base words are modeled as an oriented graph. Nodes of the graph correspond to lexemes, edges represent derivational steps between them, pointing from the base word to the derived one. Each derivative has at most one base word. Thus, a primary (unmotivated) word is the root of the tree and all its derivatives are organized according to their morphemic and semantic complexity from the simplest to the most complex ones; see the tree structure in Fig. 3.

Lexemes in DeriNet were extracted from the MorfFlex CZ dictionary [5], which covers a major part of the lexicon of contemporary Czech including proper names, archaic words, low-frequency words and regular, automatically generated coinages without respect to whether they are attested in a corpus. Derivational relations between lexemes were created semi-automatically under manual control, preferring high precision to recall.

The current version, DeriNet 2.0 [30], contains more than 1 million lexemes connected with more than 809 thousand derivational relations. The DeriNet data are available for download (Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, CC BY-NC-SA 3.0), and can be searched online by the DeriSearch tool.¹

4 A SEMI-AUTOMATIC APPROACH TO ASSIGNING WORD-FORMATION RELATIONS WITH SEMANTIC LABELS

4.1 Linguistic decisions on the design of the experiment

The task of introducing semantic labels into DeriNet was a challenge mainly due to the size of the data, which does not allow for a large-coverage manual annotation, and second, due to that the resource is still under construction (edges are either added, or deleted in course of revisions). The task was thus designed as a semi-automatic procedure and limited to only five semantic categories in this pilot phase, namely to:

- *DIMINUTIVE*: in line with the proposal of comparative semantic concepts (and unlike the corresponding semantic label in Derivancze), we assume this category to be expressed by words belonging to different part-of-speech categories in Czech by using suffixes (cf. examples for this and the other labels in Table 2),
- *FEMALE*: this category subsumes female counterparts of both masculine animate common nouns and proper nouns in Czech, both derived by suffixation,
- *POSSESSIVE*: in our experiment this semantic category is limited to denominal derivation of possessives (with the affixes *-ův* and *-in*) that relate to an individual, and is thus narrower than the respective comparative semantic concept (since not including possessives related to a group of individuals or to a species like *pes* ‘dog’ > *psí* as in *psí srst* ‘dog hair’),
- *ITERATIVE*: following up the long-lasting linguistic debate on this category (reviewed by Ševčíková and Panevová [24] with respect to the DeriNet data), this category is assumed to be limited to imperfective verbs derived from imperfectives in Czech by different suffixes,
- *ASPECT*: this label, as the only one in our experiment, has no counterpart in the repertoire of comparative semantic concepts; it relates to a previous decision made in the course of the build-up of DeriNet to include pure aspectual counterparts into the data since the category of aspect is, unlike other inflectional categories of verbs, conveyed by derivational morphemes [24]; this semantic label is applied to suffixation of verbs from verbs when changing aspect.

In accordance with the focus of DeriNet, the semantic labels are meant to reflect the structural, word-formation meaning while lexical shifts are not taken into

¹ <http://ufal.mff.cuni.cz/derinet/search>

consideration [27]. The aim of the labeling experiment was to apply the labels to the entire DeriNet data.

example	label to assign
<i>pes</i> ‘dog’ > <i>psík</i> ‘small dog’	<i>DIMINUTIVE</i>
<i>žlutý</i> ‘yellow’ > <i>žlutoučký</i> ‘yellowish’	<i>DIMINUTIVE</i>
<i>málo</i> ‘little’ > <i>maličko</i> ‘very little’	<i>DIMINUTIVE</i>
<i>spát</i> ‘to sleep’ > <i>spinkat</i> ‘to sleep’ (baby talk)	<i>DIMINUTIVE</i>
<i>učitel</i> ‘teacher’ > <i>učiteka</i> ‘female teacher’	<i>FEMALE</i>
<i>Jaroslav</i> (male first name) > <i>Jaroslava</i> (female first name)	<i>FEMALE</i>
<i>Novák</i> (male surname) > <i>Nováková</i> (female surname)	<i>FEMALE</i>
<i>učitel</i> ‘teacher’ > <i>učitelův</i> ‘teacher’s’	<i>POSSESSIVE</i>
<i>učitelka</i> ‘female teacher’ > <i>učitelčin</i> ‘female teacher’s’	<i>POSSESSIVE</i>
<i>chodit</i> ‘to walk.IPFV’ > <i>chodívat</i> ‘to walk.IPFV repeatedly’	<i>ITERATIVE</i>
<i>kupovat</i> ‘to buy.IPFV’ > <i>kupovávat</i> ‘to buy.IPFV repeatedly’	<i>ITERATIVE</i>
<i>chytit</i> ‘to catch.PFV’ > <i>chytat</i> ‘to catch.IPFV’	<i>ASPECT</i>
<i>štěkat</i> ‘to bark.IPFV’ > <i>štěknout</i> ‘to give a bark.PFV’	<i>ASPECT</i>

Tab. 2. Examples of semantic categories

4.2 Compilation of the training and test data sets

For the machine learning experiment, training and test data sets were prepared in four subsequent steps. First, relevant base-derivative pairs were extracted from existing language resources, namely from the monolingual dictionary of Czech [9] (examples of all five categories; see Fig. 1), from MorfFlex CZ [5] (instances of diminutives, possessives and female names; Fig. 2), and from the VALLEX dictionary [14] (examples to be assigned the *ITERATIVE* and *ASPECT* labels). Only those pairs were included that are attested in DeriNet and a derivational link is established in the data.

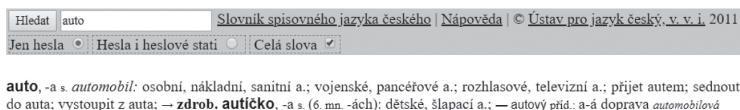


Fig. 1. Entry of the noun ‘auto’ car’ in the monolingual dictionary [9] (<https://ssjc.ujc.cas.cz/>), containing the diminutive autíčko ‘small car’ (incl. the semantic category)

word	morphological lemma	comment
astrofyzička	astrofyzička^(^FM*3k)	remove 3 letters, add 'k', and get: astrofyzik
bioložka	bioložka^(^FM*3g)	remove 3 letters, add 'g', and get: biolog
cizinka	cizinka^(^FM*2ec)	remove 2 letters, add 'ec', and get: cizinec

Fig. 2. Candidates of female nouns extracted from MorFlex CZ [5]. Semantic category (^FM for female noun) and base words are encoded in the morphological lemma (see the “comment”).

Second, diminutive and possessive suffixes that were not covered by the exploited resources were searched for in reference books (esp. [11], [17], [26]) and used to identify relevant derivatives in the DeriNet data. After a manual annotation, these instances were added to those extracted from the dictionaries.

These instances, which positively substantiated the relations under consideration, were complemented by negative examples (to be assigned none of the five semantic categories) in the third step. The negative examples were extracted from DeriNet under manual control and assigned a sixth label (*none*). In this way, a data set consisting of a total of 14,752 both positive and negative instances was compiled, see Table 3.

label	<i>DIMINUTIVE</i>	<i>FEMALE</i>	<i>POSSESSIVE</i>	<i>ITERATIVE</i>	<i>ASPECT</i>	<i>none</i>
count	3,303	1,449	2,252	1,555	3,719	2,474

Tab. 3. Portions of examples for each semantic label and the none label within the data set of a total of 14,752 instance

In the fourth step, the data set consisting of positive and negative examples was assigned features in Table 4. All features were binarized according to the labeled data, which increased dimensionality, mainly because of the n-gram features (encoded as one-hot).

The data were then divided into three data sets (with no overlaps):

- 80% of the data were used as a training data set for training the machine learning model,
- 10% of the data served as a development test data set to find adequate probability thresholds for each semantic label,
- 10% of the data were used as an evaluation test data set for evaluation of the model.

4.3 Development of the machine learning model

Starting with a preliminary set of supervised machine learning experiments using the Python 3 scikit-learn module [21], Multinomial Logistic Regression (MLR) has been chosen as the most promising method for the semantic labeling task, showing better results than Decision Tree and Naive Bayes methods.

feature	values	comment
part-of-speech category of the derivative	N (noun), A (adjective), V (verb), D (adverb)	- source: DeriNet 1.7 [29]
part-of-speech category of the base word	N (noun), A (adjective), V (verb), D (adverb)	- source: DeriNet 1.7
gender of the derivative	M (masculine animate), I (masculine inanimate), F (feminine), N (neuter)	- assigned with nouns - source: MorfFlex CZ [5]
gender of the base word	M (masculine animate), I (masculine inanimate), F (feminine), N (neuter)	- assigned with nouns - source: MorfFlex CZ
aspect of the derivative	PFV (perfective), IPFV (imperfective), B (biaspectual)	- assigned with verbs - source: MorfFlex CZ, SYN2015 corpus [12], VALLEX [14]
aspect of the base word	PFV (perfective), IPFV (imperfective), B (biaspectual)	- assigned with verbs - source: MorfFlex CZ, SYN2015 corpus, VALLEX
possessivity tag	1 (with possessives), 0 (others)	- assigned with the derivative - source: MorfFlex CZ
final n-grams of the derivative	final bi-, tri-, tetra-, penta-, hexagrams	
final n-grams of the base word	final bi-, tri-, tetra-, pentagrams	
semantic label of the derivational relation	<i>DIMINUTIVE, POSSESSIVE, FEMALE, ITERATIVE, ASPECT, none</i>	- source: monolingual dict. [9], MorfFlex CZ, VALLEX, manual annotation

Tab. 4. Features to assign with the training and test data

MLR is a generalization of the logistic regression method to multiple target tasks. For all given features of each class, MLR estimates adequate regression parameters. As an output, the MLR method returns probability values based on logistic sigmoid function for each target class ([1], [10]).

To set the MLR model for prediction of semantic labels, *newton-cg* solver was used, which is predefined in scikit-learn. The number of iterations was increased up to one thousand to converge. The goal of the MLR model to be trained was to classify examples according to the most probable semantic label, taking into account the highest possible precision.

Based on the performance of the method on the development test data set, the following thresholds were determined for individual semantic labels in order to further increase the precision: 0.75 for *DIMINUTIVE*, 0.4 for *FEMALE*, 0.4 for *POSSESSIVE*, 0.5 for *ITERATIVE*, and 0.4 for *ASPECT*. If the probability of the most probable semantic label predicted by the model was below the particular threshold, the semantic label was not accepted (changed to *none*). The results of MLR on the training and evaluation test data sets are reported in Table 5.

	accuracy	precision	recall	f1-score
training data set	0.992	0.991	0.992	0.991
evaluation test data	0.986	0.984	0.984	0.984
sample of predicted data	0.971	0.962	0.963	0.962

Tab. 5. Evaluation of the trained MLR model on the training data, evaluation test data, and manually annotated random sample of 2,000 relations from predicted data

The MLR model was applied to the previous version of the DeriNet data (DeriNet 1.7; [29]), which were previously assigned the same features as the training and test data (except for the semantic label feature; see Table 4). The MLR model assigned one of the five semantic labels to 150,521 derivational relations in total. The *POSSESSIVE* label was the most frequent one (predicted with 88,620 derivational relations), followed by the *FEMALE* label (28,510 rel.), *ASPECT* (15,459 rel.), *ITERATIVE* (11,890 rel.), and *DIMINUTIVE* (6,042 rel.).

The precision and recall of the labeling procedure were evaluated on a randomly selected sample of 2,000 relations assigned either one of the five semantic categories or the *none* label; see Table 5 for evaluation of the sample as a whole and Table 6 for details on individual labels.

gold / predicted	<i>DIMINUTIVE</i>	<i>FEMALE</i>	<i>POSSESSIVE</i>	<i>ITERATIVE</i>	<i>ASPECT</i>	<i>none</i>
<i>DIMINUTIVE</i>	62	0	0	0	0	4
<i>FEMALE</i>	1	296	0	0	0	3
<i>POSSESSIVE</i>	0	0	905	0	0	1
<i>ITERATIVE</i>	0	0	0	135	4	0
<i>ASPECT</i>	0	0	0	3	170	1
<i>none</i>	1	39	1	0	0	374

precision	0.969	0.982	0.999	0.985	0.987	0.948
recall	0.983	0.941	0.999	0.988	0.987	0.976

Tab. 6. Confusion matrix based on manual annotation of a random sample of 2,000 relations and precision and recall calculated for individual labels in the sample

Semantic labels, as assigned in the machine learning experiment, are part of the current version of the DeriNet network (DeriNet 2.0, [30]). Semantic labels can be used for searching the data by the DeriSearch tool. A sample tree containing semantic labels is displayed in Fig. 3.

5 DISCUSSION AND FUTURE WORK

The word-formation system of Czech is characterized by homonymy of affixes,² on the one hand, and synonymy of affixes, on the other. Many affixes convey more

² The term “homonymy” [17] or “polyfunctionality” [26] is preferred to “polysemy” in recent accounts.

than one meaning, cf. *-ka* is used to express the diminutive meaning in *skříňka* ‘small cupboard’ while in *učitelka* ‘female teacher’ it falls under the *FEMALE* category, but it occurs also in instrument nouns (*žehlička* ‘iron’) and other formations. From the opposite perspective, a particular meaning is usually expressed by several, formally different affixes, cf. the suffixes *-ka*, *-yně*, *-ice*, *-ová* in the *FEMALE* category.

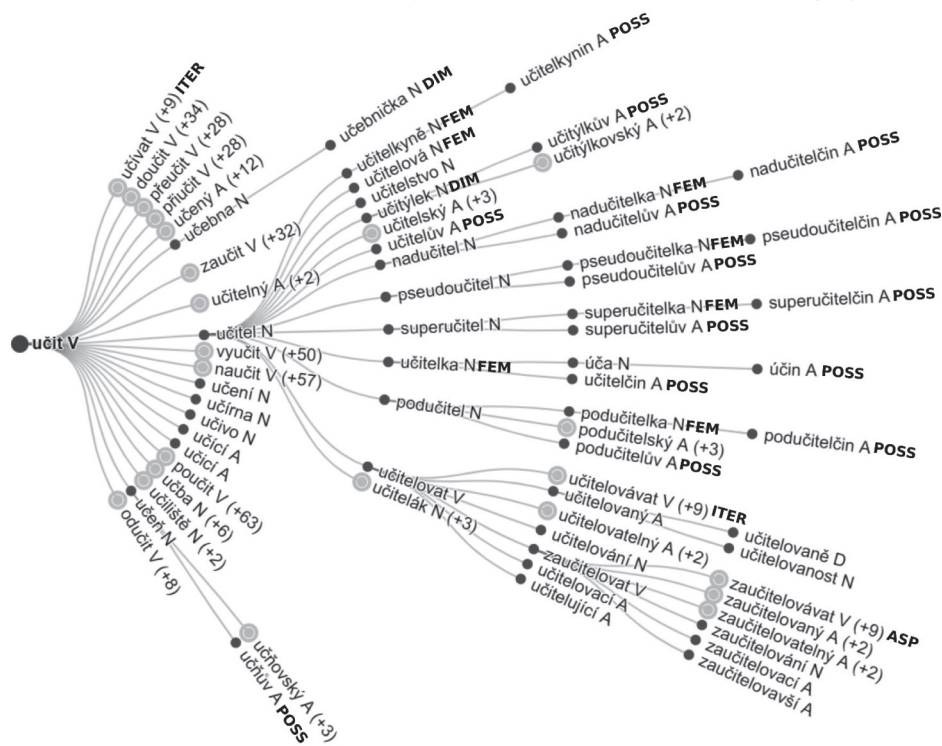


Fig. 3. The derivational tree with the root *učít* ‘to teach’ assigned with semantic labels predicted by the experiment. Semantic labels (in bold) are displayed with the derivative. Number of further derivatives, if hidden in the figure for clarity, is given in parentheses

Even though the presented labeling experiment was limited to a small number of semantic categories, its excellent results in terms of both precision and recall document, in our opinion, that the make-up of the DeriNet network (derivational families organized into rooted tree structures) and the features included in the machine learning experiment provided a sufficient basis for the resolution of homonymy in most cases.

A detailed analysis of incorrectly predicted labels draws attention to mostly peripheral or borrowed formations that are still formally and morphologically close to the correct representatives of the particular semantic categories but differ in meaning. See Table 7 for examples of relations that were incorrectly assigned the

FEMALE or the *DIMINUTIVE* label. Examples like *textař* ‘lyricist’ > *textařina* ‘profession of a lyricist’ point out the usefulness of animateness as a morphological feature of feminine nouns (not available in MorfFlex CZ) since only animate feminines are to be considered female counterparts of animate masculine nouns.

base word	derivative	incorrectly predicted label
<i>ježek</i> ‘hedgehog’	<i>ježura</i> ‘echidna’	<i>FEMALE</i>
<i>fořt</i> ‘forest warden’	<i>fořtovna</i> ‘forest warden’s lodge’	<i>FEMALE</i>
<i>profesor</i> ‘professor’	<i>profesura</i> ‘professorship’	<i>FEMALE</i>
<i>textař</i> ‘lyricist’	<i>textařina</i> ‘profession of lyricist’	<i>FEMALE</i>
<i>smrt</i> ‘death’	<i>smrtka</i> ‘Death’	<i>DIMINUTIVE</i>
<i>had</i> ‘snake’	<i>hadice</i> ‘hose’	<i>DIMINUTIVE</i>

Tab. 7. Examples of pairs with incorrect labels

The labels *ASPECT* and *ITERATIVE* were not sufficient to cover a handful of examples in which a perfective verb is captured as a derivative of another perfective in DeriNet (e.g. *oloupat* ‘to peel.PFV’ > *oloupnout* ‘to peel.PFV’, *chytit* ‘to catch.PFV’ > *chytnout* ‘to catch.PFV’). These relations correspond to the semelfactive semantic concept in Bagasheva’s set; the respective label will be included in the next round of semantic labeling.

6 CONCLUSIONS

The semi-automatic procedure introducing semantic labels into the DeriNet network, which was described in the present paper, was carried out as a pilot experiment to verify its applicability to large, specifically organized data. The approach was limited to five semantic categories that are conveyed (mainly) by ambiguous suffixes and, with the exception of derivation of possessives, do not change the part-of-speech category of the base word. The fact that the assigned categories are rooted in the proposal of comparative semantic concepts might not be obvious in this pilot phase, as we chose basic categories that are involved not only in Bagasheva’s proposal. However, the choice of a particular linguistic background is essential for perspectives of further development and usability of the data.

The labeling task started with extraction of relevant features from existing resources in order to compile high-quality training and test data sets with enough examples of each category in an efficient way. The machine learning model was designed with the aim to be replicable after any changes in the DeriNet data and to be extendable to other labels. More than 150 thousand semantic labels were predicted by the model, by achieving both an excellent precision and recall. Analysis of the data with both correctly and incorrectly predicted labels is expected to be relevant for our next steps as well as, importantly, for linguistic insights into derivations.

ACKNOWLEDGMENTS

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation and by the Student Faculty Grant UKMFF/160753/2018-2/SFG of the Faculty of Mathematics and Physics, Charles University. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH-CZ project (LM2015071, LM2018101).

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*. 2nd edition. New York, John Wiley & Sons.
- [2] Bagasheva, A. (2017). Comparative semantic concepts in affixation. In *Competing Patterns in English Affixation*, pages 33–65, Bern, Peter Lang.
- [3] Dokulil, M. (1962). *Tvoření slov v češtině: Teorie odvozování slov*. Praha, ČSAV.
- [4] Dokulil, M. et al. (1986). *Mluvnice češtiny 1*. Praha, Academia.
- [5] Hajič, J., and Hlaváčová, J. (2013). *Morfflex CZ*. LINDAT/CLARIN digital library at ÚFAL MFF UK. Accessible at: <http://hdl.handle.net/11858/00-097C-0000-0015-A780-9>
- [6] Haspelmath, M. (2010). Comparative concepts and descriptive categories in cross-linguistic studies. *Language*, 86(3), pages 663–687.
- [7] Hathout, N. (2010). *Morphonette: a morphological network of French*. CoRR, arXiv, abs/1005.3902.
- [8] Hathout, N., and Namer, F. (2014). *Démonette, a French derivational morpho-semantic network*. *Linguistic Issues in Language Technology*, 11(5), pages 125–168.
- [9] Havránek, B. (ed.; 1960–1971). *Slovník spisovného jazyka českého*. Praha, Academia.
- [10] Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd edition. New York, John Wiley & Sons.
- [11] Karlík, P. ed. (2016). *Nový encyklopedický slovník češtiny*. Praha, NLN.
- [12] Křen, M. et al. (2015). *SYN2015: reprezentativní korpus psané češtiny*. Praha, ÚČNK FF UK. Accessible at: <http://www.korpus.cz>
- [13] Kyjánek, L. (2018). *Morphological Resources of Derivational Word-Formation Relations*. Technical report no. 2018/TR-2018-61. Praha, ÚFAL MFF UK.
- [14] Lopatková M. et al. (2016). *VALLEX 3.0*. LINDAT/CLARIN digital library at ÚFAL MFF UK. Accessible at: <http://hdl.handle.net/11234/1-2307>
- [15] Mel'čuk, I. (2006). *Explanatory Combinatorial Dictionary*. In *Open Problems in Linguistic and Lexicography*, pages 225–355, Monza, Polimetrica.
- [16] Mel'čuk, I., and Žolkovskij, A. K. (1984). *Tolkovo-kombinatornyj slovar' russkogo jazyka*. Vienna, Wiener Slavistische Almanach. Sonderband 14.
- [17] Nekula, M. et al. (2012). *Příruční mluvnice češtiny*. 2nd edition. Praha, NLN.
- [18] Osolobě, K. et al. (2002). A Procedure for Word Derivational Processes Concerning Lexicon Extension in Highly Inflected Languages. In *Proceedings of LREC 2002*, pages 1254–1259, Paris, ELRA.
- [19] Pala, K., and Hlaváčková, D. (2007). *Derivational Relations in Czech WordNet*. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 75–81, Prague, ACL.

- [20] Pala, K., and Šmerk, P. (2015). Derivancze – Derivational Analyzer of Czech. In International Conference on Text, Speech, and Dialogue, TSD 2015, pages 515–523, Berlin, Springer.
- [21] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pages 2825–2830.
- [22] Sedláček, R., and Smrž, P. (2001). A New Czech Morphological Analyser ajka. In International Conference on Text, Speech and Dialogue, TSD 2001, pages 100–107, Berlin, Springer.
- [23] Straková et al. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of ACL 2014: System Demonstrations*, pages 13–18.
- [24] Ševčíková, M., and Panevová, J. (2018). Derivation of Czech verbs and the category of aspect. *Linguistica Copernicana*, 2018(15), pages 79–93.
- [25] Ševčíková, M., and Žabokrtský, Z. (2014). Word-Formation Network for Czech. In *Proceedings of LREC 2014*, pages 1087–1093, Paris, ELRA.
- [26] Šimandl, J. ed. (2016). *Slovník afixů užívaných v češtině*. Praha, Karolinum.
- [27] Štekauer, P. (2005). *Meaning Predictability in Word Formation: Novel, context-free naming units*. Amsterdam, John Benjamins.
- [28] Štícha, F. et al. (2018). *Velká akademická gramatika spisovné češtiny 1*. Praha, Academia.
- [29] Vidra, J. et al. (2018). *DeriNet 1.7*. Praha, ÚFAL MFF UK. Accessible at: <http://ufal.mff.cuni.cz/derinet>
- [30] Vidra, J. et al. (2019). *DeriNet 2.0*. LINDAT/CLARIN digital library at ÚFAL MFF UK. Accessible at: <http://hdl.handle.net/11234/1-2995>

NON-SYSTEMIC VALENCY BEHAVIOR OF CZECH DEVERBAL NOUNS BASED ON THE NomVallex LEXICON

VERONIKA KOLÁŘOVÁ¹ – ANNA VERNEROVÁ¹ – JONATHAN VERNER²

¹Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

²Faculty of Arts, Charles University, Prague, Czech Republic

KOLÁŘOVÁ, Veronika – VERNEROVÁ, Anna – VERNER, Jonathan: Non-systemic valency behavior of Czech deverbal nouns based on the NomVallex lexicon. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 424 – 433.

Abstract: In order to describe non-systemic valency behavior of Czech deverbal nouns, we present results of an automatic comparison of valency frames of interlinked noun and verbal lexical units included in valency lexicons NomVallex and VALLEX. We show that the non-systemic valency behavior of the nouns is mostly manifested by non-systemic forms of their actants, while changes in the number or type of adnominal actants are negligible as for their frequency. Non-systemic forms considerably contribute to a general increase in the number of forms in valency frames of nouns compared to the number of forms in valency frames of their base verbs. The non-systemic forms are more frequent in valency frames of non-productively derived nouns than in valency frames of productively derived ones.

Keywords: adnominal morphemic forms, Czech deverbal nouns, non-systemic valency behavior, valency, valency lexicon

1 INTRODUCTION

When describing valency behavior of Czech deverbal and deadjectival nouns, valency of their base verbs or adjectives can be taken into consideration in order to see whether the nouns' valency properties are derivable from their base words. If this is the case, they can be understood as a result of a regular process. When the valency properties of a noun are more or less independent of its base word, these can be considered to be a result of an irregular process. Applying such a comparative approach, systemic (typical) and non-systemic (special) valency behavior of nouns is distinguished.

Up to now, the distinction between systemic and non-systemic valency behavior has been intensively studied on the material of Czech deverbal nouns (Section 3), focusing on non-systemic forms of their valency complementations [4], e.g., *varovat koho*.Acc 'to warn sb' → *varování komu*.Dat 'warning to sb', i.e., *warning addressed to sb*. In this paper, we show how the non-systemic valency behavior is represented in the current version of the NomVallex lexicon (Section 2), drawing an automatic

comparison between valency frames of nouns included in NomVallex and their base verbs included in the VALLEX lexicon (Section 4). This comparison represents the first attempt to provide statistical data on the non-systemic valency behavior of Czech deverbal nouns.¹

2 THE NOMVALLEX LEXICON

NomVallex is a valency lexicon of Czech deverbal nouns, created within the theoretical framework of the Functional Generative Description (FGD, [11]) and based on corpus data (*Czech National Corpus*, subcorpus *SYNV6* [8], and *Araneum Bohemicum Maximum* [1]).² Applying the valency theory of the FGD [10], valency properties of a noun lexical unit (LU) are captured in a valency frame which is modeled as a sequence of valency slots, supplemented with their morphemic forms. The following types of complementations may fill in the individual slots of valency frames of most deverbal nouns: obligatory or optional actants, i.e., Actor (ACT), Patient (PAT), Addressee (ADDR), Effect (EFF) and Origin (ORIG), e.g., *balení dárků*.PAT *rodiči*.ACT ‘wrapping of the presents by parents’, and obligatory free modifications, especially those with the meaning of direction, e.g., *chlapcův*.ACT *pozdní příchod do školy*.DIR3 ‘boy’s late arrival to the school’. Nouns denoting quantity (a container) usually only have one valency slot in their valency frame, an actant called Material (MAT), which is in the form of prepositionless genitive, cf. *jedno balení léků*.MAT ‘one package of medicine’.

Up to now, NomVallex has focused on deverbal nouns belonging to three semantic classes, i.e. Communication (e.g. *dotaz* ‘question’), Mental Action (e.g. *plán* ‘plan’) and Psychological Noun (e.g. *nenávisť* ‘hatred’), see [7]. The lexicon captures all lexical meanings of the nouns, differentiating also basic “notional” meanings, i.e. action (e.g., *žádání* ‘asking’, *dovtipení se* ‘inferring’), abstract result of action (e.g., *žádost* ‘request’), quality (e.g., *důvtip* ‘ingenuity’), substance (e.g., *komunikace (silnice)* ‘road’), and quantity (a container, e.g., *soubor* ‘collection’). Currently, it contains more than 400 noun lexical units.³

NomVallex relates to VALLEX [9], created within the same theoretical framework. NomVallex adopts VALLEX annotation scheme and in relevant cases it also splits the lexemes into lexical units and assigns them to the relevant semantic classes according to the base verbal lexical units captured in VALLEX. As both lexicons are available as machine readable data, an automatic comparison of any valency characteristics annotated in the lexicons is possible. The links between the pairs of corresponding verbal and noun lexical units are recorded in the noun

¹ Although there are two other valency lexicons containing Czech deverbal nouns ([12] and [2]), none of them links information on valency of the nouns to their base verbs.

² The aim of the lexicon is to cover also other nominals such as adjectives and deadjectival nouns.

³ <https://logic.ff.cuni.cz/nomvallex-beta/>

entries, by indicating the verb's identification code in the attribute *derivedV*, cf. (1) and (2).

- (1) *žádat* 'to ask'
- id: blu-v-žádat-2
ACT(Nom) ADDR(Acc) PAT(*o+Acc,inf,aby,at',zda,že*)
- (2) *žádost* 'request'
- derivedV: blu-v-žádat-2
ACT(Gen,poss,*od+Gen*) ADDR(Dat,*k+Dat*) PAT(Gen,*k+Dat,na+Acc,o+Acc,po+Loc,inf,aby,at',zda,že*)

NomVallex covers both types of Czech deverbal nouns that can denote action or an abstract result of action, namely:

- i. nouns derived from verbs by productive suffixes *-ní/-tí*, e.g., *dotazování* 'asking', *namítnutí* 'objecting', called productively derived nouns;
- ii. nouns derived from verbs by non-productive suffixes, such as *-ba*, *-a*, *-ka*, e.g., *námitka* 'objection', or by the zero suffix, e.g., *dotaz* 'question'; these nouns are called non-productively derived nouns.⁴

In order to be able to compare valency behavior of these two types of nouns, NomVallex aims at creating a lexicon entry for both the productively and the non-productively derived nouns derived from one base verb, e.g., *žádání* 'asking' as well as *žádost* 'request' derived by different suffixes from the verb *žádat* 'to ask'.

3 SYSTEMIC AND NON-SYSTEMIC VALENCY BEHAVIOR OF CZECH DEVERBAL NOUNS

In this section, we specify how the systemic and non-systemic valency behavior of Czech deverbal nouns is manifested.

3.1 Systemic valency behavior

The valency behavior referred to as systemic can be observed especially with Czech productively derived nouns, if they denote action, cf. *kontrolování* 'checking' in (4). Non-productively derived nouns manifest systemic valency behavior less frequently [3], cf. *kontrola* 'check' in (4).

- (3) *policista.Nom kontroluje vozidla.Acc*
'a policeman is checking vehicles'
- (4) *kontrolování / kontrola vozidel.Gen policistou.Ins*
'checking / check of vehicles by a policeman'

⁴ The term „non-productive“ reflects esp. the fact that not all verbs have an counterpart in nouns derived by these suffixes, cf. *přemlouvat* 'to persuade' – *přemlouvání* 'persuading' – **přemluva* 'persuasion'.

When determining their valency frames, the nouns are expected to inherit all participants that are present in the valency frame of their base verbal lexical unit, including the “verbal” character of the participants such as Actor, Patient or Addressee.

Forms of adnominal participants change in case the base verbal form is either Nom or prepositionless Acc, cf. (3) and (4), or – less frequently – if a noun or an adjective in Nom or Acc are a part of an expression containing the word *jako* ‘as’, see Table 1. We also consider a possessive form corresponding to verbal prepositionless Gen to be a systemic change, cf. (5).

- (5) *zanechat studia.Gen* ‘to quit the school’ →
jeho.poss zanechání ‘its quitting’

Verbal form	Adnominal systemic form
Nom	Gen, Ins, poss, <i>od</i> ‘from’+Gen
Gen	poss
Acc	Gen, poss
<i>jako</i> ‘as’ + Nom	<i>jako</i> ‘as’ + Gen
<i>jako</i> ‘as’ + Acc	<i>jako</i> ‘as’ + Gen
<i>jako</i> ‘as’ + adj-Acc	<i>jako</i> ‘as’ + adj-Acc

Tab. 1. Systemic changes

All forms which do not change their form are regarded to be systemic. These include prepositionless cases Gen, Dat and Ins, an infinitive, prepositional groups (PGs, e.g., *k* ‘to’+Dat), conjunctions (e.g. *že* ‘that’), content clauses, an adjective in prepositionless Ins, expressions containing the word *jako* ‘as’ (*jako*+Gen, *jako*+adj-Gen, *jako*+PG), and expressions containing preposition *za* ‘as/for’ plus an adjective in prepositionless Acc (*za*+adj-Acc).

3.2 Non-systemic valency behavior

Non-systemic valency behavior of deverbal nouns is most often and most distinctly manifested by changes in properties of its valency complementations [4]. They involve three phenomena:

- i. non-systemic forms of valency complementations (e.g., Gen → Dat, *otázat se kolegy.Gen* ‘to ask a colleague’ → *otázka kolegovi.Dat* ‘a question to-a-colleague’; Sections 4.1 and 4.2);⁵
- ii. a change (esp. a reduction) of the number of slots in the valency frame of a noun (e.g., the noun *velení* in *vrchní velení* ‘the supreme headquarters’ denotes a group of people rather than a process of commanding, as in *jeho.ACT velení armádě.PAT* ‘his commanding the army’, and thus loses ACT from its valency frame; Section 4.3);

⁵ Various factors contributing to usage of non-systemic forms, including an influence of a form of a valency complementation of a light verb in light verb constructions (e.g. *dát otázku kolegovi* ‘to address a question to a colleague’), are discussed in [3].

- iii. a change of the character of a valency complementation to exclusively nominal, e.g., Material modifying nouns denoting quantity, as in *jedno balení léků*.MAT ‘one package of medicine’, in contrast to Patient in *balení kufrů*.PAT ‘packing of bags’, denoting action. This case is however extremely rare in the NomVallex data and is not dealt with in the paper.

We assume the notional meaning of a deverbal noun that displays non-systemic valency behavior is always different from action, and thus the noun denotes an abstract result of action, quality, substance or quantity.

4 AN AUTOMATIC COMPARISON OF VERBAL AND NOUN VALENCY FRAMES

Our automatic comparison of valency frames of nouns in NomVallex and valency frames of their base verbs in VALLEX obviously only covers nouns that provide a link to their base verbal lexical unit in VALLEX. First, an automatic procedure checks whether the valency frame of the given noun lexical unit corresponds to systemic valency behavior (i.e., if the number and type of valency slots is the same as in the corresponding verbal valency frame, and if their forms are either the same or correspond to a systemic change, see Section 3.1). Second, any difference from the systemic valency behavior is indicated as non-systemic one and is captured in the noun entry in the attribute *framediff* (difference in frame).

In this Section, we only focus on differences in the number or forms of actants, leaving out free valency modifications. Comparing verb-noun pairs with equal actants, we provide the general statistics on the number of morphemic forms in noun valency frames (Section 4.1), and we present distribution of non-systemic adnominal forms across the NomVallex data (Section 4.2). A difference in the number of actants is in focus of Section 4.3.

4.1 An increase in the number of adnominal forms

Noun valency structures show various limitations compared with verbal ones:

- i. Adnominal prepositionless Gen and possessive forms may be syntactically ambiguous, being a result of different systemic changes or even some non-systemic ones (cf. Table 1 and Table 4).
- ii. Noun valency patterns are subject to certain restrictions on combinations of actants expressed by particular morphemic forms, e.g., double postnominal genitives [5], including their word order, e.g., all incongruent attributes come after the noun, prepositionless Gen comes the first, then come the other forms [13].
- iii. Morphemic forms of particular actants modifying nouns denoting an abstract result of action are rather often non-systemic, e.g., *návrh na reformy* ‘a proposal for reforms’, see [4]. However, the adnominal actants often keep the systemic forms as well, e.g., *návrh reformem* ‘a proposal of reforms’.

As a result, deverbal nouns show a strong tendency to have at their disposal more morphemic forms of their actants than their base verbs, cf. (1) and (2), which enables them to use the appropriate form depending on the syntactic structure they occur in or depending on their notional meaning.

A general statistics on the number of morphemic forms in verbal and corresponding noun valency frames is given in Table 2, showing an apparent increase in the number of adnominal forms. The verbal lexical units which correspond to several noun lexical units (as in *žádat* ‘to ask’ – *žádání* ‘asking’, *žádat* ‘to ask’ – *žádost* ‘request’) are figured in the statistics as many times as many links to noun lexical units they have. The opposite case, nouns with more than one link to a verbal lexical unit, is not included in the statistical data given in Table 2.

On average, the total number of adnominal forms is more than 43% higher than the total number of the verbal forms. Counting the number of forms per a valency frame, noun valency frames contain on average 2.2 more forms than valency frames of the corresponding verbal lexical units. The most considerable increase in number of adnominal forms can be seen in valency frames of non-productively derived nouns of Communication (the total number of adnominal forms is more than 53% higher than the total number of the forms of base verbs of Communication, which brings on average 3.3 more forms in noun valency frames).

Class	Noun's suffix	Verb -noun pairs	Base verb's forms		Adnominal forms			
			Total	Number of forms per LU	Number of forms		Number of forms per LU	
					Total	Increase of %, comp. to verbs	Total	Increase, comp. to verbal LUs
Communication	prod.	71	446	6.3	569	27.6	8.0	1.7
	non-prod.	34	207	6.1	318	53.6	9.4	3.3
Mental Action	prod.	72	330	4.6	501	51.8	7.0	2.4
	non-prod.	34	151	4.4	229	51.7	6.7	2.3
Psych.	prod.	32	124	3.9	190	53.2	5.9	2
Verb / Noun	non-prod.	13	60	4.6	87	45.0	6.7	2.1
Total	prod.	175	900	5.1	1260	40.0	7.2	2.1
	non-prod.	81	418	5.2	634	51.7	7.8	2.6
	both / average	256	1318	5.2	1894	43.7	7.4	2.2

Tab. 2. An increase in the number of adnominal forms

The same data is used in Table 3 in order to pinpoint the distribution of systemic and non-systemic forms in valency frames of nouns. Looking at the average numbers of all systemic and non-systemic forms, we can see that non-systemic forms account for 15% of the total number of adnominal forms. However, taking into consideration

whether the nouns are productively or non-productively derived, the non-systemic forms account for 24.9% of all adnominal forms in valency frames of non-productively derived nouns, while non-systemic forms in the valency frames of productively derived nouns only account for 10.1%. The most significant difference in number of non-systemic forms can be seen in valency frames of nouns of Communication; while the percentage of non-systemic forms in valency frames of non-productively derived nouns is 25.2%, these forms in valency frames of productively derived nouns only account for 5.3%. The statistical data given in Table 3 confirms results of previous manual analysis carried out on corpus data [3], showing clearly that non-productively derived nouns tend to use non-systemic forms to a higher extent than productively derived nouns. At the same time, it follows from the data that in some cases also productively derived nouns use non-systemic forms and so their valency behavior cannot be considered to be purely systemic either.

Class	Noun's suffix	Verb -noun pairs	Base verb's forms	Adnominal forms				
				Systemic		Non-systemic		Total / 100%
					%		%	
Communication	prod.	71	446	539	94.7	30	5.3	569
	non-prod.	34	207	238	74.8	80	25.2	318
Mental Action	prod.	72	330	441	88.1	60	11.9	501
	non-prod.	34	151	177	77.3	52	22.7	229
Psych. Verb / Noun	prod.	32	124	153	80.5	37	19.5	190
	non-prod.	13	60	61	70.1	26	29.9	87
Total	prod.	175	900	1133	89.9	127	10.1	1260
	non-prod.	81	418	476	75.1	158	24.9	634
	both	256	1318	1609	84.9	285	15.1	1894

Tab. 3. The number of systemic and non-systemic adnominal forms

4.2 A distribution of non-systemic forms of actants

Analyzing the adnominal forms in more detail, all non-systemic forms were classified by the actant they express and ordered according to their frequency in the lexicon data (see Table 4).⁶ Our data shows that while ACT, EFF and ORIG are only exceptionally expressed by a non-systemic form, PAT and ADDR use these forms quite often, though ADDR only with nouns of Communication. Concerning PAT and ADDR, only PAT can be expressed by an infinitive or by a content clause, starting either with a conjunction (C) or without it (CONT). Regardless this difference, we can see that the most frequent non-systemic form of both PAT and ADDR is

⁶ In Table 4, NA stands for *non applicable*, i.e. for the case when no such an actant is present in valency frames of nouns representing the particular semantic class, and the number 0 means that such an actant exists but there is no non-systemic form it is expressed by. The numbers after slash signs refer to the number of particular forms.

a prepositional group, cf. (6) and [6]. The second most frequent non-systemic form of ADDR is positionless Dat, cf. (6). As for the second most frequent form of PAT, there is no clear tendency for the examined groups of nouns to use some common non-systemic forms; their valency behavior is rather idiosyncratic and should be studied case-by-case, considering their individual valency frames.

- (6) *žádat obec.ADDR(Acc)* ‘to ask the village’ →
žádost k obci.ADDR(k+Dat) ‘request (addressed) to the-village’
žádost obci.ADDR(Dat) ‘request (addressed) to-the-village’

Functor	Noun's suffix	Class		
		Communication	Mental Action	Psychological Noun
ACT	prod.	0	Gen/1; poss/1	0
	non-prod.	0	0	0
ADDR	prod.	PG/6, Dat/3, poss/2, Gen/1 The most frequent PGs: <i>pro+Acc, k+Dat</i>	0	NA
	non-prod.	PG/18, Dat/8, Gen/3, poss/3 The most frequent PGs: <i>k+Dat, pro+Acc</i>	0	NA
EFF	prod.	0	0	NA
	non-prod.	C/1	0	NA
ORIG	prod.	NA	INS/1	NA
	non-prod.	NA	0	NA
PAT	prod.	PG/17, CONT/2, inf/2, C/1 The most frequent PGs: <i>o+Loc, na+Acc, k+Dat</i>	PG/17, Gen/7, poss/4, CONT/3, Dat/3, inf/2, C/2 The most frequent PGs: <i>o+Loc, k+Dat, nad+Ins</i>	PG/22, C/2, Ins/2, poss/1, inf/1 The most frequent PGs: <i>z+Gen, nad+Ins</i>
	non-prod.	PG/34, C/15, Gen/7, CONT/7, inf/2, poss/1 The most frequent PGs: <i>na+Acc, k+Dat, proti+Dat</i>	PG/31, C/5, CONT/5, inf/1 The most frequent PGs: <i>o+Loc, k+Dat, nad+Ins</i>	PG/18, inf/4, C/3, CONT/1 The most frequent PGs: <i>z+Gen, před+Ins, nad+Ins</i>

Tab. 4. A distribution of non-systemic adnominal forms

4.3 A difference in the number of actants

Our automatic comparison of verb-noun pairs of valency frames also marks cases of a change in the number of actants in noun valency frames. Table 5 shows that these cases are rather rare. An addition of an actant is often just a result of a decision to annotate the valency frame of the noun in a different way, compared to the base verbal lexical unit in VALLEX, rather than a manifestation of non-systemic valency behavior of the particular noun lexical unit. However, besides the case of a different annotation, a deletion of an actant can indicate non-systemic valency behavior indeed (namely a change in the notional meaning of the noun, leading to losing an actant).

Most frequently, the notional meaning of a noun changes from action to a substance (a person or a group of people as in *neschopné vrchní velení armády*.PAT ‘an incompetent army’s supreme command’, losing ACT from its valency frame, or a thing as in *jednosměrná komunikace* ‘one-way road’, losing all actants of its base verb). A deletion of PAT, accompanying ‘action → quality’ change in the notional meaning of the noun, can be exemplified by the noun *důvtip* ‘ingenuity’, cf. the verbal construction in (7) and the nominal construction in (8), out of which the latter cannot be modified by PAT in any morphemic form.

(7) *generál*.ACT *se dovtípil něčeho/že*.PAT

‘a general has inferred sth/that’

(8) *důvtip generála*.ACT

‘the general’s ingenuity’

Actant	No change	An actant added	An actant deleted
ACT	282	0	11
ADDR	95	9	12
EFF	33	3	3
ORIG	8	3	4
PAT	282	3	6

Tab. 5. Changes in the number of actants in noun valency frames

5 CONCLUSION

We have presented results of the first automatic comparison of valency frames of interlinked noun and verbal lexical units, included in valency lexicons NomVallex and VALLEX.

Our data shows that the non-systemic valency behavior of Czech deverbal nouns is mostly manifested by non-systemic forms of their actants, most frequently by a prepositional group. The non-systemic forms considerably contribute to a general increase in the number of forms in valency frames of nouns compared to

the number of forms in valency frames of their base verbs. In line with our expectations, the data shows that non-systemic forms are more frequent in valency frames of non-productively derived nouns than in valency frames of productively derived ones.

ACKNOWLEDGMENTS

The research reported in the paper was supported by the Czech Science Foundation under the project 19-16633S, by Charles University Research Centre program No. UNCE/SCI/022 and by the Progres grant Q14. *Krise racionality a moderní myšlení*. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

- [1] Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In Sojka, P. et al., editors, TSD 2014. LNAI 8655, pages 247–256. Springer International Publishing.
- [2] Hajič, J. et al. (2003). PDT-VALLEX: Creating a Largecoverage Valency Lexicon for Treebank Annotation. In Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pages 57–68. Vaxjo University Press.
- [3] Kolářová, V. (2010). Valence deverbativních substantiv v češtině (na materiálu substantiv s dativní valencí). Praha, Karolinum.
- [4] Kolářová, V. (2014a). Special valency behavior of Czech deverbal nouns. In Spevak, O., editor, Noun Valency, pages 19–60, Amsterdam, John Benjamins Publishing Company.
- [5] Kolářová, V. (2014b). Nominalizované struktury se dvěma aktanty ve formě bezpředložkového genitivu. *Naše řeč*, 97(4–5), pages 286–299.
- [6] Kolářová, V., Vernerová, A., and Klímová, J. (2018). Předložková vyjádření adnominálních valenčních doplňků. *Prace Filologiczne*, 72, pages 211–223.
- [7] Kolářová, V., Vernerová, A., Klímová, J., and Kolář, J. (2017). Possible but not probable: A quantitative analysis of valency behaviour of Czech nouns in the Prague Dependency Treebank. *Jazykovedný časopis*, 68(2), pages 208–218.
- [8] Křen, M. et al. (2017). Korpus SYN, verze 6 z 18. 12. 2017. Praha, Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>.
- [9] Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., Žabokrtský, Z., and Barančíková, P. (2018). VALLEX 3.5 – Valenční slovník českých sloves. Praha, Karlova univerzita. Accessible at: <http://ufal.mff.cuni.cz/vallex/3.5/>.
- [10] Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Praha, Academia.
- [11] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.
- [12] Svozilová, N., Prouzová, H., and Jirsová, A. (2005). *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Praha, Academia.
- [13] Uhlířová, L. (2017). Slovosled nominální skupiny. In *Nový encyklopedický slovník češtiny*. Accessible at: <https://www.czechency.org/slovník>.

TOWARDS RECIPROCAL DEVERBAL NOUNS IN CZECH: FROM RECIPROCAL VERBS TO RECIPROCAL NOUNS

VÁCLAVA KETTNEROVÁ – MARKÉTA LOPATKOVÁ

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic

KETTNEROVÁ, Václava – LOPATKOVÁ, Markéta: Towards reciprocal deverbal nouns in Czech: from reciprocal verbs to reciprocal nouns. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 434 – 443.

Abstract: Reciprocal verbs are widely debated in the current linguistics. However, other parts of speech can be characterized by reciprocity as well – in contrast to verbs, their analysis is underdeveloped so far. In this paper, we make an attempt to fill this gap, applying results of the description of Czech reciprocal verbs to nouns derived from these verbs. We show that many aspects characteristic of reciprocal verbs hold for reciprocal nouns as well.

Keywords: reciprocity, deverbal nouns, lexical and syntactic reciprocal nouns

1 INTRODUCTION

Reciprocity, as language means encoding mutuality, has attracted much attention in the current linguistics, esp. from a typological perspective ([1], [2], [3]). Despite representing a rather infrequent language phenomenon [4], reciprocity plays a substantial role in the rule based generation of well-formed structures: its prominent position in this task is given by the fact that reciprocity – similarly as diathesis – brings about changes in the surface syntactic structure, see the analysis of reciprocity in generative linguistics [5] and in the dependency-oriented Meaning-Text Theory [6]. The most thorough description of reciprocity in Czech is provided by works elaborated within the Functional Generative Description ([7], [8], [9], [10]). Besides these works, reciprocity in Czech is discussed esp. in ([11], [12]).

Reciprocity in Czech can characterize verbs (1), nouns (2), adjectives (3), and adverbs (4). In contrast to verbs, the description of reciprocity with other parts of speech is rather at the beginning.

- (1) *Manželé se navzájem rušili ze spaní.*
‘Man and wife disturb each other from sleeping’
- (2) *vzájemná náklonnost Petra a Jany*
‘Peter and Jane’s mutual affection’
- (3) *hrdí na sebe*
‘pride of each other’

- (4) *kolmo na sebe*
'perpendicularly to each other'

In this paper, we provide a pilot study of Czech reciprocal nouns derived from verbs, making use of results of the analyses of Czech reciprocal verbs, esp. ([7], [8]), [9]. For their description, we take over a model of a syntactic operation of reciprocalization elaborated for reciprocal verbs [13]. As a theoretical background, the valency theory of the Functional Generative Description is applied ([14], [15], [10]). Due to the limited range of this paper, we focus on nominal structures of deverbal nouns here, while changes characteristic of employing reciprocal nouns in verbal structures, i.e., in reciprocal light verb constructions are left aside.

The paper is structured as follows. First, we classify Czech reciprocal nouns into two groups, lexical and syntactic reciprocal nouns (Sect. 2). Then we discuss the semantic and deep syntactic changes brought about by reciprocalization in nominal structures of deverbal nouns (Sect. 3). Further, we focus on morphosyntactic changes associated with reciprocalization of these nouns (Sect. 4). In Section 5, we explain the role of reciprocalization with lexical and with syntactic reciprocal nouns. Finally, Section 6 comments the distribution of the information on reciprocalization between lexicon and grammar, as two sides of the language description.

2 LEXICAL VS. SYNTACTIC RECIPROCAL NOUNS

Similarly as reciprocal verbs, reciprocal nouns can be differentiated into lexical and syntactic reciprocal nouns. Lexical reciprocal nouns contains mutuality in their lexical meaning (e.g. *dohoda* 'agreement', *podoba* 'similarity', *přátelství* 'friendship', *rozhovor* 'talk'). These deverbal nouns are typically derived from lexical reciprocal verbs, i.e., from those verbs that bear the semantic trait of mutuality in their lexical meaning [13]. This group includes also all deverbal nouns systematically derived by the derivational morphemes *-ní/-tí* from these verbs (e.g. *diskutování* 'discussing', *chození* 'dating', *oddělení/oddělování* 'isolating', *praní se* 'fighting', *rozlišení/rozlišování* 'distinguishing'), see [16].

Further, mutuality can be expressed also by nouns the meaning of which do not bear the semantic trait of mutuality, which, however, allow some of their semantic participants to enter into reciprocity (e.g. *dar* 'gift', *hrozba* 'threat', *chvála* 'praise', *soucit* 'compassion', *radost* 'joy', *strach* 'fear').¹ We refer to them as to syntactic reciprocal nouns since mutuality is primarily expressed by syntactic means with them (i.e., the syntactic operation of reciprocalization must be applied for expressing mutuality).

¹ The conditions of reciprocalization with verbs is discussed in [7].

3 SEMANTIC AND DEEP SYNTACTIC ASPECTS OF RECIPROCALIZATION

The formal model of reciprocalization in Czech has been proposed in [13]. Despite being designed for reciprocal verbs, this model explains reciprocalization with reciprocal nouns derived from these verbs as well, regardless of their type (Sect. 2).

As with reciprocal verbs, reciprocalization operates on valency frames of reciprocal nouns. Its formal model reflects that a pair of semantic participants,² referring to distinct referents, are symmetrically mapped onto valency complementations involved in reciprocity, and as a consequence, onto surface positions provided by these complementations. The complex mapping of semantic participants has both semantic and morphosyntactic effects (Sect. 4). From the semantic perspective, the reciprocal structure portrays a complex event comprising two propositions expressed in a single structure, see e.g. [17].

For example, with the noun *půjčka* ‘loan’, derived from the verb *půjčit^{pf}/půjčovat^{impf}* ‘to lend’, the semantic participants Agent and Recipient, corresponding to the ACT and ADDR valency complementations, respectively, can enter into reciprocity, see the valency frame of the noun (5)³ and examples (6). Applying the syntactic operation of reciprocalization to the valency frame of this noun leads to the complex mapping of its semantic participants onto the deep and surface syntax, see the scheme in Fig. 1.

(5) *půjčka* ‘loan’: ACT_{2,7,pos,od+2} ADDR_{2,3,pos} PAT₂

(6) *vzájemná půjčka Petra a Pavla / Petrova a Pavlova vzájemná půjčka*
≈ půjčka peněz Petrovi od Pavla a zároveň půjčka peněz Pavlovi od Petra
 ‘Peter and Paul’s loan of money’

≈ ‘Paul’s loan of money to Peter and at the same time Peter’s loan of money to Paul’

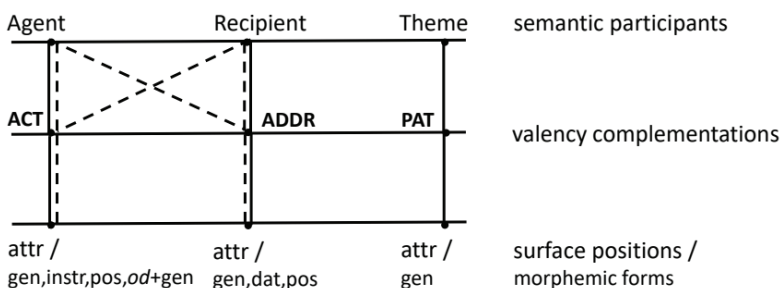


Fig. 1. The scheme of reciprocity of the noun *půjčka* ‘loan’; the solid line displays the mapping in unreciprocal structures, the dashed line depicts it in reciprocal ones.

² Reciprocity can comprise a triplet of participants as well (e.g., *Kolegové se vzájemně představili*. ‘Colleagues introduced each other to each other.’, *vzájemné představení kolegů* ‘a mutual introduction of colleagues to each other’). However, as these cases are extremely rare, we leave them aside here.

³ In valency frames of nouns and verbs, we omit the information on obligatoriness of valency complementations (as it is not relevant for our further explanation).

We can observe that reciprocalization represents the same process with nouns as with verbs. For example, the same scheme, describing relations between the set of semantic participants and the set of valency complementations in Fig. 1, characterizes reciprocalization with the verb *půjčít^{pf}/půjčovat^{impf}* ‘to lend’, see the valency frame of the verb (7) and examples (8). They differ only in changes in surface positions, given by different structural possibilities of verbs and nouns.

(7) *půjčít^{pf}/půjčovat^{impf}* ‘to lend’: ACT₁ ADDR₃ PAT₄

(8) *Petr a Pavel si vzájemně půjčovali peníze.*

≈ *Petr půjčoval peníze Pavlovi a zároveň Pavel půjčoval peníze Petrovi.*

‘Peter and Paul lent money to each other.’

≈ ‘Peter lent money to Paul and at the same time Paul lent money to Peter.’

4 MORPHOSYNTACTIC CHANGES IN RECIPROCAL NOMINAL STRUCTURES

The complex mapping of semantic participants, characteristic of reciprocalization, is reflected in morphosyntactic changes of valency complementations too. Similarly as with verbs, one surface position affected by reciprocalization is pluralized (Sect. 4.1) while the other is either deleted from the surface, or it is filled with the reflexive pronoun, or with the expression *jeden druhý* ‘each other’ (Sect. 4.2); further, reciprocal nouns can be modified by adjectives expressing mutuality (Sect. 4.3).

4.1 The pluralized surface position

The pluralized position is provided by that valency complementation of a noun that corresponds to the pluralized position of its respective base verb; this adverbial position is expressed either as the nominative subject, or as the accusative direct object [13]. As the pluralized position of nouns is obligatorily expressed on the surface, it can be considered to be the more prominent one.

The valency complementation corresponding to the pluralized position has typically morphemic forms resulting from changes of the adverbial nominative or accusative to adnominal forms: nominative typically changes into possessive forms, genitive, instrumental, or the prepositional case *od+Gen* with deverbal nouns and accusative turns into possessive forms and genitive with these nouns [18].

For example, with the noun *hádká* ‘quarrel’, reciprocalization involves ACT and ADDR (corresponding to the semantic participants *Communicator_1* and *Communicator_2*, respectively), each providing an attribute position, see the valency frame (9). From these attribute positions, the position given by ACT of the noun is the more prominent one as this ACT corresponds to the nominative ACT of the base verb *hádat se^{impf}* ‘to quarrel’, compare frame (9) with the valency frame of the verb

(11). In reciprocal nominal structures, this attribute position is pluralized. With nouns (similarly as with verbs (12a-c)), it can be pluralized by coordination (10a), by a plural noun (10b), or by a collective noun (10c). As a specific morphemic form of the pluralized complementation, the prepositional case *mezi*+Instr expands in reciprocal nominal structures, see examples (10d-e).

- (9) *hádká* ‘quarrel’: ACT_{2,pos} ADDR_{s+7} PAT_{o+4,dcc}
- (10) a. *hádká Petra_{ACT} a Jany_{ACT}*
 ‘Peter_{ACT} and Jane’s_{ACT} quarrel’
 b. *hádká kolegů_{ACT}*
 ‘quarrel of colleagues_{ACT}’
 c. *hádká výboru_{ACT}*
 ‘quarrel of the committee_{ACT}’
 d. *hádká mezi Petrem_{ACT} a Janou_{ACT}*
 ‘Peter_{ACT} and Jane’s_{ACT} quarrel’
 e. *hádká mezi kolegy_{ACT}*
 ‘quarrel of colleagues_{ACT}’
- (11) *hádat se* ‘to quarrel’: ACT₁ ADDR_{s+7} PAT_{o+4,dcc}
- (12) a. *Petr_{ACT} a Jana_{ACT} se hádali.*
 ‘Peter_{ACT} and Jane_{ACT} were quarrelling.’
 b. *Kolegové_{ACT} se hádali.*
 ‘Colleagues_{ACT} were quarrelling.’
 c. *Výbor_{ACT} se hádal.*
 ‘The committee_{ACT} was quarreling.’

Further, with the noun *izolace* ‘isolation’ (and its base verb *izolovat^{biasp}* ‘to isolate’), their semantic participants Part₁ and Part₂, mapped onto the valency complementations PAT and ORIG, respectively, see valency frame (13), can be reciprocalized. From the surface positions given by these nominal valency complementations, the attribute position provided by PAT is the more prominent one, hence pluralized (14), as PAT is in correspondence with the accusative PAT of the base verb *izolovat^{biasp}* ‘to isolate’, expressed as the direct object (15), see also example (16).

- (13) *izolace* ‘isolation’: ACT_{2,pos} PAT_{2,pos} ORIG_{od+2}
- (14) *vzájemná izolace členů_{PAT} domácnosti*
 ‘household members’_{PAT} isolation from each other’
- (15) *izolovat^{biasp}* ‘to isolate’: ACT₁ PAT₄ ORIG_{od+2,z+2}
- (16) *Technologie členy_{PAT} domácnosti vzájemně izolují.*
 ‘Technologies isolate household members’_{PAT} from each other.’

4.2 The less prominent surface position

With reciprocal nouns, the less prominent position involved in reciprocalization can remain unexpressed on the surface. If it is present, it can be optionally occupied either by the reflexive pronoun, or by the expression *jeden druhý* ‘each other’, both coreferring with the expression in the more prominent position. These possibilities are conditioned by morphemic forms of the valency complementation providing this position.

As with reciprocal verbs, if this complementation has the form of the prepositional case *s+Instr*, it is systematically deleted from the surface. The prepositional group *s+Instr* is the most frequent form of the valency complementation providing the less prominent surface position with lexical reciprocal nouns (see Sect. 2). For example, ADDR in the valency frame of the noun *dohoda* ‘agreement’ (17) is subject to reciprocalization with ACT. While ACT is pluralized, ADDR is omitted from the surface, see example (18).

(17) *dohoda* ‘agreement’: ACT_{2,pos} ADDR_{s+7} PAT_{na+6,o+6,inf,dec}

(18) *dohoda obchodníků_{ACT} na ceně kávy*
‘traders_{ACT} agreement on the price of coffee’

A complementation expressed by a simple case or a prepositional case other than *s+Inst* can be filled by the long form of the reflexive pronoun⁴ or by the expression *jeden druhý* ‘each other’, both coreferring with the more prominent position. In contrast to reciprocal verbs, however, the surface realization of this valency complementation of reciprocal nouns is only optional. For example, with the noun *podpora* ‘support’, see valency frame (19), ACT and PAT can be reciprocalized. While ACT is pluralized, PAT can be deleted from the surface (20a), or – if it is present on the surface – it is occupied by the reflexive pronoun in its respective long form (20b), or by the expression *jeden druhý* ‘each other’, from which *jeden* has the form of genitive, while *druhý* is in the respective form prescribed by PAT (excluding genitive or possessive forms) (20c).

(19) *podpora* ‘support’: ACT_{2,pos,od+2} PAT_{2,3,pos} EFF_{v+6}

(20) a. *Petrova_{ACT} a Pavlova_{ACT} vzájemná podpora*
b. *Petrova_{ACT} a Pavlova_{ACT} vzájemná podpora sobě_{PAT}*
c. *Petrova_{ACT} a Pavlova_{ACT} vzájemná podpora (jednoho druhému)_{PAT}*
‘Peter_{ACT} and Paul’s_{ACT} support for each other’

⁴ Let us emphasize that there is a difference between reciprocal nouns and reciprocal verbs. With reciprocal verbs, besides the long form of the reflexive pronoun, the clitic forms *se/si* are available in the dative or accusative case, representing positional variants of the pronoun [10]. With reciprocal nouns, only the long forms of the reflexive pronoun can occur [19].

4.3 Modifying adjectives

A reciprocal noun can be modified by the adjectives *vzájemný* or *společný* ‘mutual’. The latter one is, however, polysemous: besides the meaning “mutual” (21), it also expresses the meanings “collective, joint” (22) and “common” (23). In the meaning “mutual”, the adjective seems to be restricted to lexical reciprocal nouns. For example, while with the lexical reciprocal noun *shoda* ‘agreement’, the modifying adjective has the meaning “mutual” (21), with the syntactic reciprocal noun *radost* ‘joy’, only the meaning “common” is available (24).

- (21) *společná shoda mezi nájemníky*
‘mutual agreement between tenants’
- (22) *společný koncert Hradištanu a sboru Stojanova gymnázia*
‘a joint concert of Hradišťan and the choir of Stojanov’s grammar school’
- (23) *společný majetek*
‘common property’
- (24) *společná radost týmu z výhry*
‘common joy of the win’

As for the function of these adjectives, if the less prominent position is expressed on the surface (Sect. 4.2), the adjectives stress the meaning of mutuality (20b-c). However, if the less prominent position is not expressed on the surface, the adjective is – besides the pluralization of the more prominent position – the only marker of mutuality, removing possible ambiguity between reciprocal and unreciprocal interpretation (20a), (25a) and (27). Without the respective adjectives, these structures can be interpreted as either reciprocal, or unreciprocal with an elided valency complementation. For example, (25b) can have either the reciprocal interpretation, or the unreciprocal one with PAT of the noun *sympatie* ‘sympathy’ unexpressed on the surface, see the valency frame (26).

- (25) a. *naše_{ACT} vzájemné sympatie*
‘our mutual sympathy’
b. *naše_{ACT} sympatie*
‘our sympathy’
 \approx *naše_{ACT} vzájemné sympatie* vs. *naše_{ACT} sympatie k ostatním_{PAT}*
 \approx ‘our_{ACT} mutual sympathy vs. our_{ACT} sympathy for others_{PAT}’
- (26) *sympatie* ‘sympathy’: ACT_{2,pos} PAT_{3,k+3,pro+4,s+7,vůči+3}
- (27) *společná dohoda EU a USA*
‘mutual agreement of EU and USA’

5 ROLE OF RECIPROCALIZATION WITH LEXICAL VS. SYNTACTIC RECIPROCAL NOUNS

With lexical and syntactic reciprocal nouns, reciprocalization plays different roles. With syntactic reciprocal nouns, it is a necessary condition for expressing mutuality. However, with lexical reciprocal nouns, which already bear mutuality in their lexical meaning, its role is different: it allows to make the semantic participants involved in reciprocity equal with respect to their participation (in terms of *figure* and *ground*) in the event expressed by a noun, see esp. [20] and [13], stressing that the mapping of participants onto valency positions is not random, compare (28a-b).

For example, the noun *rozchod* ‘split-up’ is characterized by two semantic participants, Part_1 and Part_2. As the noun contains mutuality in its lexical meaning, it expresses a mutual event even if its semantic participants are not reciprocalized. In this case, the participant in the more prominent position can be interpreted as more active in the event than the other expressed in the less prominent position; compare examples (30a) with (30b) in which each time a different participant, *hráč* ‘player’ or *trenér* ‘trainer’, occupies the more prominent position provided by ACT of the noun, see its valency frame (29). However, it does not change the fact that they both are involved in a mutual event. In contrast, when these participants are subject to reciprocalization, their participation in the event is presented as equal (30c).

(28) a. *Jak Petr rostl, byla jeho podoba s otcem stále zřetelnější.*

‘As Peter was growing up, his similarity with his father was more and more visible.’

b. *?Jak Petr rostl, byla otcova podoba s ním stále zřetelnější.*

‘As Peter was growing, father’s similarity with him was more and more visible.’

(29) *rozchod* ‘split-up’: ACT_{2,POS} PAT_{s+7}

(30) a. *hráčův_{ACT} rozchod s trenérem_{PAT}*

‘the player’s split-up with the trainer’

b. *trenérův_{ACT} rozchod s hráčem_{PAT}*

‘the trainer’s split-up with the player’

c. *rozchod hráče_{ACT} a trenéra_{ACT}*

‘split-up of the player_{ACT} and the trainer_{ACT}’

6 RECIPROCALIZATION OF NOUNS IN THE LANGUAGE DESCRIPTION

Formal theories attempting for generation of well-formed structures carefully distribute the information between lexicon and grammar; the former stores those individual properties of language units that are not predictable from their semantic or

morphosyntactic features while the latter captures their recurrent patterns which can be described in the form of rules.

As for reciprocity, three types of information should be provided by *the lexicon* as it is conditioned by semantic and partially by pragmatic factors which are not reflected in the language structure:

- the information on the type of a noun (lexical or syntactic reciprocal noun),
- its valency structure, and
- the information on individual pairs of the valency complementations that can be subject to reciprocalization.

In contrast, surface syntactic changes follow from morphemic forms of the valency complementations involved in reciprocity – they are regular enough to be captured by formal rules stored in *the grammar*. In addition to morphosyntactic changes of valency complementations, these rules should describe their lexical expression (Sect. 4.1 and 4.2) and the role of adjectives (Sect. 4.3).

7 CONCLUSION

In this paper, we have explained principles underlying generation of well-formed reciprocal structures of deverbal nouns that cover their semantic, deep as well as surface syntactic structures. We show that valency frames of both lexical and syntactic reciprocal nouns must be stored in the lexical component of the language description, including the information on those valency complementations which can be reciprocalized. Then detailed rules describing changes in their nominal structures caused by reciprocalization and closely cooperating with rules governing surface formation of unreciprocal structures must be provided by the grammar component.

ACKNOWLEDGEMENTS

The research reported in this paper has been supported by the GAČR grant No. 18-03984S, *Between Reciprocity and Reflexivity: The Case of Czech Reciprocal Constructions*. This work has been using language resources distributed by the LINDAT/CLARIN project of the MŠMT ČR, No. LM2015071.

References

- [1] Nedjalkov, V. P. (ed.) (2007). *Reciprocal Constructions*. Amsterdam/Philadelphia, John Benjamins.
- [2] Evans, N., Gaby, A., Levinson, S. C., and Majid, A. (eds.) (2011). *Reciprocals and Semantic Typology*. Amsterdam/Philadelphia, John Benjamins.

- [3] König, E., and Gast, V. (eds.) (2008). *Reciprocals and reflexives: cross-linguistic and theoretical explorations*. Berlin/New York, Mouton de Gruyter.
- [4] Maslova, E. (2008). Reflexive encoding of reciprocity. In E. König, V. Gast (eds.), *Reciprocals and reflexives: cross-linguistic and theoretical explorations*, pages 227–257, Berlin/New York, Mouton de Gruyter.
- [5] Reinhart, T., and Siloni, T. (2005). The Lexicon-Syntax Parameter: Reflexivization and Other Arity Operations. *Linguistic Inquiry*, 36(3), pages 389–436.
- [6] Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. Albany, State University of New York Press.
- [7] Panevová, J. (1999). Česká reciproční zájmena a slovesná valence. *Slovo a slovesnost*, 60(4), pages 269–275.
- [8] Panevová, J. (2007). Znovu o reciprocitě. *Slovo a slovesnost*, 68(2), pages 91–100.
- [9] Panevová, J., and Mikulová, M. (2007). On reciprocity. In *The Prague Bulletin of Mathematical Linguistics*, 87, pages 27–40.
- [10] Panevová, J., Hajičová, E., Kettnerová, V., Lopatková, M., Mikulová, M., and Ševčíková, M. (2014). *Mluvnice současné češtiny 2, Syntax na základě anotovaného korpusu*. Praha, Karolinum.
- [11] Medová, L. (2009). *Reflexive Clitics in the Slavic and Romance Languages. A Comparative View from an Antipassive Perspective*. PhD thesis, Princeton, Princeton University.
- [12] Grepl, M., and Karlík, P. (1999). *Skladba češtiny*. Olomouc, Votobia.
- [13] Kettnerová, V., and Lopatková, M. (2018). Lexicographic Potential of Syntactic Properties of Verbs: The Case of Reciprocity in Czech. In XVIII EURALEX International Congress, *Lexicography in Global Contexts*, pages 685–698, Ljubljana, Ljubljana University Press.
- [14] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel.
- [15] Panevová, J. (1994). Valency Frames and the Meaning of the Sentence. In P. A. Luelsdorff (ed.), *The Prague School of Structural and Functional Linguistics*, pages 223–243, Amsterdam/Philadelphia, John Benjamins.
- [16] Dvořák, V. (2017). Verbální substantivum. In P. Karlík, M. Nekula, J. Pleskalová (eds.), *Nový encyklopedický slovník češtiny*. Praha, Nakladatelství Lidové noviny.
- [17] Evans, N., Gaby, A., and Nordlinger, R. (2007). Valency mismatches and the coding of reciprocity in Australian languages. In *Linguistic Typology*, 11, pages 541–597.
- [18] Kolářová, V. (2010). *Valence deverbativních substantiv v češtině (na materiálu substantiv s dativní valencí)*. Praha, Karolinum.
- [19] Dvořák, V. (2017). Dějové substantivum. In P. Karlík, M. Nekula, and J. Pleskalová (eds.), *Nový encyklopedický slovník češtiny*. Praha, Nakladatelství Lidové noviny.
- [20] Gleitman, L. R., Gleitman, H., Miller, C., and Ostrin, R. (1996). Similar, and similar concepts. *Cognition*, 58, pages 321–376.

PROCESSING OF DERIVATIONAL FEATURES FOR (SEMI)AUTOMATIC
CREATION OF DICTIONARY DEFINITIONS IN THE USER INTERFACE
(CZEDD) FOR LEARNING CZECH AS A SECOND LANGUAGE:
SUFFIX *-tel* AND *-ista*

ERIK CITTERBERG – ADRIANA VÁLKOVÁ
Faculty of Arts, Masaryk University, Brno, Czech Republic

CITTERBERG, Erik – VÁLKOVÁ, Adriana: Processing of derivational features for (semi)automatic creation of dictionary definitions in the user interface (CZEDD) for learning Czech as a second language: suffix *-tel* and *-ista*. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 444 – 455.

Abstract: This work-in-progress paper presents the tool CZEDD which enables the user to learn how to predict the meaning of words. The CZEDD consists of (semi) automatic definitions for derived words because a lot of these words have predictable lexical meaning. The tool will be intended for foreigners who learn the Czech language and it could be useful as a dictionary and/or translator in which the definitions based on the word's structure are stored. Two detailed case examples (the suffix *-tel*, and the suffix *-ista*) illustrate the approach.

Keywords: derivational morphology, Czech for foreigners, suffixes, lexical meaning, structural meaning, dictionary

1 INTRODUCTION

The Czech language is a Slavic language with richly developed morphology. Foreigners who learn the Czech language are confronted with it from the beginning. Next to inflectional morphology which studies how the forms of lexemes are created by morphemes (e.g. from the noun *pes*¹ 'a dog' forms *psovi* 'to a dog', *psi* 'dogs' and e.g. from the verb *mít* 'to have' forms *měli jsme* 'we had', *máš* 'you have' (form in singular)) it is necessary for complete knowledge of Czech to know how some of these morphemes build other different lexemes, not just their forms (e.g. noun *knihovna* 'a library' derived from noun *kniha* 'a book'). This is a part of the derivational morphology and the word-formation in general. However, there are many studies, from codification grammars to online tools, which handle inflectional morphology for

¹ In this text, we translated to English just those derived words which we have found in English dictionary Glosbe (see [6]), we did not try to create the neologism. When we don't find the shape in the dictionary, we write the verb which is semantically related to the agent name (e.g. *vychovatel* which means 'one who nurtures especially children' so we write it like this: *vychovatel* ← *vychovat* 'to nurture').

Czech language and which show how to work with it in the teaching and/or learning of Czech language for foreigners, the textbooks included. The information of meaning, so-called structural meaning, of these word-formation morphemes are mostly found only in the specialized books about the word-formation (see e.g. [4], [5], [12]), in newer Czech grammars (see e.g. [2], [16]) and in online dictionaries (see e.g. [7], [14]). In this field of study are tools which show the derivational relations, for example, the DeriNet (see [13], [18]), Deriv or Derivancze (see [11]) and Morfio (see [3]) which show more of formal relations than semantic relations.

At this moment, there is no study which focuses on the predictability of lexical meaning of Czech derived words, especially for how much concrete suffixes are predictable or not, and its use for teaching.

In the following sections, we present the tool CZEDD (Czech electronic derivational dictionary) and processing of suffixes which are used in this application. CZEDD provides an option of working with word-formation as a part of grammar which may play a key role in learning (acquisition) the Czech language by the clearly determined meaning of affixes.

2 MOTIVATION

Native speakers can predict the meaning of words they have never heard before or they can subconsciously create a “new” word for the specific context using word-formation morphemes. This is because they know the meaning of a suffix analogically based on already known words, e.g. *publikovatel* ‘a person who publishes something’ ← *publikovat* ‘to publish’ with the analogy to words ending with *-tel*: *učitel* ‘a teacher’, *cestovatel* ‘a traveller’, etc.). The lexical meaning is a complex of the historical, social and other influences, and for its complete understanding the structural meaning is insufficient. On the other hand, a rough estimate of unknown word meaning could prove to be of value for the fluency of communication. We think the foreigners who will periodically use this app might become more aware of the structure of words. Moreover, the morphemes with word-formation function carry specific semantic information (e.g. *knih-ovna* ‘a library’): *-ovna* is a name for a place) and more specific grammatical information of a part of speech and its properties (e.g. *knihovna*: noun, feminine, noun paradigm *žena* ‘a woman’). However, foreigners find it difficult to recognize the paradigm of nouns.

Students who learn Czech as a second language speak at least one other language (their mother tongue). With the knowledge of suffixes similar to that of native speakers, the foreigners should be able to understand the approximate meaning of an internationalism which is adapted to the Czech language by suffixes. Moreover, for Slavic students with similar mother tongue to Czech, it is possible to expect a quick understanding of these adaptations and such students should be able to acquire the derivation rules intuitively.

3 PROCESSING OF AFFIXES FOR CREATING DEFINITIONS IN THE CZEDD

3.1 Processing affixes

We focus on the most frequent and productive suffixes used for deriving nouns. We have processed nouns derived by adding monofunctional suffixes *-tel* and *-ista*. We have found the possible meaning of nouns derived by these suffixes according to the information about them in the online dictionary and from the specialized books, mentioned in the Introduction. We have tested this meaning on data of written Czech corpora SYNv6 (SYNv7) (see [8]) which enables us to find and work with the most frequent of them. The queries are specified in Corpus Query Language (CQL). For the words from the corpus, we have compared their structural meaning and the meaning found in the online dictionaries in Lexiko (see [17]) and evaluated the correspondence between them in percent. For the words for which the lexical and structural meaning are in acceptable correspondence, we are trying to find the most general definition.

3.2 Evaluation data from corpus

For suffix *-tel* we have processed 1 129 lemmas, i.e. all lemmas for the corpora query [tag="N.*"& lemma="*.tel"]² and we have found out that for words with lower frequency the structural meaning corresponds to their lexical meaning. Therefore, for the next suffix, suffix *-ista* specified by query [tag="N.I.*"& lemma="*.ista"], we have processed only the 200 most frequent word forms. The word-formation research in corpus is described in e.g. [9], [10].

Suffix *-tel*

Out of all found lemmas, those which are not derived (e.g. *epitel* ‘an epithelium’) have been manually removed and 1 129 lemmas have been chosen to be further processed. This number includes the unprefixated and prefixated forms (prefix *do-*, *na-*, *o-*, *ob-*, *od-*, *po-*, *pod-*, *pro-*, *pře-*, *před-*, *při-*, *roz-*, *s-*, *u-*, *v-*, *vy-*, *vz-*, *z-*, *za-*)³.

General and simplified structural meaning found in books: suffix *-tel* means, in general, an agent of some action with semantic features [+Person], [+Masculine], [+Animate] [+Agents]. This action is represented by the verb from which the noun is derived.

56,07% of 1 129 lemmas were found in the dictionaries. The lexical meaning does not correspond to the structural meaning in 3,01%, for e.g. *nakladatel* ‘a publisher’, *věřitel* ‘a creditor’, *buditel* ‘a revivalist’ and we have found there is a group of the impersonal nouns:

² We use regular expressions occurring in the SYN corpus. “.*” is interpreted as any character repeated from zero to potential infinity.

³ For this corpus research, we have worked with prefixated nouns derived from prefixated verbs, i.e. we did not process the nouns with prefix *nad-* (e.g. *nadučitel* (freely translated ‘more than teacher’) and prefix *pod-* in words like *podučitel* (freely translated ‘less than teacher’).

- [-Person], [+Masculine], [-Animate], [+Agens]: typically, mathematics and business names *jmenovatel* ‘a denominator’, *dělitel* ‘a divisor’, *čítatel* ‘a numerator’, *násobitel* ‘a multiplier’, *menšitel* ‘a subtrahend’, *úročitel* ‘an interest rate’, *odmocnitel* ‘a square root’, *umořovatel* ‘one payment in a series of installments’, *odůročitel* ‘a discount rate’, *součinitel* ‘a coefficient’
- 3 words which could be [+Person] [+Masculine] [+Animate], [+Agens] or [-Person], [+Masculine], [-Animate], [+Agens]: *činitel* ‘a factor/an agent’, *ukazatel* ‘a pointer’, *zaměstnavatel* ‘an employer’⁴.

We have found that the lexical meaning is more specific in 4,34% (e.g. *spisovatel* ‘a writer’) and concurrently we have not found the nouns for which the lexical meaning is more general than their structural meaning.

The nouns derived by the suffix *-tel* from perfect verbs (e.g. *vydražitel* ‘an auctioneer’ ← *vydražit* ‘to auction off’) expresses the action that has been done or will be done. The nouns derived from imperfect verbs (*vyšetřovatel* ‘an investigator’ ← *vyšetřovat* ‘to investigate’) means the action is in progress.

We have found that 74,3% nouns are derived from imperfect verbs and 25,7% nouns from 1 129 lemmas are derived from perfect verbs. But in 5,31% we have found

a. the nouns derived from perfect verbs behave like nouns derived from imperfect verbs:

- names for professions (e.g. *vychovatel* ← *vychovat* ‘to nurture’, *zastupitel* ‘a representative’ ← *zastoupit* ‘to deputize’)
- name for a person for whom this action is typical of (*zastavitel* ← *zastavit* ‘to pawn’, *chovatel* ‘an animal keeper’ ← *chovat* ‘to keep’) but not typical as a job

and b. nouns derived from imperfect verbs, but behave like the nouns derived from perfect verbs:

- e.g. *pachatel* ‘an offender’ derived from imperfect verb *páchat* ‘to offend’ but with the definition for the perfect verb *pachatel = ten, kdo spáchal* ‘one who committed a crime’, *zakladatel* ‘a founder’ is *ten, kdo něco založil* ‘one who founded an organization’, *zastupitel* ‘a representative’ is *ten, kdo zastupuje* ‘one who represents’.

Most of the nouns are derived from verbs of III–V⁵ verbal classes, though we have found two exceptions: *přistihitel* ‘one who caught an offender in an act’ derived

⁴ The target of the next study will be finding the context of words which can be animate and inanimate as well and we want to find which is the predominant interpretation. We could not use the tags of the SYN corpus tagset to recognize it, because of inaccuracies of the disambiguation.

⁵ According to traditional division of Czech verbs based on their forms in present tense into five verbal classes.

from verb *přistihnout* ‘to catch’ belonging to the II verbal class; and noun *přemožitel* ‘one who overcame someone or something’ derived from verb *přemoci* ‘to overcome’ belonging to the I verbal class.

Suffix *-ista*

General (and simplified) structural meaning: name for a person with semantic features [+Person], [+Masculine] and [+Animate].

We have found that in the group of the 200 most frequent lemmas, it is not possible to predict the lexical meaning in 50,5% of the lemmas, while the structural meaning can be applied in 49,5% of the lemmas:

- 27,5% nouns derived from nouns ending with *-ismus* (e.g. *fašista* ‘a fascist’ ← *fašismus* ‘a fascism’)
- 9% names for instrumental players (e.g. *kytarista* ‘a guitarist’ ← *kytara* ‘a guitar’)
- 8,5% names for sports players (e.g. *fotbalista* ‘a footballer’ ← *fotbal* ‘football’)
- 4,5% for nouns derived from words *-istika* (e.g. *cyklista* ‘a cyclist’ ← *cyklistika* ‘cycling’)

3.3 Generation of definitions

We have created definitions for nouns derived by adding the suffix *-tel* depending on verbal aspect. We can distinguish between the perfect and imperfect verbs thanks to the DeriNet. Definitions have been created by specifying the endings and according to the existing or not existing prefix for both suffixes, i.e. *-tel* and *-ista*.

First step – verbal aspect recognition

At first, we focused on verbal form without the prefix⁶ in two previous steps

- e.g. *zpracovatel* ‘a processor’ ← *zpracovat* ‘to process’ ← *pracovat* ‘to work’
definition *ten, kdo zpracoval nebo zpracuje* ‘one who processed or processes’

Also, based on the existence of imperfective verbs in the scope of two previous derivational steps, we have found the nouns derived from secondary imperfective forms:

- e.g. *dotazovatel* ← *dotazovat* ← *dotázat* ← *tázat* ‘to ask’
definition *ten, kdo dotazuje* ‘one who asks’.

Second step – creating the definition

Suffix *-tel*

Figure 1 shows the steps used for generating definitions. There are a few exceptions: *chovatel* ‘an animal keeper’, *klovatel* ← *klovat* ‘to peck’, *snovatel* ← *snovat* ‘to weave’, *plovatel* ← *plovat* ‘to float’, *kovatel* ← *kovat* ‘to smith’, which are individually specified.

⁶ Prefix *ne-* is not computed, because it does not change the verbal aspect.

[^h ch]ovatel	prefix	NO	„ten, kdo . ^u je“		
		YES	the string contains a verb . ^u it or . ^u nout or [aá]t?	YES	„ten, kdo . ^u je“
				NO	„ten, kdo . ^o val or . ^u je“
[^o]iíyáá]vatel			„ten, kdo . ⁱ yá]vá“		
[^o]ěvatel			„ten, kdo . ⁱ vá“		
-itel	prefix	NO	the string contains a verb . ^u .it?	„ten, kdo . ^u .í“	
			the string contains a verb . ^u [^u].[eě]t?	„ten, kdo . ⁱ “	
			the string contains a verb . ^u .ovat a . ^u .it?	„ten, kdo . ^o u.il nebo . ^o u.í“	
		YES	the string contains a verb . ^o u.it?	„ten, kdo . ^o u.il nebo . ^o u.í“	
			the string contains a verb . ^o ovat & not existing the verbal form . ^u it?	„ten, kdo . ^o val“	
			the string contains a verb . ^u [eě]t?	„ten, kdo . ^u [eě]l nebo . ⁱ “	
			the string contains a verb . ^u [^{ou}]. ^u it?	„ten, kdo . ⁱ l nebo . ⁱ “	
-[^z ^b]atel	prefix	NO	„ten, kdo . ^a “		
		YES	„ten, kdo . ^a l nebo . ^a “		
-zatel	prefix	NO	„ten, kdo . ^z e“		
		YES	„ten, kdo . ^z al nebo . ^z e“		
-batel	prefix	NO	„ten, kdo . ^b á (. ^b e)“		
		YES	„ten, kdo . ^b al nebo . ^b á (. ^b e)“		
-p[ií]satel	prefix	NO	„ten, kdo piše“		

Fig.1. Rules for definition generation for nouns *-tel*

Suffix *-ista*:

We have created a definition for nouns derived from nouns ending with *-ismus* (e.g. *komunista* ‘a communist’ ← *komunismus* ‘communism’ and for nouns derived from nouns ending with *-istika* (e.g. *cyklista* ‘a cyclist’ ← *cyklistika* ‘cycling’). We have also created a definition for foreign adapted words with meaning “name for sports players”, but they are derived from the base word, not as in two previous groups. Definitions are created according to base word endings (e. g. *hokejista* ‘a hockey player’ ← *hokej* ‘hockey’ and *fotbalista* ‘a footballer’ ← *fotbal* ‘football’):

- noun derived from nouns *-ismus*: definition *stoupenec* [*.*ismu*] (‘a follower of [*.**]’)
- noun derived from nouns *-istika*: definition *ten, kdo se zabývá* [*.*istikou*] (‘one who is an enthusiast of [*.**]’)
- adapted words of foreign origin ending with *-ej*: *hráč* [*.*eje*] and *-al*: *hráč* [*.*alu*] (‘a player of [*.**]’)

We have found the suitable correspondence in meaning and word structure for words with meaning “instrumental players” but we have not found the way how to write a rule which will apply to most of these words.

4 TECHNICAL REALIZATION

4.1 Technical realization

The CZEDD could be considered both a conventional dictionary with automatically generated definitions of words, organized according to predefined typology of derivation, and a user interface built for interaction with the DeriNet with extended functionality of implementing Majka [15] and Ajka [1] morphological analyzers.

As a wrapper for all technologies, the Flask Python framework has been used. The simplest form of the CZEDD is pregenerated database with derivational and morphological information for chosen words contained within the Derinet network.

To interact with the database a web application has been created to connect the CZEDD database, the Derinet and Ajka. This web application serves as a user-friendly interface for searching within the database.

The user can enter a word or a text as an input which is then checked against the database and the DeriNet. Additional information is then provided from Ajka.

The database was generated from the DeriNet with series of filters, mainly in the form of regular expressions, which enabled us to find words for which we have sufficient rules to create a suitable definition as well as provide additional information about them. Results of this process was a list of words divided into types, which were further analyzed with rules defined for specific derivational types and then saved into a table in database which can then be queried by users via the CZEDD web application.

Lemmatization and morphological analysis has been done with Majka and for missing words, a function that was checking Ajka api was implemented.

In case the user searches for an unknown word which is not included neither in our CZEDD database, nor in the DeriNet network, it is saved for future review if labeled as one of processed derivational types.

The concept of processing each word was based on creating class objects in Python for words.

As a base, class 'Word' has been created. In the first step based on word endings, an internal pseudo-derivational type has been determined. Then the word has been tagged by Majka and in case it was not found within its data an http request to Ajka has been submitted to retrieve a morphological tagging from there. Both morphological analyzers use tag format developed at Masaryk University. Tag enabled us to determine several attributes: lemma, gender, number, paradigm, part of speech.

The DeriNet was then queried for base word and retrieval of derivational branch up to second verb. This also enabled us to check if a prefix could be identified. English translation of each word within the derivational branch has been extracted by sending an http request to Glosbe API [6].

In case a word has been identified to be within our derivational typology, a class based on Word class has been created. This class was named TypeWord. Additional attributes include definitions in Czech and English languages created by applying rule based substitutions based on their prefix and pseudo-derivational type attribute.

For text input containing multiple words a Text class has been created which is an object of Word and TypeWord objects with additional dictionary attributes for storing original unprocessed words with their lemma as a value.

All this information has been provided for all words within the DeriNet network and the CZEDD database has been then generated. The web application serves as a user interface for this database as well as a searching tool for words within the DeriNet network.

4.2 DeriNet – Derivational network

The DeriNet is a lexical network, which comprises core word-formation relations. The network is currently limited to derivational relations. The network has been extracted from an existing corpus of contemporary Czech and semi-automatically generated using existing data resources (corpora and lexical resources).

Generated candidate pairs of a derived word and its base word were checked manually before creating an edge in the network, unless they came from a highly reliable resource.

The relations between derived words and their base words are modeled as an oriented graph. Nodes of the graph correspond to lexemes. Edges represent derivational steps between lexemes. The orientation of edges reflects the word-

formative process: the edge points from a base lexeme to a derived lexeme. Each lexeme can have at most one base lexeme.

The DeriNet is publicly available on the Internet at <http://ufal.mff.cuni.cz/derinet>. It can be used under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License (CC-BY-NC-SA). The data is available in a simple line-oriented format as well as in a self-documenting XML-based format.

4.3 Majka – Morphological tool

For lemmatization and morphological analysis in the CZEDD web application and in the CZEDD database generation two morphological analyzers were used. These tools contain different data. When possible, Majka was used due to much higher speed and Ajka was used for words not contained within Majka.

It has several functionalities including lemmatization, morphological tagging as well as generating word forms for given lemma based on given tag.

4.4 Glosbe – Online dictionary

There are many commercial solutions available for bilingual Czech to English dictionary. However, in case of open source solutions for simple Czech to English dictionary, the only on-going project that we know of is the Glosbe online dictionary.

Glosbe is a simple multilingual online dictionary that runs as a community project managed by a small development team based in Poland. Its purpose is to create an extensive polylingual general purpose dictionary. Among other things, it contains examples of usage for words taken from several sources as well as general definitions for certain suffixes.

Data included in the Glosbe dictionary are under various licenses: CC-BY-SA, FDL and custom license. Data source is always indicated next to data if it is needed due to the license.

5 CZEDD – CZECH ELECTRONIC DERIVATIONAL DICTIONARY

5.1 What is the CZEDD?

The CZEDD is a user interface which enables to understand the principle of semantic and formal connections between Czech derived words. There are two basic functions – 1. insert word and 2. insert text. The CZEDD works like a special bilingual dictionary with definitions based on word structure (see the CZEDD as a dictionary). The function “**Insert text**” provides the processing of text in which derived words are colour marked and for the colour marked words the same process is applied as in the first function.

The CZEDD can be used in the teaching of Czech word-formation, as an e-learning tool, especially for more advanced students. A different way how to use the CZEDD is for translations from Czech to English (especially for beginners) and as a translator from Czech to Czech (for advanced students).

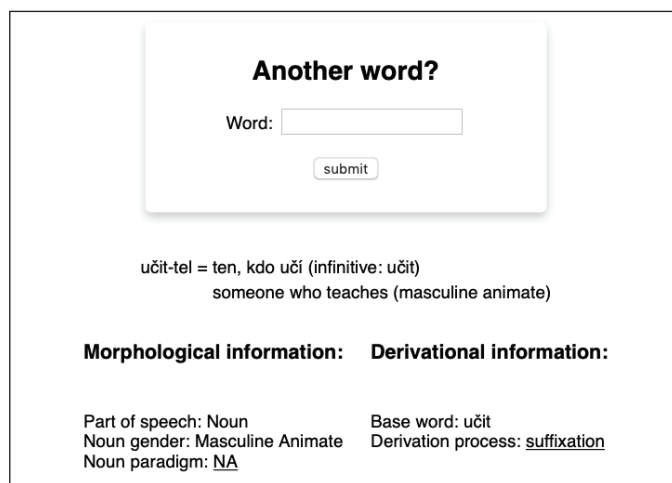
5.2 CZEDD as a translator

CZEDD can be used as a translator, especially for the newly created words, neologisms or the words created just for the concrete situation (context). Most translators cannot work with these types of words. CZEDD provides bilingual translator interface: from Czech to English.

5.3 CZEDD as a dictionary

In CZEDD you can find a grammatical information about searched word (see Figure 2). This dictionary is available through the “**Insert word**” function.

1. definition
2. part of speech
3. noun gender
4. noun paradigm
5. base word
6. derivation process



The screenshot shows a web interface for the CZEDD dictionary. At the top, there is a search box with the text "Another word?" and a "submit" button. Below the search box, the word "učit-tel" is displayed with its definition: "ten, kdo učí (infinitive: učít)" and "someone who teaches (masculine animate)". Below this, there are two columns of information: "Morphological information" and "Derivational information". The morphological information includes "Part of speech: Noun", "Noun gender: Masculine Animate", and "Noun paradigm: NA". The derivational information includes "Base word: učít" and "Derivation process: suffixation".

Fig. 2. Processed word

6 DISCUSSION AND FUTURE WORK

We have tried to create definitions for derived words from their structural meaning. We have processed 1 129 nouns derived by suffix *-tel* and the first 200 most frequent nouns derived by suffix *-ista*. General definition was created for most nouns derived by *-tel* according to the verbal aspect: *ten, kdo [dělá] nebo [udělal/udělá]* (‘one who [does] or [has done/will do]’). However, it was necessary

to separate two groups of nouns derived by *-ista* depending on their base word: a. *-ismus* (*komunista* ‘a communis’ ← *komunismus* ‘communism’): *stoupenec* [.*ismu] (‘a follower of ...’), b. *-istika* (*cyklista* ‘a cyclist’ ← *cyklistika* ‘cycling’): *ten, kdo se zabývá* [.*istikou] (‘one who is an enthusiast of...’). We have also processed nouns with meaning “sports players” according to the endings of their respective base (non-derived) words: *-al, -ej* (*fotbalista* ‘a footballer’ ← *fotbal* ‘football’; *hokejista* ‘a hockey player’ ← *hokej* ‘hockey’ with definition *hráč* [.*alu/*.eje] ‘a player of...’).

The lexical meaning was not in correspondence with the structural meaning for nouns derived by the suffix *-tel*, which amounts to 3,01%. This percentage contains 10 nouns which are not primarily animate: *jmenovatel* ‘a denominator’, *dělitel* ‘a divisor’, *čítatel* ‘a numerator’, *násobitel* ‘a multiplier’, *menšitel* ‘a subtrahend’, *úročitel* ‘an interest rate’, *odmocnitel* ‘a square root’, *umořovatel* ‘one payment in a series of installments’, *odúročitel* ‘a discount rate’, *součinitel* ‘a coefficient’; and three nouns which could mean a person, or an inanimate object: *činitel* ‘a factor/an agent’, *ukazatel* ‘a pointer’, *zaměstnavatel* ‘an employer’.

As expected, most of nouns (74,3%) are derived from imperfect verbs and only 25,7% nouns are derived from perfect verbs. Nouns derived from perfect verbs have an identical meaning as the nouns derived from imperfect verbs in 5,31%.

In the future, we want to add online exercises which will enable students to strengthen their knowledge of Czech word-formation, especially the derivation. It will be adjusted to their concrete language level due to Common European Framework (CEFR). We plan to provide examples of use by adding the sentences from the corpus.

In the future, the easiest way of extending the scope of CZEDD would be naturally, via using already created scripts for regenerating more complete databases as its source materials, such as the DeriNet, keep growing.

With continuous work on defining more derivation types, we will be able to find new rules for generating not only more automatic definitions, but also using rule-based approach with cross verification with other resources for further extension of the DeriNet network.

Further didactic functionality is also possible as well as making the results and created materials more accessible with our own API. This approach will make possible both further extensions via third party applications as well as creating a more user-friendly iterations of CZEDD itself.

ACKNOWLEDGMENTS

This work was supported by the project of specific research Czech language in unity of synchrony and diachrony – 2019 (MUNI/A/1061/2018).

References

- [1] Brno Morphological Analyzer Ajka. Accessible at: <https://nlp.fi.muni.cz/projekty/ajka/ajkacz.htm>
- [2] Čechová, M. (1996). *Čeština – řeč a jazyk*. Praha, ISV nakladatelství.
- [3] Cvrček, V. and Vondříčka, P. (2013). *Morfio – aplikace pro analýzu slovtvorných vztahů*. Praha, FF UK. Accessible at: <http://morfio.korpus.cz>.
- [4] Daneš, F., Dokulil, M., and Kuchař, J. (eds.) (1967). *Tvoření slov v češtině 2. Odvozování podstatných jmen*. Praha, Academia.
- [5] Dokulil, M. (1962). *Tvoření slov v češtině. 1, Teorie odvozování slov*. Praha, Nakladatelství Československé akademie věd.
- [6] Glosbe – multilingual online dictionary. Accessible at: <https://cs.glosbe.com/cs/en>.
- [7] Karlík, P., Nekula, M., and Pleskalová, J. (2016). *Nový encyklopedický slovník češtiny*. Praha, Nakladatelství Lidové noviny. Accessible at: <https://www.czechency.org/slovník/>.
- [8] Křen, M. et al. (2018). *Korpus SYN, verze 7*. Praha, Ústav českého národního korpusu FF UK. Accessible at: <https://www.korpus.cz>.
- [9] Osolsobě, K. (2011). *Morfologie českého slovesa a tvoření deverbativ jako problém strojevé analýzy češtiny*. Brno, Masarykova univerzita.
- [10] Osolsobě, K. (2011). *Korpus jako zdroj dat pro studium slovtvorby*. In Petkevič, V. – Rosen, A. (eds.), *Korpusová lingvistika Praha 2011 – 3. Gramatika a značkování korpusů*, pages 10–23, Praha.
- [11] Pala, K., and Šmerk, P. (2015). *Derivancze – Derivational Analyzer of Czech*. In *TSD 2015*, pages 515–523. Accessible at: https://link.springer.com/chapter/10.1007/978-3-319-24033-6_58.
- [12] Skoumalová, Z., Dokulil, M., and Panevová, J. (1997). *Obsah – výraz – význam: Výbor z lingvistického díla Miloše Dokulila I*. Praha, Univerzita Karlova, Filozofická fakulta.
- [13] Ševčíková, M., and Žabokrtský Z. (2014). *Word-Formation Network for Czech (LREC)*. Accessible at: http://www.lrecconf.org/proceedings/lrec2014/pdf/501_Paper.pdf.
- [14] Šimandl, J. (ed.) (2016). *Slovník afixů užívaných v češtině*. Praha, Karolinum. Accessible at: <http://www.slovníkafixu.cz/>.
- [15] Šmerk, P. (2007). *Fast Morphological Analysis of Czech*. In Petr Sojka and Aleš. *Proceedings of Third Workshop of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*, pages 13–16, Brno, Masaryk University.
- [16] Štícha, F. et al. (2013). *Velká akademická gramatika spisovné češtiny*. Praha, Academia.
- [17] *Webové hnízdo o novodobé české slovní zásobě a výkladových slovnících LEXIKO*. Accessible at: <https://lexiko.ujc.cas.cz/heslare/>.
- [18] Žabokrtský, Z., Ševčíková, M., Straka, M., Vidra, J., and Limburská, A. (2016). *Merging Data Resources for Inflectional and Derivational Morphology in Czech (LREC)*. Accessible at: http://ufal.mff.cuni.cz/~straka/papers/2016lrec_derinet.pdf.

CONCEPTION AND DEVELOPMENT OF AN OPEN DATABASE SYSTEM ON HISTORICAL MULTILINGUALISM IN AUSTRIA

KATHARINA PROCHAZKA¹ – LUDWIG MAXIMILIAN BREUER^{2,3,4} – AGNES KIM¹

¹ Department of Slavonic Studies, University of Vienna, Austria

² Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Vienna, Austria

³ Department of German Studies, University of Vienna, Austria

⁴ Centre for Translation Studies, University of Vienna, Austria

PROCHAZKA, Katharina – BREUER, Ludwig Maximilian – KIM, Agnes:
Conception and development of an open database system on historical multilingualism
in Austria. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 456 – 466.

Abstract: This paper discusses the development and structure of an online information system, which aims to gather and visualize data on historical multilingualism in Austria (German: *historische Mehrsprachigkeit in Österreich*, short: MiÖ), with a particular focus on Slavic languages. The database tracks the development of multilingualism over time, its distribution in space and its representation in literature, therefore allowing to examine its dynamics and change. As an example, we investigate the area of the so-called Marchfeld (č./sk. *Moravské pole*). The paper further discusses how the database is embedded into the collaborative research platform of the Special Research Program “German in Austria (DiÖ)” as well as its technical realization and the possibility to include data from other related research projects.

Keywords: online information system, historical multilingualism, language contact, Austria, Austria-Hungary

1 INTRODUCTION: HISTORICAL MULTILINGUALISM IN AUSTRIA

Present day Austria is generally perceived and constructed as a monolingual German-speaking country, with the exception of a few so-called autochthonous minorities (Slovenian in Carinthia and Styria, Burgenland Croatian, Hungarian and Romani in Burgenland and Vienna, Czech and Slovak in Vienna). This paper explores the alternative side of this assumption, focusing explicitly on (historical) multilingualism in Austria. Furthermore, the paper presents the development of a database, which allows to trace the dynamics and change of individual and societal multilingualism over time. The research is embedded within a larger project, the Special Research Program (SFB) “German in Austria (DiÖ). Variation – Contact – Perception” (Austrian Science Fund/FWF F 60, [2]). The database presents the joint work of the project parts of Task Cluster C of the SFB (PP05 and PP06, both concerned with language contact of German with other languages, particularly with Slavic languages. The project is being developed with PP11, which realizes the

collaborative online research platform “German in Austria” (for another relevant collaboration of these projects see [4]).

Historically, Austria was part of the Habsburg Empire, which was both multinational and multilingual. Following the Compromise of 1867, the Habsburg Empire was divided into two parts, the Austrian half and the Hungarian half. These were also known by their unofficial denotations: Cisleithania and Transleithania. Together, they were referred to as Austria-Hungary. The two parts were largely politically independent and pursued separate language policies. While Transleithania legally mandated Hungarian as the only official language, Cisleithania adopted a comparatively liberal acceptance of multilingualism. Multilingualism in Cisleithania encompassed individual and societal multilingualism, meaning that it can be classified as polycentric (see [9, p. 534]). Depending on the local hegemonic and linguistic constellations, it resulted in different forms of diglossia, specific to each crownland. German always played a role in the development of these diglossic situations, as a linguistic majority or minority in that area, or simply by virtue of being the *lingua franca* of Cisleithania (see [8, p. 314]). Other languages, however, also participated in various diglossic situations. Czech, for example, was not only spoken in Bohemia, Moravia and Austrian Silesia, but also in parts of Lower Austria. Within the latter, its status varied considerably between being the language of working migrants and the language of a village’s majority (see [6]).

In this context, the projects of Task Cluster C study the contact of German in Austria mainly with Slavic languages, based on the core assumption that the complex historical multilingualism of Cisleithania played a significant role in shaping the current monolingual view of Austria and German in Austria. Because no typical scenario of multilingualism can be identified in Cisleithania, the two projects conduct case studies to gain a comprehensive perspective of the many facets of language use in Cisleithania. In order to achieve this, they work qualitatively as well as quantitatively. Multiple data sources are combined to overcome the so-called ‘bad data problem’ (inherent to historical sociolinguistics).¹

In the following sections we will present the lexicographic database, representation of linguistic annotation.

2 MIÖ: AN INFORMATION SYSTEM ON (HISTORICAL) MULTILINGUALISM IN AUSTRIA

2.1 Aims and goals

A central aim of the SFB “German in Austria” is sustainability regarding data collection, processing, and provision for the scientific as well as the general public on a collaborative online research platform. Therefore, the data collected and processed in PP05 and PP06 are made accessible as a part of that platform. This part

¹ For further information on the methodology and sources of both projects, see [7].

is concerned explicitly with multilingualism and is thus referred to as the Information system on (historical) multilingualism in Austria (German: *Informationssystem zur [historischen] Mehrsprachigkeit in Österreich*, short: MiÖ).

Currently, there is no comprehensive database encompassing data on (historical) multilingualism in Austria. MiÖ aims to close this gap by providing access to data related to historical multilingualism and its distribution. This includes information on the language skills of individuals and groups, as well as the sociolinguistic context (e.g. legal documents governing the teaching of languages in schools). MiÖ intends to facilitate further research by consolidating all the necessary information. Additionally, it equips its users with the ability to critically evaluate historical documents and sources of any kind by providing a comprehensive bibliography.

The aim of MiÖ is to present and visualize multilingualism in Austria along three axes that can be queried by the user. These constitute the common core of information for any data included (see Fig. 1):

1. its **development** over time: the dynamics and change of multilingualism can be viewed on a time axis, e.g. by searching for a year or a time span.
2. its **distribution** in space: data is geolocated to the respective places and/or regions to enable the visualization of linguistic data associated with a place (e.g. census data, data from linguistic questionnaires).
3. its **representation** in literature: a commented and indexed bibliography is available for any kind of data incorporated into MiÖ.

MiÖ is designed to be an open system that allows the integration of data on (historical) multilingualism from other research projects within separate modules and beyond the end of the SFB “German in Austria”.

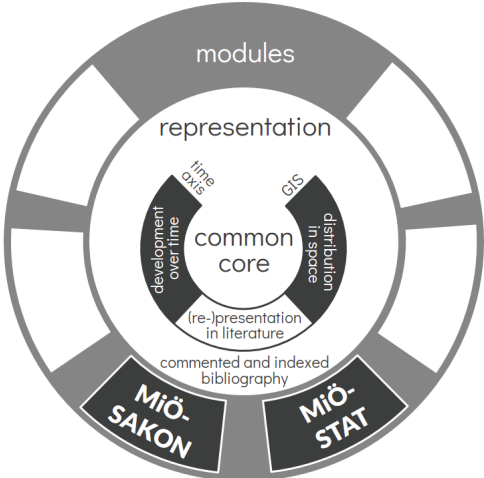


Fig. 1. Schematic overview of MiÖ; white spaces indicate the possibility to integrate further modules.

2.2 Modular structure

In addition to the common core described above, MiÖ is organized in modules to retain maximal flexibility for expansion. In the first project phase of the SFB “German in Austria” (2016–2019), the realization of two modules is planned: MiÖ-STAT (German: *Statistische Informationen zur Mehrsprachigkeit in Österreich*, ‘Statistical Information on Multilingualism in Austria’) and MiÖ-SAKON (German: *Sprachliche Areal- und Kontaktphänomene im Deutschen in Österreich*, ‘Areal linguistic and contact phenomena in German in Austria’). These modules reflect the research foci of the two project parts involved, PP05 and PP06. The influence of contact with other languages on German in Austria is examined in terms of the linguistic layer (linguistic phenomena and variation, PP06) and from a sociolinguistic perspective (language contact scenarios, attitudes towards as well as perception and regulation of multilingualism, PP05). MiÖ-STAT is the first module to be realized within the MiÖ database. MiÖ-SAKON, the second module to be implemented, will provide a catalog of linguistic phenomena in varieties of German in Austria, which are typically ascribed to language contact (particularly with Slavic languages). It will allow users to check, whether a phenomenon can plausibly be explained by language contact or whether such a contact explanation is better described as a language myth. In the following sections, this paper will set aside MiÖ-SAKON and focus instead on the realization of MiÖ-STAT.

As its name suggests, MiÖ-STAT collects statistical information on language use in Austria(-Hungary) from a variety of sources. These data can be rather general (such as the census), or domain-specific (such as data on the linguistic background of students in elementary schools). The time axis of MiÖ-STAT commences in 1867, the year of the Austro-Hungarian Compromise. The covered area extends outside the borders of today’s Austria, also encompassing parts of Cisleithania. MiÖ-STAT is aware of the bias which is inherent to language questions in statistical surveys. Both the phrasing of the question, as well as the political circumstances, considerably influence the self-reported behavior of the informants. Aside from this, surveys may be subject to irregularities, such as the forging of data (for the census in Cisleithania see [1]). MiÖ-STAT acknowledges these biases and consequently contextualizes the data sources with relevant literature.

Additionally, MiÖ-STAT collects and connects information from various sources, thus enabling users to compare them for individual places or entire regions. As shown in exemplary studies (see [5], [6]) such comparison allows for a transparent and more reliable reconstruction of the linguistic situation at a certain place at a specific time.

2.3 Technical realization

MiÖ is embedded within the larger collaborative online research platform of the SFB “German in Austria”. While the other project parts provide recent self-collected linguistic data (recordings of oral speech data and perception data from listener’s judgment tests) for inclusion in the database, MiÖ is the only section

working primarily with written historical documents at this stage. For this reason, the database structure has to be extended and adapted to manage the challenges of working with historical data. This mainly concerns the implementation of “time” in terms of an analyzable variable and an integrative component of the modeled information structure (e.g. change of names for places or regions, see below). Finally, all data generated or collected within the SFB “German in Austria” should be connected, to ensure that the data (historical and at some point also contemporary) on MiÖ can be analyzed in the context of the linguistic data on DiÖ and vice versa.

MiÖ integrates various *types of data*: (scientific) literature, statistical data (MiÖ-STAT, see above), information on linguistic phenomena (MiÖ-SAKON, see above) and their classification, among others. These can be described with regard to the *information types* reflected in the common core of MiÖ (see Fig. 1). As the data sources integrated into MiÖ are diverse, the database allows both the incorporation of the original documents (as scans or images where possible) and the machine-readable digitization of the information contained within.

The underlying database management system (DBMS) of the DiÖ online research platform is PostgreSQL². This is implemented with the Django web framework³ that allows to model the data structure on a separate abstract layer, independent of the back-end. PostgreSQL is a widely used and well documented DBMS that provides many important functions (e.g. GIS extension, JSON integration). Its entire development is open source and published on GitHub⁴, which allows developers to share their tools directly with the community, simultaneously strengthening the sustainability of these tools, since it provides the possibility for maintaining and developing them even after the project is finished.

2.4 Current status

So far, the collaboration has focused primarily on the implementation of the back-end for the common core and the module MiÖ-STAT. MiÖ will be available via the browser interface of the DiÖ online research platform which is under development as of July 2019. In this section, we provide an example of how MiÖ-STAT may support the research process concerning questions about sociolinguistic aspects of historical multilingualism in Austria.

As noted above, MiÖ-STAT includes statistical information that refers to certain places and stems from various sources. Fig. 2 shows a simplified data model of the underlying database. The core information, i.e., temporal, bibliographical and geographical information, is highlighted in grey. Additionally, Fig. 2 indicates that the module is embedded into MiÖ and the DiÖ research platform and how the information in MiÖ-STAT is linked to the DiÖ data.

² <https://www.postgresql.org/>

³ <https://www.djangoproject.com/>

⁴ <https://github.com/german-in-austria/>

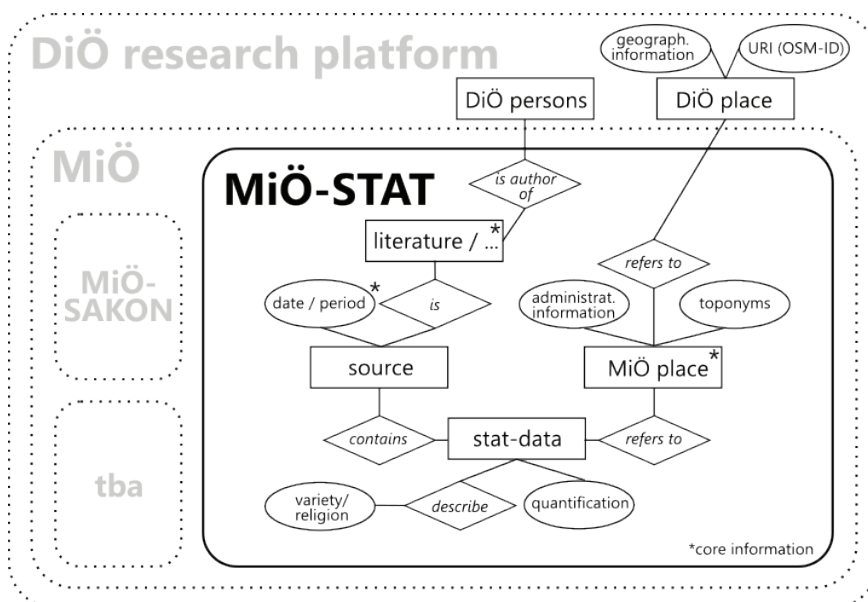


Fig. 2. Simplified entity-relationship model of MiÖ-STAT and its embedding into MiÖ and the DiÖ research platform

The model links contemporary places⁵ (in the DiÖ place-DB) to their historical equivalents (in the MiÖ place-DB) and allows to map their changing administrative affiliation and name changes. This is essential for the inclusion of data with varying resolution, as some sources do not provide information on single localities but only on the municipality or judicial district level. If possible, events (such as a change in administrative affiliation) are assigned exact dates (in the format DD.MM.YYYY) or time spans between two exact dates. Where exact temporal information is not available, as is usually the case with the publication date of books, the time span in which the event must have taken place is indicated.

Regarding the statistical information, MiÖ-STAT also strives to depict the original data source as accurately as possible. For example, the census for Cisleithania did generally not distinguish Czech from Slovak and referred to speakers of these languages as *Čecho-Slaven* ‘Czecho-Slavs’ in 1880. Therefore, the information is not directly linked to a language (variety, see [4]) or religion (denomination) but rather to the name for the respective language or religion used in the source (see [a] in Fig. 3).

⁵ Regarding place terms, we distinguish several administrative levels (from the largest to the smallest, with Austrian terminology): country, crownland / state, political district, judicial district, municipality, locality. The smallest, i.e. the locality level, is our main point of reference. The larger entities are described as the sum of the smaller entities they comprise. Contemporary place terms are defined as such that officially existed on Jan. 1st, 2018.

The input masks are optimized for easy entry of the statistical data and automatically sum up the entered numbers in order to enable immediate self-checks during the entry process (see [b] in Fig. 3).

VZ: lit: 1880_noe (1883-01-01 - 1883-12-31) - ID: 1

Ort: noe

Ort: Loimersdorf, Bezirk Gänserndorf, Niederösterreich, 2292, Österreich - ID: 33

Bezeichnung	Anzahl	Kommentar
gesamte Bevölkerung	481	
Religionen [a]		
Kathol.	477	
Protest.	0	
Israeliten	4	
Andere	0	
Varietäten		
Deutsche	170	
Čecho-Slaven	13	
Andere	265	Serbo-Kroaten

Varietäten: Gesamt: 481 entspricht nicht der Summe: 448 (Differenz: -33)
 Religionen: Alles OK

VZ Daten Speichern

Fig. 3. Input mask for data from a specific census (1880) for a specific locality (Loimersdorf)

In the short run, MiÖ-STAT aims to achieve comprehensive high-resolution coverage of selected multilingual regions of the former Habsburg monarchy. In the long run, Austria within its current borders should, at least, be covered comprehensively. This goal is achieved by including census data per locality, which is subsequently enriched by other statistical data sources. In the area of the so-called Marchfeld (č./sk. *Moravské pole*) between Vienna and Bratislava, which has a size of approx. 900 km², this yields 84 locations and hence 336 data points, if only the census data from 1880, 1890, 1900 and 1910 are included.

Queries made using the online accessible front-end will provide answers to questions such as: Where was Czech/Slovak⁶ spoken at a certain point in time? How did the percentage of Czech/Slovak speakers in certain places develop over a certain time period (see Fig. 4 for 1880 and 1910 for the Marchfeld)?

Maps, such as Fig. 4, help to visually identify places of interest at a certain point in time. Kim/Prochazka [7] have proposed a mathematical method to

⁶ The Cisleithanian census did not differentiate between Czech and Slovak, but subsumed both languages under the glottonyms *čechoslavisch* 'czechoslavic' (1880) or *böhmisch-mährisch-slovakisch* 'Bohemian, Moravian, Slovak' (1890, 1900, 1910).

formally determine places of interest across various points in time. It assumes that atypical changes in the percentage of speakers of other languages than German and foreigners⁷ shed light on the migration and assimilation history of a certain place. Moreover, locations with fluctuating patterns in the census results of the Habsburg monarchy may indicate a high degree of individual and societal bilingualism.

We have identified remarkable developments between 1880–1910 by calculating the change of percentage of speakers using a language other than German and foreigners between two subsequent censuses. If the sum of the absolute values for all three time steps (1880–1890, 1890–1900, 1900–1910) is larger than 20%, we can assume that the development is remarkable. Regarding the Marchfeld, this procedure yields 26 places of special interest, i.e. 40% of the complete sample (see Fig. 5).

The inclusion of sources, other than the Cisleithanian census, such as earlier population counts (e.g. [3]), census data from the Inter-War period (1934) or linguistic questionnaires (e.g. Wenker's questionnaires, see [6]) provides new perspectives. The possibility to retrieve the data from the system will give the opportunity for further analyses. Thus, MiÖ-STAT will provide essential information to stimulate a closer and interdisciplinary scrutiny of historical multilingualism in Austria.

2.5 Expandability and outlook

An important goal is to design the DiÖ research platform and therefore MiÖ as an open system, which provides a link between the DiÖ/MiÖ data and alternative data and data types from other projects. Therefore, a standardized unique identifier (as object identifier) is essential to blend additional data with existing data and (geographical) entities. To this end, we employ longitude-/latitude-coordinates for identifying points or shapes in a geographical coordinate system (i.e., the World Geodetic System 1984, WGS84), as well as IDs provided by OpenStreetMap⁸ (as an open source tool with a transparent license) in combination with the object type (village, municipality, street etc.). This enables a convenient and globally unique identification of a geopolitical entity, not only in the position or shape of the entity but also regarding its (socio-)political information.⁹

⁷ The Cisleithanian census only gave information on the so called *Umgangssprache* 'everyday language' of its citizens.

⁸ <https://www.openstreetmap.org/>

⁹ Another strategy to guarantee a high reusability and integration of the tools previously mentioned in section 2.3 is the implementation as a *docker* container (<https://www.docker.com/>), which can easily be integrated in various IT infrastructures without having to install all necessary dependencies.

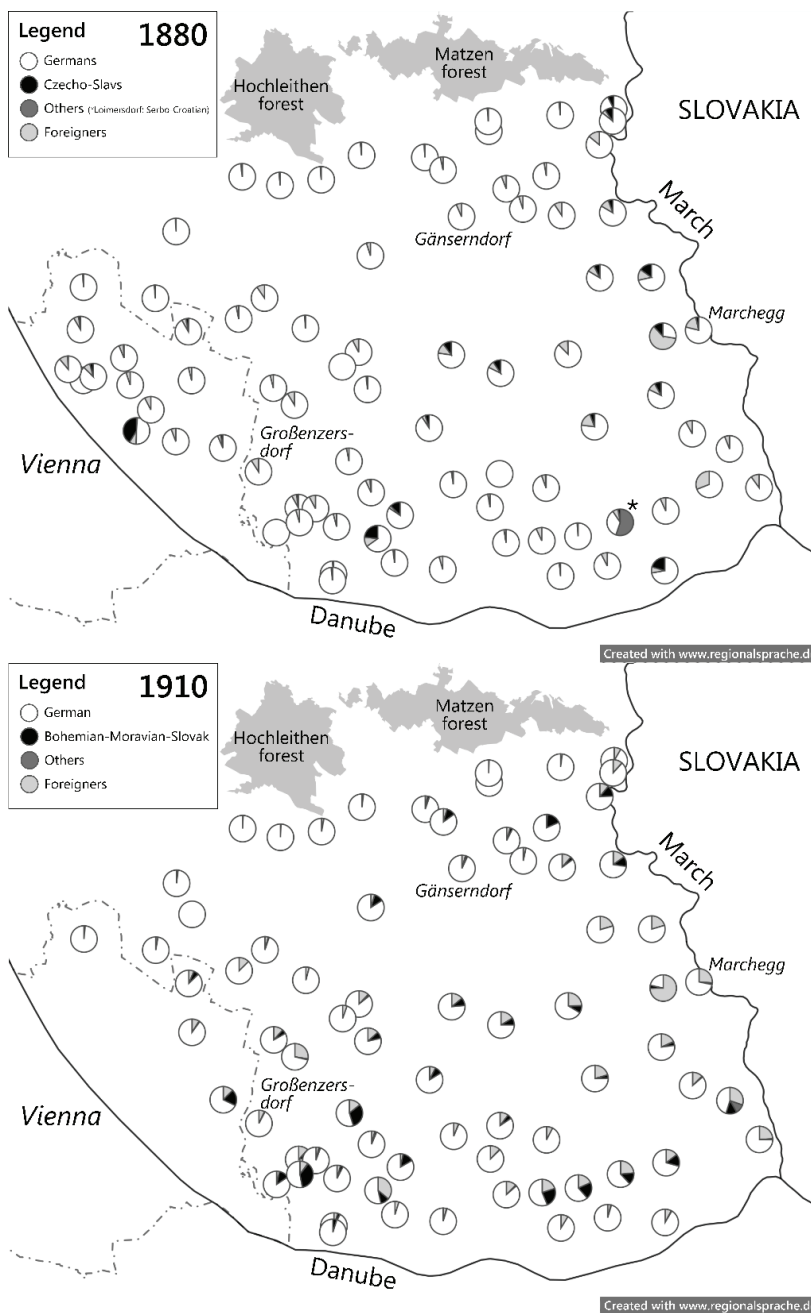


Fig. 4. Census results for all localities in the Marchfeld 1880 and 1910 (exemplary illustrations)

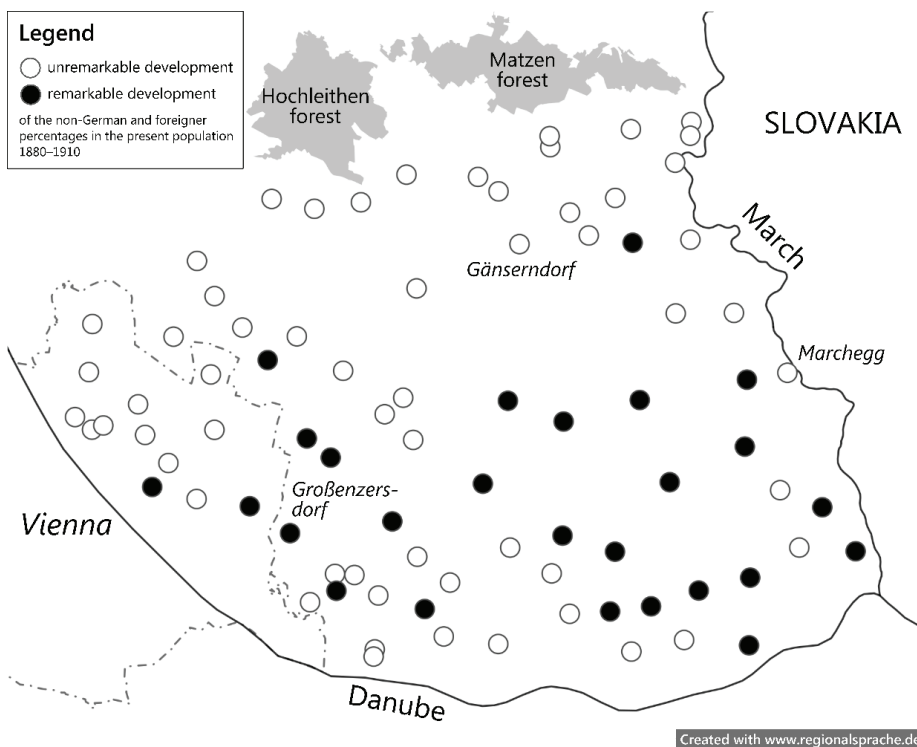


Fig. 5. Places of special interest in the Marchfeld (exemplary illustration)

Due to its open design, the database potentially allows for the direct input of data by other researchers (which can then be private or public and linked to the aforementioned unique identifiers). This is, however, a sensitive legal topic. Therefore, individual cases and the applicable licenses must be thoroughly explored before implementing this possibility.

As noted in section 2.1, the aim of the MiÖ (and DiÖ) database is to stimulate further research by using the data available in MiÖ (and the DiÖ research platform). To ensure flexibility and reusability for the users, the database will include an export function in various formats (CSV, Excel, JSON) to ensure that the data can be easily analyzed and visualized in various contexts and research projects.

Ultimately, MiÖ strives to collect various types of data on (historical) multilingualism in Austria beyond its current borders. It aims to provide information for the reconstruction of historical, sociolinguistic contact scenarios and offer information regarding the seemingly factual knowledge concerning languages and multilingualism in Central Europe.

ACKNOWLEDGMENTS

This work is supported financially by the Austrian Science Fund (FWF): F 60 “German in Austria (DiÖ). Variation – Contact – Perception”. The MiÖ database is a collaboration between PP05 (“German in the context of the other languages in the Habsburg State [19th century] and 2nd Austrian Republic”, F 6005; PI: Stefan Michael Newerkla) and PP06 (“German and the Slavic languages in Austria: Aspects of language contact”, F 6006, PI: Stefan Michael Newerkla). It is being developed collaboratively with PP11 (“Collaborative Online Research Platform”, F 6011; PI: Gerhard Budin).

The authors wish to thank Maria Schinko for help in digitizing and checking census data, Lena Katzinger and Katherine Jackson for proofreading and Stefan Michael Newerkla and Wolfgang Koppensteiner for comments on the manuscript.

References

- [1] Brix, E. (1982). Die Umgangssprachen in Altösterreich zwischen Agitation und Assimilation. Die Sprachenstatistik in den zisleithanischen Volkszählungen 1880 bis 1910. Wien, Böhlau.
- [2] Budin, G., Elspaß, S., Lenz, A. N., Newerkla, S. M., and Ziegler, A. (2018). Der Spezialforschungsbereich „Deutsch in Österreich (DiÖ). Variation – Kontakt – Perzeption“. *Zeitschrift für Germanistische Linguistik* 46(2), pages 300–308.
- [3] Czoernig, K. (1857). *Ethnographie der oesterreichischen Monarchie*. Wien, Kaiserl. koenigl. Direction der administrativen Statistik.
- [4] Kim, A., and Breuer L. M. (2017). On the development of an interdisciplinary annotation and classification system for language varieties. Challenges and solutions. *Jazykovedný časopis* 68(2), pages 191–207.
- [5] Kim, A. (2018). Von „rein deutschen“ Orten und „tschechischen Minderheiten“. *Spracheinstellungen und bevölkerungspolitisches Bewusstsein in den Wenkerbögen*. In Philipp, H., Ströbl, A., Weber, B. and Wellner, J. (eds.). *Deutsch in Mittel-, Ost- und Südosteuropa*, pages 275–318, Regensburg, Universitätsbibliothek Regensburg.
- [6] Kim, A. (2019). Multilingual Lower Austria. Historical sociolinguistic investigations on the Wenker questionnaires. In Bülow, L., Fischer, A.-K. and Herbert, K. (eds.). *Dimensionen des sprachlichen Raumes. Variation – Mehrsprachigkeit – Konzeptualisierung*, pages 187–211, Frankfurt am Main, Peter Lang Verlag.
- [7] Kim, A., and Prochazka, K. (2019). Slawisch und Deutsch in Österreich. Methodische Ansätze zur Rekonstruktion historischen Sprachkontakts und seiner Einflüsse auf das Deutsch in Österreich. In *Wiener Slavistisches Jahrbuch*. N.F. 7, pages 1–27.
- [8] Rindler Schjerve, R. (eds.) (2003). *Diglossia and power. Language policies and practice in the 19th century Habsburg Empire*. Berlin/New York, Mouton de Gruyter.
- [9] Stewart, W. A. (1968). A sociolinguistic typology for describing national multilingualism. In Fishman, J. A. (eds.): *Readings in the Sociology of Language*, pages 531–545, Berlin, de Gruyter.

ON POSSIBILITIES AND METHODS OF ANALYSIS OF THEMATIC EXPRESSIONS IN SPOKEN TEXTS

PETR POŘÍZKA

Faculty of Arts, Palacký University, Olomouc, Czech Republic

POŘÍZKA, Petr: On possibilities and methods of analysis of thematic expressions in spoken texts. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 469 – 480.

Abstract: The treatise focuses on mutual comparison of three methods of detection of prominent text units (prominent in relation to the contents of the text). The methods are: 1) analysis of key words based on comparison of source and referential corpora, 2) thematic concentration and h-point, and 3) the TF*IDF method. We try to thematize their pros and cons and, using the results of the carried out analyses, propose the optimal method for the extraction of thematic words from the spoken texts the frequency structure of which differs distinctly from the frequency structure of written texts.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 INTRODUCTION

Quantitative linguistics disposes of methods that are used to recognize main topic(s) of texts or keywords in the texts. Methods of extraction of these so-called *prominent units* are tested on texts of different genres and they are predominantly used to analyze written texts. This study intends to find out to what extent the selected methods of analysis can be used to extract prominent units in spoken texts. It is well known that in its form spoken language often differs distinctly from written language. From the quantitative viewpoint, the difference is evident even if we compare frequency vocabulary of spoken and written texts. Spoken dialogues have a specific frequency structure and a clearly distinct frequency distribution of individual parts of speech (henceforth POS). This fact can have relevant consequences since these methods of analyzing prominent units are based on word lists (or on the comparison of those lists) and on frequency structure of texts. Let us now see frequency structure of POS in large corpora of written Czech included in the Czech National Corpus (CNC; the column *CNC-written* represents the average values of POS of the SYN line of corpora) and in representative spoken corpora of CNC (the column *CNC-spoken* represents the average values of POS of ORALv1 and ORTOFONv1 corpora). In the table, relative frequency in per cent is stated.¹

¹ Partial corpora of the SYN line contain approximately 100 million words, ORALv1 includes about 5.5 mil. words and ORTOFONv1 about 1 mil. words.

POS	CNC-written	CNC-spoken
Noun	30.53	11.41
Adj	11.48	3.50
Pron	10.48	20.27
Num	3.17	2.04
Verb	16.86	20.15
Adv	7.10	12.84
Prep	10.55	5.67
Conj	7.56	11.48
Part	0.99	8.38
Interj	0.05	0.42
<i>resp+hes</i>	---	2.15
<i>uncomp</i>	---	1.04
<i>unknown</i>	1.26	0.65

Tab. 1. Frequency distribution of parts of speech in written and spoken corpora of CNC. Number represent relative frequency in per cent. Legend: *resp+hes* = response and hesitation; *uncomp* = uncompleted words; *unknown* = expressions not recognized by a tagger

As we can see, the differences are manifested most significantly in the distribution of *nouns*: the frequency of their appearance in spoken texts is distinctly lower than in written texts (approx. 30% written vs. 11% spoken); a similarly distinct decrease is documented in the distribution of *adjectives* (approx. 11.5% vs. 3.5%) and *prepositions* (approx. 10.5% vs. 5.5%). On the other hand, the frequency of *adverbs* (7% vs. 13%) and *particles* (1% vs. 8%) rises.² As we will demonstrate, thematic expressions are extracted from nouns, adjectives and verbs. And nouns, as expressions signifying *substances*, are undoubtedly significant for any method the aim of which is to detect prominent text units. On the basis of these differences we intend to find out to what extent the perceptibly lower frequency distribution of nouns (and possibly even other differences) will be manifested in our analyses carried out with the use of selected methods.

2 DATA, METHODS, TOOLS

For our probe we chose two of currently often used methods of extraction of prominent units: 1) *analysis of keywords* and 2) the method of measuring *thematic text concentration*, namely the part of the method in which thematic words are detected. The third method is 3) TF*IDF method (*Term Frequency vs. Inverse Document Frequency*), used in semantic analysis of texts. In the text analyses, following freely available software tools were applied: (ad 1) *KWords* [1], (ad 2) *QUITA* [2], and (ad 3) *KER – Keyword Extractor* [3].

² Among particles even hesitation and response sounds might be included (the category of *resp+hes* in Table 1); thus their proportional representation would rise by 2% to the final proportion of 10%.

The analyzed data were formed by 20 spoken texts randomly selected from the so-called *Olomouc spoken corpus* (henceforth OSC) [4]. We used orthographically normalized/standardized versions of transcripts that were further purified in order to suit our intentions. We removed all their parts that could affect textual analysis: particularly marks of individual speakers (before all lines) and all meta-textual marks and commentaries. Individual transcripts contained between 2,300 and 4,500 words (the average of 3,135 words in a transcript); the overall size of the dataset was 62,694 words.

The transcripts were subsequently lemmatized for *KWords* and *KER* with the use of *MorphoDiTa*, a morphological analyzer and tagger [5]. While working with QUITA we used a morphological analyzer *Majka* [6].

Quantitative analysis of so-called *keywords* (further on also KWs) ([7], [8]), based on the comparison of the source text (SourceC) with so-called referential corpus (RefC) is certainly one of the most commonly used methods of content analysis of texts. For keywords we take the words the frequency of which is remarkably higher in the SourceC than in the RefC. Nevertheless, the choice of the RefC influences even the overall result of the analysis and it is therefore recommended to choose textually neutral databases that reflect common language usage. For the detection of statistic relevance of differences two statistical tests are used: *log-likelihood* and *chi-squared test*. Even this exact method has its difficulties that have to be faced, namely with respect to appropriate combination of computing parameters. It is primarily necessary to set *the level of statistic significance* of the test (most frequently to 0.05, 0.01, 0.001, or even more) and sometimes other parameters (see below Sec. 3.2). It is also possible to apply so-called *stop-lists* on the text; by stop-lists we mean the lists of words or word groups that are *a priori* excluded from the analysis of KWs. Among problematic aspects of this kind of analysis belongs the fact that the analysis produces quite large lists of detected KWs (sometimes containing hundreds or even more words) that have to be in some way reduced in order to be used in subsequent analysis and interpretation of the text. Such reductions are often arbitrary, based on some *ad hoc* criteria: most frequently only the group of the initial 20, 50 or 100 words is taken from the list of all detected KWs and applied in the interpretation. That is why even the position of a certain keyword in the final list is important, the position reflecting a simple principle: the higher in the list the KW appears, the more relevant it is for the contents and topic(s) of the text. In this way KWs are hierarchized; KWs can certainly be sorted out according to the coefficient of the main statistical test. We can also use any index reflecting the relevance of different distribution of the word in the SourceC and in the RefC, or the index applied in order to neutralize the different sizes of source and referential corpora. For example, in the latest version (3.5.8) of a concordance tool *AntConc* [9] 10 indexes of this kind are implemented.³

³ Compare individual indexes in the menu of *Keyword Effect Size Measure*.

The above stated characteristics of *keyword analysis* show that this is a relatively demanding procedure during which one must set many parameters that affect the process and the resulting list of KWs. Researchers therefore look for other ways and methods leading to the revelation of main topics of texts. Recently, namely the analysis of thematic words has been tested and developed that utilizes measuring of *thematic text concentration* (further on also TC) [10]. The method is based on simple extraction of thematic words (TW) from a word list; to detect thematic words one needs no external database nor further mathematical modeling of the text that would prefer certain words to others and modify their position in the word list. The method considers as thematic the words that occupy the positions above so-called *h*-point in the word list, while the *h*-point is defined as a position in which the rank of the word equals the frequency of the word.⁴ The *h*-point concurrently represents an indistinct borderline between autosemantic and synsemantic POS: all autosemantic expressions, with the exception of adverbs and certain verbs (see below) that appear above the *h*-point are subsequently considered as main topics of the text.

Nevertheless, in practice the use of the method often results in empty TW sets. The texts with an empty set of thematic words are subsequently considered as thematically neutral while the texts in which one detects TWs are thematically determined. In order to eliminate the cases of empty TW sets, the so-called STC (secondary thematic concentration) was implemented in the method which means that the TC value is multiplied by 2 in order to shift the *h*-point lower in the word list and to increase the chance of finding some prominent units. We consider this solution as rather problematic since it is quite arbitrary and it leaves without explanation why TC values are multiplied by 2 and not by other numbers. But there is also a question: Isn't the choice of *h*-point arbitrary in itself?

The choice of an elementary text unit is methodologically relevant as well. Shall it be the word form, a lemma, or even other unit? It is common to take a *text form* as the elementary unit of the analysis but it is evident (from previous analyses) that in case of a strongly inflective Czech *lemma* is definitely a more appropriate choice since it represents all text forms of a lexeme.

3 ANALYSIS AND INTERPRETATION OF ITS RESULTS⁵

3.1 TC and thematic expressions

Lemma is the elementary unit of our analyses. Besides their lemmatization we annotated the texts even morphologically – we assigned the mark of its affiliation with a particular part of speech to each text unit. In Table 2 below we indicate frequency distribution of individual POS in our specimen of data in comparison with

⁴ For comments to the formula and to the calculation of the *h*-point see [10] (pp. 11nn).

⁵ Here we will restrict ourselves to interpretational remarks. Complete resulting lists of KWs are available and can be freely downloaded at: <http://corpus.upol.cz/system/files/KWs-lists.zip>.

morphologically annotated spoken CNC corpora. Since we applied two different taggers (see *Sec. 2*) both variants of annotation are presented in the table:

POS	OSCsample20 <i>MorphoDiTa</i>	OSCsample20 <i>Majka</i>	ORAL v1	ORTOFON v1
Noun	13.59	12.19	11.63	11.18
Adj	4.03	4.02	3.63	3.38
Pron	20.37	20.33	20.86	19.67
Num	1.89	1.81	1.77	2.31
Verb	21.8	21.67	20.46	19.84
Adv	14.37	15.5	12.93	12.74
Prep	5.92	5.96	5.66	5.69
Conj	11.84	11.67	11.51	11.46
Part	5.3	4.51	8.13	8.63
Interj	0.91	0.77	0.43	0.4
<i>resp+hes</i>	---	---	1.64	2.67
<i>uncomp</i>	---	---	0.75	1.33
<i>unknown</i>	0	1.62	0.6	0.7

Tab. 2. Comparison of frequency distribution of POS in the analyzed specimen of spoken data (*OSCsample20*) and in the spoken CNC corpora. The numeric values signify relative frequency in per cent.

he comparison enables us to suppose that the selected specimen of spoken data can be considered as representative since the frequency distributions of POS correspond with those in much larger databases (*OSCsample20*: $N \doteq 63$ thousand; *ORTOFON*: $N 1.03 \doteq$ million; *ORAL*: $N \doteq 5.5$ million words). It is significant, namely with respect to the TC method and its POS limitation of thematic words. We notice certain deviations (for example in the frequency distributions of *particles*, *adverbs* or *nouns*) but they are only minute (avg. 1.5% in case of nouns, 2.1% in case of adverbs, and 3.5% in case of particles) and therefore they cannot affect the analysis of thematic words. The proportional representation of nouns in *OSCsample20* is even slightly higher than in the CNC corpora.

The results of the analysis of thematic words carried out with the use of *QUITA* tool are presented in Table 3:

DOC	TWs according to TC
1	vědět 'to know', jít 'to go'
2	0
3	vědět 'to know', říkat 'to say'
4	0
5	vědět 'to know'

6	vědět ‘to know’
7	vědět ‘to know’, hrát ‘to play’, dělat ‘to do’
8	0
9	říkat ‘to say’, vědět ‘to know’
10	rok ‘year’, fotbal ‘soccer’
11	vědět ‘to know’, koupit ‘to buy’
12	vědět ‘to know’
13	jít ‘to go’
14	vědět ‘to know’, jet ‘to go’, jezdit ‘to go’
15	0
16	vědět ‘to know’
17	0
18	jet ‘to go’, vědět ‘to know’
19	vědět ‘to know’
20	dobrý ‘good’, vědět ‘to know’

Tab. 3. Results of the analysis of thematic words in spoken texts (OSCsample20). TWs are arranged according to their ranking.

In 5 out of 20 texts, i.e. in 25% of cases, no thematic words were found – they are texts Nos 2, 4, 8, 15 and 17 (mind their absence in Table 3). In all remaining documents only 11 different prominent units were found. We suppose that only some of them can be considered as real thematic words. Particularly they are these: *hrát* ‘to play’, *koupit* ‘to buy’, *rok* ‘year’, *fotbal* ‘soccer’. They are marked by bold print in Table 3 and they were found only in 3 out of 20 texts. Other lexemes rather indicate deviation from semantic (thematic) to pragmatic use (we verified the character of their behavior with the use of concordances in our corpus). It is true namely in case of the verb *vědět* (‘to know’, a verb of mental action, communication) that appeared in 13 out of 15 texts or in the cases of *říkat* (‘to say’, v. dicendi, communication) and *dobrý* (‘good’, an evaluative adjective). As prominent units only two other verbs were detected: *dělat* (‘to do’, v. faciendi), and *jít/jet/jezdit* (‘to go’, v. movendi).

A question arises whether it is possible to take the lexeme with noticeably pragmatic use for thematic expression. In spoken texts such lexemes often function as phatic, conative or emotional/expressive words while real thematic words should function as referential units (that signalize the relation to the topic).

The situation slightly improves in case of the STC index (Table 4). Nevertheless, we consider (as we stated above) STC as methodologically problematic. Besides, the authors of TC consider the texts without TWs as thematically neutral and the texts

with TWs as thematically determined. A paradoxical situation thus arises in which the same texts with originally empty TW sets suddenly, thanks to STC, become thematically determined.

DOC	TWs according to STC
1	(vědět 'to know'), jít 'to go', dělat 'to do', (říci 'to say'), (říkat 'to say')
2	baterka 'flashlight' , dát 'to give', (vědět 'to know'), udělat 'to do', (říkat 'to say'), třešeň 'cherry'
3	(vědět 'to know'), (říkat 'to say'), (řici 'to say')
4	(vědět 'to know'), potřebovat 'to need', lednička 'fridge' , dát 'to give', udělat 'to do', koupit 'to buy' , jet 'to go'
5	(vědět 'to know'), (říkat 'to say'), (myslit 'to think'), jet 'to go', dělat 'to do'
6	(vědět 'to know'), (hezky 'pretty'), (myslit 'to think'), (krásný 'beautiful'), Krkonoše
7	(vědět 'to know'), hrát 'to play' , dělat 'to do', dát 'to give', statistika 'statistics' , (řici 'to say'), kluk 'boy' , zápas 'match' , jít 'to go', (mhm), (myslit 'to think')
8	(vědět 'to know'), sval 'muscle' , jet 'to go', mozek 'brain' , (říkat 'to say'), dělat 'to do'
9	(říkat 'to say'), (vědět 'to know'), jít 'to go', (dobry 'good'), dívat 'watch', napsat 'to write' , psát 'to write'
10	rok 'year' , fotbal 'soccer' , hrát 'to play' , (myslit 'to think'), jít 'to go', řada 'row' , celý 'all' , hráč 'player' , (řici 'to say')
11	(vědět 'to know'), koupit 'to buy' , libra 'pound' , jít 'to go', dát 'to give', (myslit 'to think')
12	(vědět 'to know'), vidět 'to see', (myslit 'to think'), (říkat 'to say')
13	jít 'to go', (říkat 'to say'), (vědět 'to know'), pamatovat 'to remember' , dítě 'child' , chodit 'to go'
14	(vědět 'to know'), jet 'to go', jezdit 'to go', (říkat 'to say'), jít 'to go', psát 'to write', týden 'week' , škola 'school' , Honza 'Johnny' , přijet 'to come', spát 'to sleep'
15	(vědět 'to know'), jít 'to go', dělat 'to do', jet 'to go', člověk 'man' , (dobry 'good')
16	(vědět 'to know'), jít 'to go', (říkat 'to say'), (dobry 'good'), (řici 'to say')
17	Martin , (dobry 'good'), (vědět 'to know')
18	jet 'to go', (vědět 'to know'), jít 'to go', (dobry 'good'), Skotsko 'Scotland'
19	(vědět 'to know'), jít 'to go', přijít 'to come'
20	(dobry 'good'), (vědět 'to know'), fotka 'photo' , jméno 'to go' name , vidět 'to see', dívat 'to watch'

Tab. 4. Thematic words according to STC (OSCsample20).
TWs are arranged according to their ranking.

This time thematic words were detected in all partial documents of the dataset. Even if the STC caused the growth of detected lexemes they are actually verbs (or adjectives) again, functioning as pragmatic (phatic) words. The words can further be gathered in groups that share the same word-formation base or form pairs in which one verb is imperfective and the other one perfective: *říci–říkat* ‘to say’, *dělat–udělat* ‘to do’, *psát–napsat* ‘to write’, *jít–přijít–chodit* ‘to go on foot’, *jet–jezdí* ‘to go’. Among adjectives we can find increments with the same meaning and function and belonging to the same category (evaluative words): *dobrý* ‘good’, *hezký* ‘pretty’, *krásný* ‘beautiful’.

On the basis of the behaviour of all prominent units in the spoken texts verified with the use of corpus concordances the prominent TC/STC words can be divided in three zones/categories:

- 1) *non-thematic expressions* with pragmatic function (such as *dobrý* ‘good’, *hezký* ‘pretty’, *myslet* ‘to think’, *vědět* ‘to know’, *říkat* ‘to say’) – in Table 4 they are stated in parentheses;
- 2) *a broad transitional zone of borderline expressions*: namely *verbs* and *adjectives* that can be recognized as both thematic and pragmatic (for example *vidět* ‘to see’, *dívat se* ‘to watch’, *potřebovat* ‘to need’); these expressions appear repeatedly in most analysed texts;
- 3) *truly thematic expressions* (for example *baterka* ‘torch’, *lednička* ‘fridge’, *zápas* ‘match’, *fotbal* ‘soccer’, *škola* ‘school’ etc.) – they are almost solely *nouns* – in Table 4 they are stated in bold print.

If we sum the results of our analyses up they seem to suggest that, in case of spoken texts, the TC/STC method fails. It may be caused by the fact that spoken texts differ from the written texts significantly: they have a specific frequency structure of the text/vocabulary, they contain many pragmatically used expressions, functioning as phatic, conative or emotive words.

3.2 *KWords* and key-lemmas

We used the *KWords* tool and carried out the analysis with following settings:

- stop-list: pronouns, prepositions, conjunctions, numbers
- methods: *log-likelihood*
- significance level (α): 0.0001
- minimal frequency: 3
- percentage of registered keywords: all significant types
- referential corpus: SYN2015

The list of keywords can be arranged according to DIN that signalizes the relevance of differences in KWs in the SourceC and RefC. We limit the list of lemmas to the units of high and highest prominence ($DIN > 95$).⁶

⁶ For the calculation formula and more detailed description of DIN values see [1].

We analysed 5 documents of the *OSCsample20* set; three of them (Nos 2, 4, and 8) had an empty TC set while in case of the remaining documents (Nos 7 and 11) the set was not empty. Given the extent of the study and the fact that the resulting list of keywords are rather large, we will limit ourselves merely to brief remarks and possible conclusions that follow from our analyses:

- The DIN index functionally and effectively reduces the number of keywords and it also hierarchizes KWs.
- If we limit the list of lemmas to the units of high (DIN: 95–97) and highest (DIN 98–100) prominence, the resulting lists will contain approx. 40 up to 60 words in the texts (cf. Below):

DOC	DIN 95–97	DIN 98–100	DIN 95–100
2	20	36	56
4	16	38	54
7	30	37	67
8	13	43	56
11	29	35	64
MEAN	21.6	37.8	59.4

Tab. 5. The number of keywords in *KWords* tool

- Only very few pragmatically used words appear in the lists: they are following particles (*ano, jo, hm, no, tož*) or interjections (*aha, hele, jé*) and should here be regarded not as prominent units but rather as pragmatically applied words.
- Frequency POS distribution of resulting keywords suggests that the highest positions in the list are actually occupied by thematically significant expressions:

POS	FREQ	FREQ %
Noun	134	44.97
Verb	110	36.91
Adj	22	7.38
Part	18	6.04
Adv	6	2.01
Interj	6	2.01
Num	2	0.67
Total	298	100.00

Tab. 6. Frequency distribution of POS in *KWords* tool

- Unlike in TC, certain adjectives (such as *dětský* ‘childish’, *infekční* ‘infectious’, *levoruký* ‘left-handed’, etc.) and verbs (such as *lyžovat* ‘to ski’, *pršet* ‘to rain’, *vyléčit* ‘to rain’, etc.), i.e. the words that can truly be regarded as thematically prominent.

It seems that the *analysis of keywords* is more suitable for the detection of prominent units in spoken texts than the method based on *thematic concentration of texts*.⁷

3.3 *KER* and TF*IDF method

TF*IDF method [11] compares the frequency of the word in the analysed text with the “reversed” frequency of the word in all documents. IDF expresses the “relevance” of the word: the more frequently a particular word appears in the documents the less relevant it is for the analysed text. From mathematical viewpoint the method is relatively simple:

$$\text{TF}(t) = (\text{Number of times term } t \text{ appears in a document}) / (\text{Total number of terms in the document})$$

$\text{IDF}(t) = \log_e(\text{Total number of documents} / \text{Number of documents with term } t \text{ in it})$.

The demo version of *KER – Keyword Extractor* has certain limitations. We therefore carried out analyses with following settings:⁸

- TF*IDF threshold level: 0.05
- maximum number of keywords: 25

This setting has turned out as optimal in majority of the analysed texts: the resulting number of KWs is lower than the pre-set maximum limit (5 out of 20 texts reached the maximum limit). Moreover, it turned out that the detected number of KWs does not depend directly on the length of the text. Texts Nos 1, 4, and 14 that reached the maximum limit of 25 KWs do not even contain the average number of words. By the way of contrast, texts Nos 4 and 10 (approx. 2,300 words) and texts Nos 13 and 17 (approx. 3,000 words) are almost equally long. Nevertheless, the numbers of KWs that were found in texts of the same length differ diametrically: doc4: 25 × doc10: 4; doc13: 25 × doc17: 7. Cf. below:

DOC	TOKENS	KWs
1	2561	25
2	3667	14
3	3333	11
4	2358	25
5	3183	13
6	2835	15

⁷ We should point out that we compared spoken texts (SourceC) with written ones (RefC). Therefore, we would like to examine the possible influence of the reference corpus (different register) by means of further analyses in the future.

⁸ For example when set to more than 25 KWs, the application signalizes failure of the database and it stops the whole process.

DOC	TOKENS	KWs
7	3949	20
8	2449	20
9	4537	13
10	2316	4
11	3547	20
12	2306	14
13	3042	25
14	3706	25
15	2860	11
16	3916	25
17	2935	7
18	3874	21
19	2530	21
20	2790	14
MEAN	3134.70	17.15

Tab. 7. The resulting number of KWs in KER

The TF*IDF method generates approximately the same amount of words as analysis of keywords, 340 vs. 300, but the resulting structures of POS differ significantly (compare Tables 6 and 8). TF*IDF actually detects only *nouns* (85%) and *adjectives* (14%), with the exceptions of *hm* (a particle) and *ježiš* (an interjection).

POS	FREQ	FREQ (%)
Noun	291	84.84
Adj	49	14.29
Interj	2	0.58
Part	1	0.29
Total	343	100.00

Tab. 8. Frequency distribution of POS in KER

The results document an important characteristic of TF*IDF: the method truly effectively eliminates all phatic expressions, hesitations, responses, and other phenomena that occur in spoken texts very frequently. In the final list, even certain autosemantic POS are missing, particularly verbs and adverbs. Generally we can conclude that the TF*IDF method appears as the most promising; the extracted words can certainly be considered as thematically relevant, their number is not too high and it needs no reduction (necessary if analysis of keywords is applied). During testing we observed that the results were influenced by the length of the analysed text (the setting of elementary parameters was constant, TF*IDF threshold level + max. number of KWs): the longer the text was the less words appeared in the list of Kws.

4 CONCLUSION

The TF*IDF seems to be a good alternative that can solve or eliminate the drawbacks of respective variant methods. Analysis of KWs generates an extensive list of prominent units that needs reduction while the TC method often results in very short or even empty lists of thematic words. We are aware of the fact that more analyses will have to be carried out, testing more extensive materials and various types of texts (prepared vs. unprepared spoken texts) in order to map out the character of the TF*IDF method and to find optimal settings of the key parameters.

References

- [1] Cvrček, V., and Vondříčka, P. (2013). KWords. Praha. Accessible at: <http://kwords.korpus.cz>.
- [2] Matlach, V., Kubát, M., and Čech, R. (2014). QUITA – Quantitative Text Analyzer. Olomouc. Accessible at: <https://code.google.com/archive/p/oltk/>.
- [3] Libovický, J. (2016). KER – Keyword Extractor. Praha. Accessible at: <https://lindat.mff.cuni.cz/services/ker/>.
- [4] Pořízka, P. (2009). Olomouc Corpus of Spoken Czech: characterization and main features of the project. *Linguistik online*, 38(2), pages 35–43.
- [5] Straka, M., and Straková, J. (2014). MorphoDiTa: Morphological Dictionary and Tagger. Praha. Accessible at: <http://lindat.mff.cuni.cz/services/morphodita/>.
- [6] Šmerk, P. (2009). Majka – Morphological Analysis of Czech. Brno. Accessible at: <https://nlp.fi.muni.cz/czech-morphology-analyser/>.
- [7] Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford.
- [8] Scott, M., and Tribble, Ch. (2006). *Textual Patterns. Key words and Corpus Analysis in Language Education*. Amsterdam – Philadelphia.
- [9] Anthony, L. (2019). *AntConc*. Tokyo. Accessible at: <http://www.laurenceanthony.net/software>.
- [10] Čech, R. (2016). *Tematická koncentrace textu v češtině*. Praha.
- [11] Rajaraman, A., Ullman, J.D. (2011). Data Mining. In Leskovec, J. et al., *Mining of Massive Datasets*. pages 1–17.

IDENTIFICATION OF SPONTANEOUS SPOKEN TEXTS IN SLOVAK

RÓBERT SABO – PETER KRAMMER – JÁN MOJŽIŠ – MARCEL KVASSAY

Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

SABO, Róbert – KRAMMER, Peter – MOJŽIŠ, Ján – KVASSAY, Marcel: Identification of spontaneous spoken texts in Slovak. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 481 – 490.

Abstract: We propose a text classification method for the purpose of creating a language model for automatic recognition of spontaneous spoken speech. Transcripts from our departmental speech database served as spontaneous spoken texts. Using supervised machine learning methods, we have created multiple classification models (including neural networks), that were able to distinguish them from written texts with high accuracy. We subsequently verified the accuracy of our trained models on a database of texts containing direct speech extracted from newspaper articles.

Keywords: spontaneous speech, text classification, supervised machine learning, neural networks, Slovak language

1 INTRODUCTION

The automatic speech recognition technology is currently used in various areas of life. A few years ago, it was used mostly in justice, medicine and automated dialogue systems where the limited domain and established rules of text creation enabled it to achieve high recognition accuracy. Nowadays, thanks to the use of neural networks, automatic speech recognition is introduced into areas where spontaneous speech is used. Automatic recognition of spontaneous speech requires special approaches not only to the acoustic modeling but also to the language modeling [1]. In this paper we focus on the acquisition of text material for training a language model for automatic recognition of spontaneous speech.

In text classification, researchers mostly focus on classification by content, theme, genre, and so on. The use of classification methods to distinguish the style or the form of text (spoken versus written) is less common.

In Slovak, only text categorization techniques (latent Dirichlet allocation) were used in an article by D. Zlacký et al. [2], which led to an increase in accuracy. However, this method is not suitable for our purpose.

In principle, we could use approaches based on human-designed features [3] or those without them [4]. We have opted for defining our own features because we made assumptions about important characteristics of our texts but did not have sufficiently representative data for training the classifier.

The aim of our study is not to create a better language model for the particular task of automatic recognition of spontaneous speech, but to articulate a choice of suitable methods for classifying spontaneous spoken texts and to reveal the typical features of spoken text that could help to identify it.

2 TEXTS ACQUISITION

For our purposes (i.e., creating language models) we have chosen the texts from our departmental text database.

2.1 Spontaneous spoken texts acquisition

For the purpose of creating a model we used a language model for spontaneous speech recognition we have chosen the annotations (annotated transcripts) from our departmental speech database as the spontaneous texts. Their number is quite limited, but other available texts, such as movie subtitles or direct speech extracted from written texts are somehow modified and therefore not fully authentic.

Specifically, we have chosen the annotations of interviews from the portal “100názorov” [5]. The portal collects short interviews (about 10 minutes each) with personalities of cultural and social life on various topics. We created two databases: the database “100n_all” with all the interviews and the database “100n_polit” only with the interviews on the subject of politics.

Special annotation labels for various acoustic events (background noise, hesitations, breaths) were removed from texts but punctuation was left intact for further processing. All annotated segments were linked together and then divided into rows (records), each row containing 160 words, corresponding to the average article length of the “Nový Čas” journal, which was our source of written texts (see next subsection). As a result, the database “100n_all” consisted of 2575 lines (representing 270 interviews) and the database “100n_polit” of 315 lines (33 interviews).

2.2 Written texts acquisition

Written texts were obtained using a standard web crawling technique focusing on the Cas.sk news portal (domestic politics category) [6] as our data source. We installed Web Scraper (<https://www.webscraper.io/>) in our Chrome browser, created a template based on the Cas.sk structure and then extracted the article texts. A total of 7702 articles were obtained, dated between 2019/02/13 and 2013/12/12 (inclusive). The articles contained 160 words on average. Like our spoken texts, these articles have been linked to create a file containing 7702 lines. Both files (spoken texts and news texts) were further preprocessed as detailed in the next section.

3 TEXT PREPROCESSING

Preprocessing was dependent on the type of text. There were 7 different preprocessing steps: 1. surnames removal, 2. abbreviations and numbers removal, 3. names removal, 4. dots, commas and lone letters removal, 5. extra spaces removal, 6. lowercase conversion, 7. lemmatization. For written news article texts, there was one extra preprocessing step: since Web Scraper stores each article in fragments (based on template), these fragments had to be linked together in the right order.

Surnames removal is a dictionary filter containing the surnames of well-known people (e.g. politicians). Abbreviations and number removal was also a dictionary filter, populated with well-known universities and organizations (web addresses were also removed). Names removal filtered out all words starting with a capital letter but not placed at the beginning of a sentence. This preprocessing sometimes produced residual isolated letters with no semantic meaning. Dots, commas and lone letters filter removed all these. Moreover, extra spaces may have been generated by each previous filter that replaced unwanted words with spaces in order to prevent artificial joining of the remaining words and letters. The removal of extra spaces was followed by a lowercase letter conversion and by lemmatization, respectively. Some of these filters relied on regular expressions (like names or spaces removals).

Lemmatization was performed using the online service “Morfologická dezambiguácia” [7]. The service is based on the open source MorphoDiTa tool [8], which combines a tagger, an entity recognizer, and a text analyzer. A local MorphoDiTa client was created and queried via HTTP POST requests.

“**Morfologická dezambiguácia**” service also provided text analysis, which we have used to obtain additional information about the relevant verbs, such as their grammatical person and number.

4 METHOD

4.1 Selection of typical ngrams

Since our goal was to identify spontaneous spoken texts, we used aggregated word frequency statistics from the corpus “Hovor” [9] to define our classification attributes. This corpus contains 6,5 million words (tokens) from different areas and can serve as a representative corpus of Slovak spoken speech.

We used the most common unigrams, bigrams and trigrams from this corpus as candidate indicators of the “spokenness” of text. For each row (record) in our datasets, we calculated how many of its n-grams occurred in the “Hovor” corpus and what was their frequency there (this tells us how typical the n-gram is for spoken speech). In order to avoid potential overfitting of our models through topic-related words, we only used unigrams with corpus frequency of at least 719 (which represent the 600 most frequent words), as well as bigrams with at least 30 occurrences and trigrams

with at least 11 occurrences in the corpus. In what follows we refer to these three categories as “the frequent spoken n-grams” in Slovak.

4.2 Classification models

Experiment 1

The training of our classification models was realized in Weka [10]. We used primarily Radial Basis Functions (RBF), Multi Layer Perceptron (MLP), Support Vector Machine (SVM SMO), Random Forest, and Linear Discriminant Analysis (LIDA), with various parameters and settings. We then used F-measure Score (F1 Score) and Area Under Curve Receiver Operating Characteristic (AUC ROC) as our model accuracy criterion. Finally, 20-fold cross validation was used in order to obtain sufficiently objective accuracy estimates.

In the first phase, classification models were created based only on the following 12 numerical attributes:

- the first 3 attributes (per1sg, per2sg, per3sg) contained the counts (frequencies) of verbs in the singular of the 1st, 2nd and 3rd grammatical person, respectively, in each record;
- the next 3 attributes (per1pl, per2pl, per3pl) contained the corresponding plural counts;
- further 3 attributes (1-gram, 2-gram, 3-gram) contained aggregated n-gram frequencies (n = 1, 2, 3) in each record for the frequent spoken n-grams; and the last 3 attributes (1-count, 2-count, 3-count) contained weighted sums of n-gram frequencies (n = 1, 2, 3) in each record for “the frequent spoken” n-grams, with their frequencies in the corpus serving as weights. In order to make all these attributes mutually comparable, we standardized them.

Our trained models achieved the accuracies listed in Tab. 1. The significance of each input attribute (feature) is shown in Tab. 2, expressed through several alternative metrics used in [12], such as Information Gain (InfoGain), Gain Ratio [11] (GainRatio), Correlation Coefficient (Correl), Chi Square [11] (Chi2), and Signification Evaluation [12] (SignEval). Overall, the most significant attribute was (unsurprisingly) the indicator of the grammatical first person (singular) (“per1sg”), followed by its plural counterpart (“per1pl”) and the count of the frequent spoken bigrams (“2-count”). For more details, see the “Discussion” section.

Model Type	F1 Score	AUC ROC
MLP Classifier	0,987	0,998
Random Forest	0,983	0,997
RBF Classifier	0,965	0,991
Voted Perceptron	0,956	0,936
SVM SMO	0,947	0,919
LIDA Classifier	0,934	0,984

Tab. 1. Comparison of performance for models using the 12 input attributes

Attribute	InfoGain	GainRatio	Chi2	SignEval	Correl
per1sg	0,5016	0,3952	7174,570	0,755	0,6462
per2sg	0,0261	0,0927	426,434	0,406	0,1844
per3sg	0,0551	0,0239	670,714	0,195	0,1892
per1pl	0,2677	0,2240	4105,102	0,655	0,5335
per2pl	0,0127	0,0376	205,771	0,195	0,0846
per3pl	0,0158	0,0101	238,122	0,126	0,1484
1-gram	0,4012	0,1529	4997,830	0,495	0,2130
2-gram	0,4085	0,1360	5312,871	0,500	0,4095
3-gram	0,1433	0,1148	2208,187	0,378	0,4022
1-count	0,3884	0,1554	4846,620	0,475	0,3110
2-count	0,4907	0,1979	6598,433	0,561	0,6217
3-count	0,1431	0,0821	2193,327	0,382	0,3599

Tab. 2. Attribute importance metrics for our 12 numerical features

In the next phase, we tried an alternative approach based on the bag-of-words representation of the texts themselves, which we repeated twice: with and without lemmatization. In Slovak, lemmatization has a profound effect on the size of the vocabulary. In our case the vocabulary was reduced by more than 60% (from 87838 words to 34023). Even after this reduction it was clear, however, that not all words would be significant in distinguishing the two classes. Therefore, we used Principal Component Analysis (PCA [13]) to identify and extract 200 most significant components (linear combinations of individual word representations).

We then trained several different classifiers on these principal components as input attributes. Tables 3 and 4 show the accuracies achieved (with and without lemmatization, respectively). Again, 20-fold cross-validation was used.

Model Type	F1 Score	AUC ROC
MLP Classifier	0,996	1,000
LIDA Classifier	0,995	1,000
SVM SMO	0,996	0,995
RBF Classifier	0,987	0,998
Random Forest	0,981	0,998
Voted Perceptron	0,993	0,993
SVM SMO + RBF Kernel	0,991	0,989

Tab. 3. Comparison of classification model performance with lemmatization

Model Type	F1 Score	AUC ROC
MLP Classifier	0,996	1,000
Voted Perceptron	0,991	0,989

Model Type	F1 Score	AUC ROC
LIDA Classifier	0,985	0,997
SVM SMO	0,988	0,981
Random Forest	0,976	0,997
RBF Classifier	0,937	0,965

Tab. 4. Comparison of classification model performance without lemmatization

From the models we have trained so far, we can summarize the following best settings for each model type:

- RBF Classifier: Number of RBF functions = 8, Tolerance = 1.0e-6, Number of decimal places = 6, using Conjugate Gradient Descent, without using Normalized Basis Functions
- MLP Classifier: Number of hidden units = 8, Ridge = 0.01, Activation Function = Approximate Sigmoid
- SVM SMO: Complexity Parameter = 1.0, epsilon for Round off error = 1.0e-12, Tolerance parameter = 0.001
- Random Forest: Number of iteration = 100, maximal depth = unlimited
- LIDA Classifier: Number of decimal places = 6, Ridge = 1.0e-6

Experiment 2

The results from Experiment 1 (AUC ROC and F1 Score) are definitely encouraging, given that they were achieved in a rather demanding setting, since only political articles were included in Class 1, but a mixture of topics in Class 0. Subsequently, we decided to test the robustness of our models by testing them on data from a different source.

Therefore, in this second experiment, we created a test set with 19631 records. Of these, 16929 records came from a new (SITA) data source that contained direct speech (class 0). Next 2720 test set records were added from Cas.sk (written texts belonging to class 1). As in the 1st experiment, the training set consisted of data from the 100n and Cas.sk datasets (with the number of records from Cas.sk reduced to 5000). Creation of an independent test set allowed us to perform Hold-Out validation in addition to cross validation.

The same model types as in experiment 1 were used for classification on the basis of the 12 attributes listed in Table 2. Validation results for individual models are shown in Table 5.

Model Type	20 Fold Cross Validation		Hold Out Validation			
	F1 Score	AUC	F1 Score	AUC	Precision	Recall
MLP Classifier	0,974	0,992	0,956	0,987	0,961	0,954

Model Type	20 Fold Cross Validation		Hold Out Validation			
	F1 Score	AUC	F1 Score	AUC	Precision	Recall
RBF Classifier	0,970	0,994	0,934	0,987	0,950	0,929
Random Forest	0,981	0,998	0,918	0,988	0,943	0,909
SVM - SMO	0,942	0,930	0,908	0,928	0,937	0,898
LIDA Classifier	0,925	0,983	0,840	0,976	0,918	0,814

Tab. 5. Comparison of classification model performance

5 DISCUSSION

Since our goal in this paper was to identify the style rather than topic, in the preprocessing stage we have removed from our texts all proper names and abbreviations, which carry primarily content-related information. The methods chosen for classification did not take into account the relationship between words (context), thus affecting the semantics (meaning), but in our case, this was an advantage.

In consequence, already in the first experiment, we have achieved encouraging classification results. Using just the 12 numerical input attributes to characterize our input data greatly reduced the computational complexity of our models, and yet the F1 scores of the two best ones (MLP Classifier and Random Forest) surpassed 0.98. Overall, the most influential attributes contributing to this result were the two frequency indicators of the grammatical first person in verbs (singular and plural), followed by the attributes derived from the counts of the frequent spoken unigrams and bigrams.

The significance of each input attribute is listed in Table 2, which also shows relatively lower significance for the indicators of the grammatical 2nd and 3rd person (especially in plural).

Regarding the significance of attributes derived from the frequent spoken n-gram counts, we can see that for unigrams and 3-grams, the non-weighted attributes (1-gram, 3-gram) reach significances similar to their weighted counterparts (1-count, 3-count). Somewhat surprisingly, for bigrams, the 2-count weighted attribute consistently outperformed its non-weighted counterpart (2-gram) across all the monitored significance criteria.

For the classification based on the bag-of-words models, slightly better results were achieved (Tables 3 and 4). The best models achieved the F1 score above 0.99 (both with and without lemmatization). In this case, we used 200 most important components from PCA analysis. Lemmatization reduced the dictionary from 87838 to 34023 words.

By clubbing together all the forms of a given word, lemmatization allows its frequency to be estimated more objectively. On the downside, it also removes the signals of the grammatical person which appeared to be significant in our previous models. However, the latter loss is more than outweighed by other signals of spoken speech that still enabled our “lemmatized” classifiers to perform very well.

An important advantage of the approach based on the 12 numerical attributes is that it achieved similar classification performance as the bag-of-words approach, which required lengthy and complex PCA analysis. Its second advantage is the ease of determining the significance of each attribute.

Of course if we process only one dataset, it could be argued that it would take longer to define those 12 numerical attributes than to blindly run the 200-principal component model without trying to interpret its components. But if we consider that more datasets could be processed through our 12 numerical attributes, their benefits would then multiply.

In all the cases discussed so far, the best results were achieved by the MLP Classifier, while the Random Forest model produced solid results too. Surprisingly, the RBF Classifier achieved somewhat weaker results; in some cases it was even worse than Random Forest.

In the second experiment with SITA data in the test set, there was a more pronounced decline in F1 scores. This was due to the different type of the “spoken” SITA texts (quoted direct speech) than those in the training set (transcripts of spontaneous spoken speech). Table 5 shows this decrease for Hold-out validation with the SITA data. The accuracy is still quite high, however. To identify spontaneous spoken texts in a larger corpus, it is important to have high accuracy. As evident from Table 5 for Hold-Out validation, precision is higher than recall for each method. Overall, the MLP Classifier was the most accurate. Although it was outperformed by 2 models (RBF Classifier, Random Forest) in cross validation, it was the top performer in the more challenging Hold-Out validation.

Overall, the distinctive suitability of the MLP model for this type of task is clear. Another interesting aspect is the way in which it is possible to distinguish spontaneous spoken texts in Slovak. With the 12 numerical input attributes (summarizing the information about the grammatical person, number and n-gram counts) the dominant factor is the first person information obtained from the verbs (a characteristic typical of the Slovak language). In lemmatized text, the grammatical person information is lost as the verb is modified to its basic form. However, that information can still be obtained from personal pronouns remaining in the text. However, the high number of different word forms results in a more extensive dictionary as well as lower frequencies (and thus lower representativeness) of the training set, which could be problematic in some contexts.

6 CONCLUSIONS AND FUTURE WORK

This paper dealt with the distinction between spontaneous spoken and written texts in the Slovak language. In the process, a number of aspects (grammatical person, personal pronouns as well as typical n-grams for spontaneous speech) were revealed as relevant for successful classification. An important factor was the influence of grammatical person, which can be identified from personal pronouns, but in Slovak also from verbs. Therefore, in three distinct cases (lemmatized text, non-lemmatized text, and classification based on 12 numerical attributes), significantly similar results were achieved, with relatively high success. An important advantage of the approach with the 12 numerical attributes is the considerably faster training and classification compared to the 200-principal component model. Its second advantage is the easy determination of the significance of each attribute (since its attributes are not transformed by the PCA).

In our experiments, the MLP classifier achieved the highest accuracy in most cases, which indicated its preferability for this type of classification. Noteworthy results were also achieved by the Random Forest model, given its simplicity and speed of training. Both types of models achieved remarkable F1 scores between 0.97 and 0.99 (verified by 20-fold cross-validation). Their validation on another Hold-Out dataset of a slightly different character (SITA), reduced their accuracy somewhat, but their F1 scores, precision, and recall still remained above 0.95, which we consider a very good result.

As an alternative, we would also consider comparing our approach with pre-trained models such as word2vec or fasttext.

As the next step we plan to create a specialized language model from the texts we have classified as spoken and evaluate the accuracy of automatic speech recognition for spontaneous spoken speech.

In the future, we also plan to use the classification to differentiate the text topic (e.g. sport, politics, religion, etc.). Subsequently, we will distinguish whether the text is written or spoken for each topic separately, which, we hope, will help us develop even more accurate language models for spontaneous speech recognition.

ACKNOWLEDGMENTS

The research leading to the results presented in this paper was supported by grants VEGA 2/0161/18, VEGA 2/0155/19, and U-COMP APVV-17-0619.

References

- [1] Chou, W., and Juang, B. H. (Eds.). (2003). Pattern recognition in speech and language processing. CRC Press.

- [2] Zlacký, D., Staš, J., Juhár, J., and Čižmár, A. (2014). Text categorization with latent Dirichlet allocation. *Journal of electrical and electronics engineering* 7(1), pages 161–164.
- [3] Haddoud, M., Mokhtari, A., Lecroq, T., and Abdeddaïm, S. (2016). Combining supervised term-weighting metrics for SVM text classification with extended term representation. *Knowledge and Information Systems*, 49(3), pages 909–931.
- [4] Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.
- [5] 100 názorov. Accessible at: <http://100nazorov.sk/>
- [6] Politika. Nový čas. FPD Media. Accessible at: <https://www.cas.sk/spravy/politika>
- [7] Garabík, R.: Morfológická dezambiguácia. Accessible at: <https://morphodita.juls.savba.sk/>
- [8] Straková, J., Straka, M., and Hajič, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18.
- [9] Slovenský hovorený korpus – s-hovor-6.0. Bratislava, Jazykovedný ústav Ľ. Štúra SAV 2017. Accessible at: <http://korpus.juls.savba.sk>
- [10] Hall, M., Frank, E., Holmes, G., Pfahringer, B. Reutemann, P. and Witten, I. H. (2016). *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, 11(1), 2009. E. Frank, M. A. Hall, and I. H. Witten: *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Fourth Edition.
- [11] Novaković, J., Strbac, P., and Bulatović, D. (2011). Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms, *Yugoslav Journal of Operations Research* 21, pages 119–135.
- [12] Ahmad, A., and Dey, L. (2004). A feature selection technique for classificatory analysis. Accessible at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.520.6722&rep=rep1&type=pdf>
- [13] Guan, Y., and Dy, J. (2009). Sparse Probabilistic Principal Component Analysis. Accessible at <http://proceedings.mlr.press/v5/yue09a/yue09a.pdf>

AFFORDABLE ANNOTATION OF THE MOBILE APP REVIEWS

MAREK GRÁC – MARKÉTA MASOPUSTOVÁ – MARIE VALÍČKOVÁ
Department of Czech Language, Faculty of Arts, Masaryk University, Brno,
Czech Republic

GRÁC, Marek – MASOPUSTOVÁ, Markéta – VALÍČKOVÁ, Marie: Affordable annotation of the mobile app reviews. *Journal of Linguistics*, 2019, Vol. 70, No 2, pp. 491 – 497.

Abstract: This paper focuses on the use-case study of the annotation of the mobile app reviews from Google Play and Apple Store. These annotations of sentiment polarity were created for later use in the automatic processing based on machine learning. This should solve some of the problems encountered in the previous analyses of the Czech language where data assumptions play a greater role than annotation itself (due to the financial constraints). Our proposal shows that some of the assumptions used for English do not apply to Czech and that it is possible to annotate such data without extensive financing.

Keywords: sentiment polarity, topics analysis, annotation

1 INTRODUCTION

The Internet expansion was followed by various business models, including online stores, applications and other online services. The impact of the user feedback (and virality) might be so significant that it might make the difference between success and failure. In order to process such feedback correctly, it is necessary to monitor discussions and reply to users. In case of a very small user base, it is possible to read every single comment by an expert, but such approach is too naïve for applications with a bigger audience. In such case, it is necessary to do some form of data aggregation which is later processed by an expert. This aggregation might be done manually or automatically. However, automatic supervised methods usually require a non-trivial amount of high-quality annotated data, so manual annotation seems more reasonable as a first step in the majority of the projects.

The issue of niche languages like Czech is that manually annotated data are usually not available and therefore we need to train our model on the similar (available) data. In the previous projects focusing on the Czech language, the models were trained on the data that allow their automatic classification ([1], [2]), e.g. price comparison website Heureka containing reviews of thousands of different products from users who assess positive and negative sides of the product. Such “pre-annotated” data limits the

scope of our research because we have to rely on the existing data or data derived from it. In chapter 4, we will demonstrate that some of these assumptions might not be precise enough in comparison with the annotated data.

In our project, we chose to manually annotate the data by experts and not to rely on the publicly available data. Our work focuses on sentiment polarity (and topics analysis¹). Because these two areas can help businesses understand the needs of customers as much as possible, it is expectable they will prefer an automatic (and also cheaper) solution in the future, but even partial results can be used during the annotating phase. The aim of this article is to provide insights into the annotation and compare our results with previously widely accepted assumptions.

2 STATE OF THE ART

Analysis of the sentiment and particularly the sentiment polarity is a heavily investigated area. The best approaches usually compete in SemEval challenges [3], where the sentiment analysis on Twitter is one of the challenges. The situation of the Czech language is very close to other Slavic languages. There are several proprietary technologies from global and local companies that have never been properly benchmarked and several smaller published projects.

The most notable projects are [2] where the corpus of 10,000 Facebook posts was annotated by two annotators, *Cohen's κ* 0.66 on the document-level annotation. Most of these conflicts were cases of disagreement between neutral and bipolar. Pre-annotated data where sentiment polarity is not explicit are in the order of magnitude larger, but their quality is questionable.

The other important source for the sentiment analysis in Czech is [1], where the polarity sentiment is split into five categories: *negative*, *non-negative*, *neutral*, *non-positive* and *positive*. Even though sentiment analysis can be done on various levels, it was shown that document level text analysis relies heavily on the redundancy and various hints, which is very difficult for automatic analysis. On the sentence-level, the situation is simpler but we need to derive the overall polarity based on the polarity of particular sentences. This derivation, in general, is a complex problem as one single negative sentence can override several of the positive ones. The longer the document, the harder the problem.

In their project, they used three different datasets. The first dataset was compiled from 12 randomly chosen opinion articles from a news server *aktualne.cz*. Two annotators annotated 410 segments of texts (6.868 words and 1.935 unique lemmas) with the result of *Cohen's κ* 0.63. The second dataset was compiled from the reviews of movies on *csfd.cz*. It was created to compare the outcome with the previous one. In total, there were 405 segments with the result of *Cohen's κ* 0.66. The last dataset

¹ However, the topics analysis is not the subject of this article.

was not annotated by the annotators. It was taken from a retail server Mall.cz which also includes user product reviews.

3 DATA AND ANNOTATION PROCESS

In our project, we were working with the reviews of the B2C (business to customers) mobile applications of Czech companies. Data were obtained from public sources from both major platforms: Google Play and Apple Store. Only reviews written in Czech were annotated and they will be used as training data. All reviews were also anonymized as identification of the author is not important for our project. An average review consists of a few sentences only, so texts are relatively short (Figure 1) and the majority focuses on the application itself. It makes our situation quite similar to the analysis of tweets that are heavily investigated by ([4], [5]).

Each review was annotated for sentiment analysis on the sentence-level. Our first attempt was to annotate on the segment-level but we found out that even if we can obtain acceptable inter-annotator agreement (IAA), the process is too expensive for the annotation. Our annotators were able to annotate 250 reviews on the segment-level in 10 hours, while the annotation of the same reviews on the sentence-level was finished in 4 hours. Annotation on the document-level was rejected directly based on the results of [6].

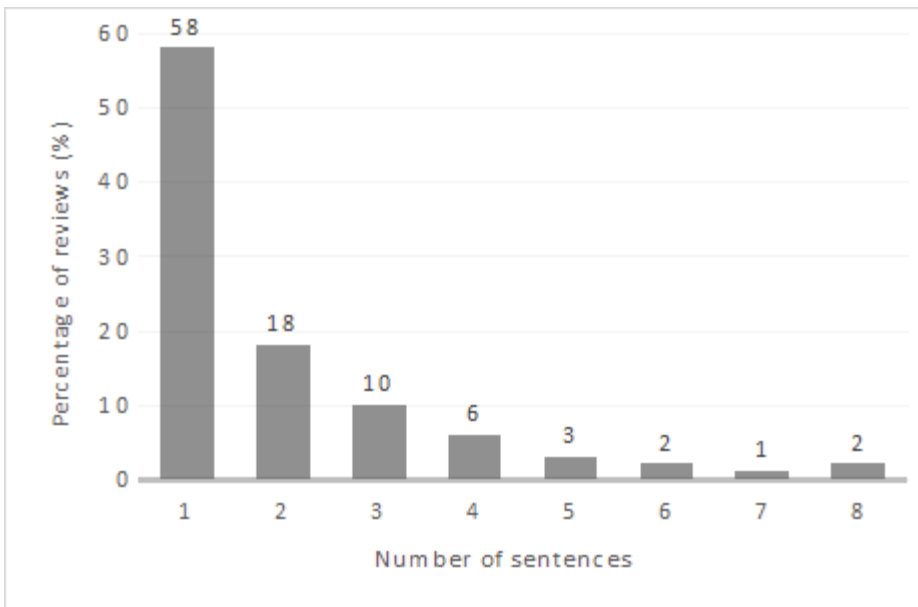


Fig. 1. A number of sentences in reviews

Every review is automatically split into sentences by using the Czech module of Punkt system [7] that can be edited by annotators to fulfil the annotation guidelines. Punkt language module has built-in language model which was used without modifications. Our original assumption was that it will be necessary to retrain the model to suit our domain better. We found out that the majority of the errors end in a situation where sentences are not split correctly. Those situations usually contain symbol ... where it is not clear where sentence boundaries are. Annotators work with each sentence separately, therefore, there is a chance of missing some inter-sentence information when a review is split into too many parts.

Such annotation of the training data is directly usable for business owners even when only part of the data are annotated. Customers can use it almost immediately and it makes the industry more open to funding such research projects.

4 Annotation of the Sentiment Polarity

The reviews are rather short; the majority of them contains only up to three sentences. Annotators annotate each of them with a preferred sentiment. After discussion, we have selected the simplest solution even though we see the benefits of using *non-positive* or *weakly positive* labels. Our primary interest was high consistency (IAA) and annotation cost. These labels were selected: *positive*, *neutral*, *negative* and *mixed*². At the first iteration, we were able to reach *Cohen's κ* 0.58 (on sentence-level), which is close to the published numbers in ([1], [2]). In the confusion matrix in Table 1, the differences between annotators are observable.

	+	±	-	0
+	18	0	0	0
±	0	30	7	1
-	0	4	125	2
0	1	2	35	5

Tab. 1. Inter-annotator agreement

As sentiment analysis is subjective even when annotation guideline is used, the small differences are acceptable. The only relevant issues are 35 (out of 250) differences where the first annotator chose neutral label but the other one preferred negative one. After investigating these cases, we found out that the majority of the problems was caused by very short reviews that contained neutral words only. One of the annotators was annotating what was written, but the other one was annotating what the user meant, e. g. *Reklama* ‘Advertisement’ means that there are too many of the ads in the application. We shifted the annotation to annotate what the user meant even if it is likely to decrease the results of our project. This is a different kind of a problem than reported in the previous papers where discrepancies occur mainly between neutral and mixed polarity.

² Mixed polarity is used when part of the sentence is positive and the other part is negative.

After we have annotated sentences, it is possible to derive sentiment of the whole document/review. It is rather difficult to differentiate the sentiment polarity on a whole document (e.g. on a news article). Thanks to our short reviews, we can select the resulting polarity based on the existence of at least one positive/negative sentences (where mixed sentiment is count as both positive and negative polarity, and neutral sentences are ignored). This approach is not reasonable in longer documents because mixed polarity will prevail as documents usually contain at least one positive and negative sentence. In our case, this does not occur. The number of documents with mixed polarity ranges between 15 and 26 % in each set of annotated data.

In order to manage the labels better, we added one more label: irony, for figurative meanings, irony and sarcasm for the Czech user is rather ironic and sarcastic. Usually, it is not noticeable from just one sentence (e.g. *Je to přesně tak, jak má být. K nejhorsí bance s nejhorsími a nejdrazšími službami patří neodmyslitelně i bezkonkurenčně nejhorsí aplikace.* ‘It’s exactly as it should be. One of the worst banks with the worst and most expensive services has inherently the worst application.’). We annotate these reviews but they are not used for training/testing.

In other projects [1], the author took a huge amount of data from the evaluation sites or online stores. They rely on the quantity of data annotated by random users and take the text written in “plus” section as positive and the text written in “minus” section as negative. However, the reviews often contain irrelevant information, at least in the case of the Czech reviews (e.g. under “minus” section is frequently used word *nic* ‘nothing’ *Nevím.* ‘I don’t know.’ or *Nic mě nenapadá.* ‘Nothing comes on my mind’). It is important to take into consideration that some features can be positive for one product, but negative for the other (e.g. loudness could be beneficial in the case of a mobile phone but unpleasant for a hover). Thus, this is not the best approach.

Another issue of this data is that we rely on the assumption that reviews with a high rating are positive reviews and with a low rating are negative. We found out (Figure 2) that this assumption is valid in the case of negative ratings, but even reviews with a perfect rating (5 stars out of 5) are positive only in approx. 60 % of cases. The rest of the reviews is evenly split between negative, mixed and neutral reviews. It is possible that a similar pattern can be seen also on different data. It could explain a gap between the results of the Czech and English language.

5 CONCLUSION

This article presented the annotation of the sentiment polarity on the reviews of the mobile applications of the given Czech B2C applications. We are annotating more than thousand reviews on a monthly basis and we are happy that we can offer part of them under CC-BY-NC license for other researchers. Those reviews are annotated on a sentence-level for the sentiment polarity and are available on <https://github.com/bedeep/mobile-app-sentiment-data>.

In our future work, we will focus on automatic detection of the sentiment polarity of a text. We plan to re-test existing approaches for the Czech language and re-implement state of the art techniques in the near future. If any of these approaches succeed, the annotators will switch from adding information to the sentences to the validation of proposed information. This should help us obtain more data even cheaper.

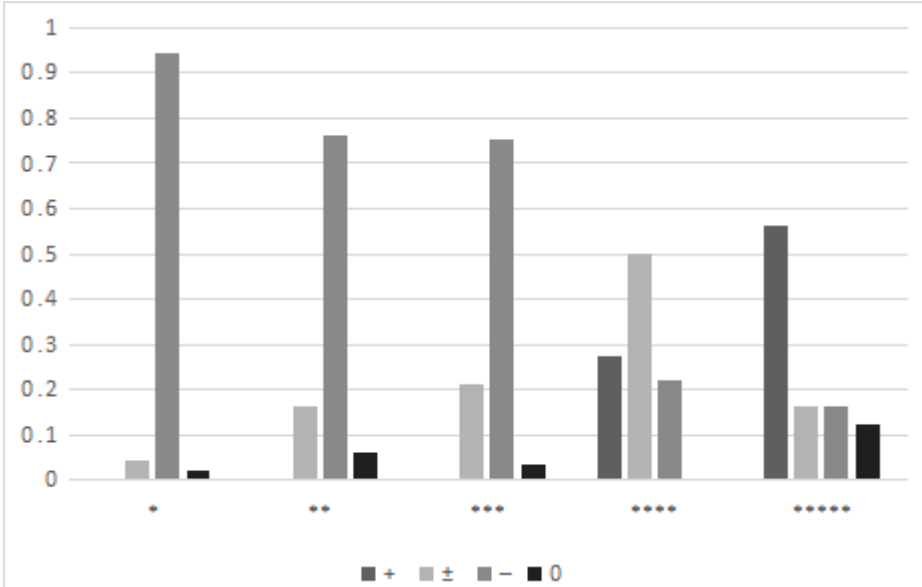


Fig. 2. The relation between rating and sentiment.

References

- [1] Veselovská, K. (2017). Sentiment analysis on Czech. Praha, Karolinum.
- [2] Habernal, I., Ptáček, T., and Steinberger, J. (2013). Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 65–74, Atlanta, Georgia: Association for Computational Linguistics.
- [3] Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 502–518, Vancouver, Canada: Association for Computational Linguistics.
- [4] Anta, A. F., Chiroque, L. N., Morere, P., and Santos, A. (2013). Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques. In Procesamiento del lenguaje natural, pages 45–52.

- [5] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment Analysis of Twitter Data. In Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, Portland, Oregon: Association for Computational Linguistics.
- [6] Wiegand, M., and Dietrich, K. (2009). The role of knowledge-based features in polarity classification at sentence level. In Proceedings of the 22nd International FLAIRS Conference.
- [7] Kiss, T., and Strunk, J. (2006). Unsupervised Multilingual Sentence Boundary Detection. In Computational Linguistics, pages 485–525.