Slovenská akadémia vied
Jazykovedný ústav Ľudovíta Štúra

Computer Treatment of Slavic and East European Languages

Fourth International Seminar
Bratislava, Slovakia, 25–27 October 2007
Proceedings

Editors
Jana Levická
Radovan Garabík

Tribun

Bratislava 2007

Slovak National Corpus
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia 2007
http://korpus.juls.savba.sk/∼slovko/

Table of Contents

Foreword

The fourth edition of the biannual conference SLOVKO, focused on NLP, computational lexicography and terminology, only partially recalls on the founding seminar in 2001 when Slovak computational linguistics was literally in its infancy. The first edition only paid attention to Slovak and Czech and more or less highlighted the 30-year Czech tradition of this field, the second and third one kept enlarging in topics and countries and thus the proceedings of SLOVKO 2007 reflect the international character of this scientific event as it offers a greater and richer scope of the computational processing issues concerning not solely the Slavic languages.

Moreover, papers by Slovak authors, covering basic and applied research, indicate the revolutionary changes that Slovak computational and corpus linguistics have undergone since then. In 2001 the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences in Bratislava created a corpus lingustics department that has already managed to present the sixth version of a fully lemmatised and annotated general corpus containing 350 million tokens. Apart from the general corpus, the Department has been working on partial projects of parallel corpora, terminology database and at present an oral corpus.

The topics of the fourth edition include but were not limited to:

- theoretical issues of computational lexicography and terminography
- bilingual lexicography and terminography
- dictionary structure and dictionary entries compilation
- corpus development: data collection, annotation and processing, word sense disambiguation, co-occurrence analysis and pertinent collocations of lexicographical and terminographical relevance
- new methods in data extraction and terminology mining from corpora
- terminology databases and terminology management systems
- linguistic components of information systems

We would like to thank all authors for their effort and willingness to present results of their research within the SLOVKO 2007 conference. A word of appreciation and gratitude goes also to the program committee for helping to choose, evaluate and correct submitted abstracts. We hope that our conference will enhance fruitful discussions and mutual cooperation as well as projects, whose presentation we will heartily welcome at SLOVKO 2009.

Organisers

Edičná poznámka

Štvrtý ročník bienálnej konferencie SLOVKO, zameraný na NLP, komputačnú lexikografiu a terminológiu, len čiastočne pripomína zakladajúci seminár z roku 2001, keď bola slovenská počítačová lingvistika doslova v plienkach. Prvý ročník mal v centre pozornosti iba slovenčinu a češtinu a viac-menej prezentoval predovšetkým 30-ročnú českú tradíciu tohto odboru, druhý a tretí ročník sa postupne rozširovali tematicky aj teritoriálne tak, že zborník SLOVKO 2007 už odráža medzinárodný charakter podujatia, ponúkajúc širší a bohatší záber počítačového spracovania nielen slovanských jazykov.

Príspevky slovenských autorov, pokrývajúce základný aj aplikovaný výskum, poukazujú na prevratné zmeny, ktoré slovenská počítačová a najmä korpusová lingvistika za ten čas zaznamenala. Od roku 2002 sa v Jazykovednom ústave Ľudovíta Štúra SAV v Bratislave vybudovalo korpusové pracovisko, ktoré už stihlo sprístupniť odbornej verejnosť šiestiu verziu lematizovaného a anotovaného všeobecného korpusu v rozsahu 350 miliónov tokenov. Popri tom oddelenie pracuje na čiastkových projektoch paralelných korpusov, terminologickej databázy a najnovšie aj hovoreného korpusu.

Hlavné témy 4. ročníka konferencie, ktorých sa zahraniční a slovenskí autori mohli pridŕžať, sú:

- teoretické otázky komputačnej lexikografie a terminografie
- bilingválna lexikografia a terminografia
- slovníková štruktúra a tvorba slovníkových hesiel
- tvorba korpusov: zber dát, anotácia a spracovanie, dezambiguácia, kookurenčná analýza a lexikograficky alebo terminograficky relevantné kolokácie
- nové metódy v extrahovaní dát a získavanie terminológie z korpusov
- terminologické databázy a systémy terminologického manažmentu
- lingvistické súčasti informačných systémov

Chceli by sme sa poďakovať všetkým autorom za úsilie a ochotu prezentovať výsledky svojej práce práve na konferencii SLOVKO 2007. Poďakovanie patrí zároveň vedeckému výboru za pomoc pri výbere, hodnotení a korigovaní abstraktov. Veríme, že naša konferencia napomôže vzájomnú diskusiu a nové spolupráce a projekty, ktorých prezentáciu uvítame na stretnutí SLOVKO 2009.

Organizátori

The Possibilities of the Lexicographic Description of Terms in the Lexical Database LEXIKON 21[1]

Edith Birkhahnová and Věra Chudomelová

Institute of the Czech Language of the ASCR, v. v. i.
{birkhahnova,chudomelova}@ujc.cas.cz

Abstract. Electronic treatment of the word stock of a language provides numerous new possibilities, including a more systematic description of technical vocabulary, which makes it possible i. a. to provide a more accurate and complex description of the characteristics of the terms at the levels of the explanation of meaning, encyclopaedic commentary, submeaning, exemplification as well as additional explanation in the form of independent tools, among them a list of the specialised fields and areas in particular.

1 Introduction

In terms of the treatment of terminology in a dictionary in the form of an electronic database, we are at this point still at the very beginning of the conceptual decisions, which must be first preceded by a detailed and precise investigation based chiefly on a sufficient amount of material analyses.

The issue of terminology and its dictionary treatment is considerably complicated, involves a number of unresolved questions, requiring detailed and profound analysis of their theoretical underpinnings, material bases as well as treatment of technical terms in existing monolingual dictionaries of Czech, or also in the dictionaries of other languages (especially Slavonic). It is necessary not only to take a stance on the treatment of terminology in our established lexicographic tradition but also to reflect the new possibilities which the creation of our electronic dictionary database entails (unprecedented in the Czech environment).

In our article, we would like to outline only a few selected component terminological problems which are to be solved progressively: the selection of terms for the list of entries, tools for describing the terms, definition of the terms, place of a technical term in a polysemic entry, usage of the technical synonyms and antonyms.

1 This paper was created within the research plan of the ICL of the ASCR, v. v. i. *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (AV0Z90610521).

In the past Czech lexicographic tradition, especially in the *Slovník spisovného jazyka českého* (Dictionary of the Standard Czech Language) [13], hereinafter as SSJČ, and in the *Slovník spisovné češtiny pro školu a veřejnost* (Dictionary of Standard Czech for Schools and the General Public) [12], hereinafter as SSČ, various discrepancies occurred in selection, designation by field qualifiers and treatment of the terminology of diverse fields, for – as shown by the conducted analyses – especially the terminology of linguistics was thoroughly and systematically treated, unlike the terms of other fields. Our aim is to minimise this disproportional representation of terms.

2 Selection of terms for the list of entries

Currently we are operatively building on the list of entries in the *Frekvenční slovník češtiny* (Frequency Dictionary of Czech) [4], herinafter FSČ. However, we have to take into consideration that frequency on the basis of one corpus is not a sufficient criterion for inclusion in our list of entries. In FSČ for example, we do not find the word *goniometrie*, which is one of the fundamental types of mathematics, the basics of which children learn already at elementary school. For this reason, we will focus on an analysis of the terminology of secondary-school textbooks as well as textbooks for elementary schools according to individual subjects and fields, and the information ascertained will be compared with the representation of the selected technical terms in the Czech National Corpus.

3 Tools for describing the technical terms

Among the tools which we have proposed for the treatment of terms in the form are:

1. A field qualifier, which will be possible to use at various levels of the entry: within the explanation of the meaning, in exemplification, with synonyms, possibly also elsewhere (on a list of field qualifiers, see below);
2. The card 'Význam/Podvýznam' (Meaning/Submeaning), which contains the tools 'Výklad významu' (Explanation of Meaning), 'Poznámka k výkladu významu' (Note on the Explanation of the Meaning), 'Encyklopedická poznámka' (Encyclopaedic Commentary) and 'Předvýklad' (Preliminary Explanation);
3. Exemplification (on exemplification of mini-entries see below).

The tool for the explanation of meaning is separated into the part of the Explanation of Meaning, where a definition understandable for a layperson is written also for technical terms (thus not using terms that are too technical), and the separate Encyclopaedic Commentary, where other data can be complemented.

E.g.: The explanation of the terminological meaning with the headword oak (field: botany) will be resolved with a qualifier within submeaning and the explanation: 'botanický rod stromů *Quercus*' ('botanical genus of the tree *Quercus*')
Encycl. Comm.: rod dvouděložných rostlin z čeledi bukovitých, pův. v Sev. Americe, ve vých. Asii a částečně v Evropě, známo asi 600 druhů (genus of dicotyledonous plants from the beech, or *Fagaceae*, family, coming from North America, East Asia and partly Europe, approximately six hundred species known).

4 List of field qualifiers

In the tools proposed for 'Lexikon 21' (hereinafter only L 21), the list of field qualifiers is hidden under the icon 'OBOR' (FIELD), which is placed on the card for the entire entry, on the card for the individual meaning and finally also in the exemplification tool, and designates the individual exemplification blocks. The first possibility – for the entire entry – will probably not be used with technical terms, because the meanings of the polysemic entries belonging to more fields will not be marked until the level of individual meanings. Thus we have gradually come to the conclusion that the designation by a field qualifier will not be relevant at the level of the entire entry. Another possibility of this division lies in the individual card 'Podvýznam' (Submeaning), which has the same tools as the card 'Význam' (Meaning), so it suffices to label only this submeaning by a field qualifier. We intend to make use of this systematically chiefly with fields that have a nomenclature, i. e. in botany, zoology and chemistry, where e.g. the labelling of genus (for example, the entry *dub* /oak/ – the 1st meaning: '/sturdy/ deciduous tree whose fruit is haycorn'; submeaning*:* 'botanical genus of the tree *Quercus*'), or class (e.g. the entry *savec* /mammal/ – the 1st meaning: 'vertebrate whose young are nourished with milk from the mammary glands'; submeaning: 'the zoological class *Mammalia*') can be dealt with at the level of submeaning. The field qualifier is not used here until at the level of submeaning. The card Submeaning seems to be usable in these cases but needs to be further tested and verified in practice.

We originally designed the list of fields to be hierarchical with the possibility to click on links when at higher levels and thus work one's way through to lower levels (e.g. natural sciences → biology → zoology), an advantage of which could have been a more general placement of a term which is used in more related sciences without the necessity of listing each field individually. However, this approach was demanding on the part of the compilers and the resulting arrangement unclear, so we decided to create only one list, sorted alphabetically. Items which were more general stayed at the same level

as those that had been classified in more detail. Thus the natural sciences, biology as well as botany are now here alongside each other, which provides us with the possibility to choose as needed.

Apart from this, the electronic version allows us to leave the titles of fields untruncated, which is a clear advantage as the full form is clearer and more explicit.

Currently our list has roughly 80 items, and was narrowed on the basis of topicality and relevance of the individual fields for the contemporary users of Czech. The essential fields from the areas of science, technology, economics, defence, law and others are included here. During its creation, we were building on a detailed analysis and comparison of the lists of fields in Czech monolingual dictionaries, both general and specialised (SSČ, SSJČ, *Příruční slovník jazyka českého* (Reference Dictionary of the Czech Language) [11], *Akademický slovník cizích slov* (The Academic Dictionary of Loanwords in Czech) [1], *Nová slova v češtině. Slovník neologismů 1* (New Words in Czech: A Dictionary of Neologisms 1) [8] and *Nová slova v češtině. Slovník neologismů 2* (New Words in Czech: A Dictionary of Neologisms 2) [9]), as well as some encyclopaedic dictionaries (e.g. *Malá ilustrovaná encyklopedie* (Small Illustrated Encyclopaedia) [5]). We further took into consideration also the list of fields in the Czech National Corpus and in foreign dictionaries (for example *Duden Deutsches Universalwörterbuch* [2] and *New Oxford Dictionary of English* [7]), we additionally drew from bilingual or multilingual dictionaries and internet lexical databases. For comparison, we also used other sources of a non-dictionary character, like, for example, lists of international decimal classification systems, which we took into account along with the lists of OKEČ – "Odvětvová klasifikace ekonomických činností" (Sector Classification of Economic Activities) and the list of disciplines within the scientific conception of the ASCR. In the end, not everything became projected into the final list, as some classifications were not suitable for dictionary treatment. We have attempted to compile the fullest list of fields possible, which would however also be simultaneously pliable during the subsequent complementation of the titles of further fields. During the gradual building of the database, it has already been revealed that some fields will need to be added while others possibly renamed, therefore we are still working on the list and elaborating it. Linguistics appeared as first in the list of items in need of complementation, as we had originally only planned on the superordinate philology.

5 Definition of terms

Terms are words which belong to the terminological system of some field, or more fields, but which are simultaneously in most cases also words of the common word stock. They thus often belong to two systems, between which the processes of terminologisation and determinologisation occur.

Therefore, we incline towards explanations that would be as understandable as possible, which will be acceptable for laypersons without deeper knowledge of the given field, while placing possible further information in the Encyclopaedic Commentary or in the Exemplification. All the words which will be used in the explanations of meaning should be inventoried as a component of the metalanguage of description and progressively also processed as headwords.

We will aim for explanations to be understandable for non-specialists, presented in the common language but at the same time precise and providing the modern user with a sufficient amount of technical information. Such properties that are well-known and most prominent for distinction from other words designating similar entities, which are characteristic or motivate metonymy (e.g. with the entry *čočka* /in Czech 'lentil' as well as 'lens'/ the shape of the seed of the plant of such a name is emphasised which became motivating for labelling similar things, for example contact lenses), should be emphasised. When forming a general definition, it is necessary to realise which words designate like things in order to ensure that their explanations differ. For example, we do not consider the explanation shown in SSČ for the headwords *sněženka* (snow-drop) 'a spring plant with snow-white flowers' and *bledule* (snowflake) 'a spring plant with milk-white flowers' as sufficient. For our database L 21, we propose to describe in further detail the appearance and colour of the flowers: for *bledule* – 'a plant flowering in spring with drooping white flowers in the shape of a bell with yellow specks on the edges', whereas for *sněženka* – 'a plant in flowering in spring with drooping snow-white flowers and a green spot in the centre'.

Multi-word terms which will not be treated as separate entries will be possible to show within Exemplification in the form of so-called mini-entries. Mini-entry ('miniheslo') is a separate exemplification block for a fixed terminological phrase with its own explanation and cited evidence:

E3 k V1 MINIHESLO dub letní *Quercus robur*

Odborný výraz v oboru botanika

Syn2000|doc.txtype=PUB,doc.temp=1996,doc.opus=bv-1| Dub letní, zvaný křemelák bývá mohutnější, ale to asi není to nejlepší pravidlo. Oproti dubu zimnímu neboli drnákovi má žaludy na delší stopce.

6 Place of a technical term in a polysemic entry

Within this problem, we proceeded mainly from the treatment of terms in SSJČ and in SSČ (see also FILIPEC 1995, p. 43 [3]) while simultaneously taking the specific possibilities of our lexical database into consideration. On this basis, we separated out several types in terms of the feasibility of their treatment in monosemic and polysemic entries:

6.1 Separate terminological entry

A) Single-meaning

a) If the meaning was considered as genuinely terminological, such entries, e.g. *sčítanec, sarkom, pepsin, perfektivum, fluoreskovat, harmonizovat*, contained the abbreviation of the relevant field/s. Also in L 21, we propose labelling these words as '*odborný výraz*' (technical term) while showing the relevant field.

b) If a headword is not for its being commonly known understood as exclusively terminological, e.g. *cedr, tuleň, beton, reaktor, cyklón*, it is shown in SSJČ and in SSČ without any signalisation of technicality. In L 21, it will be possible to mark potential technical usage by a qualifier within the Exemplification; for animals, plants, chemical substances, etc., it is possible to cover the technical usage in the Encyclopaedic Commentary, or to allot a Submeaning for it.

B) Double-meaning

If a term has two related meanings in various fields, in SSJČ and in SSČ it mostly appears with the relevant abbreviations of the fields and separate explanations, e.g. *drén* – 'tech.' and 'lék.' (med.), *erupce* 'geol.' and 'hvězd.' (astr.). We plan to cover these words as polysemic, labelling them in both meanings as *technical term* and complement with a field qualifier.

6.2 Polysemic headword

A) If none of the meanings was for its being commonly known no longer understood as exclusively terminological, e.g. *fotografie, tarif, sága, raketa*, the terminologicality was not signalled in SSJČ or SSČ. It is possible to treat the technicality in L 21 using the Encyclopaedic Commentary or as Submeaning; it is also possible to indicate the potential technical usage in Exemplification (cf. the word *tarif*).

SSJČ: tarif -u m. (z it. < arab.) *sazba 1; systematicky uspořádaný soubor sazeb, dávek ap. (sazebník) a předpisy pro jejich užívání:* (zaplatit) dopravní, poštovní t.; zvýšit t-y; denní, noční t. (za elektrický proud); *dopr.* pásmový, prostorový, nákladní, osobní t.; *ekon.* časový, úkolový, mzdový t.; – železniční, plavební, celní t.; t. správních, soudních poplatků; *dopr.* místní, průvozní, vnitrostátní, mezinárodní t.

B) If one of the meanings is understood as genuinely terminological, e.g. *mandle, čípek, céva, rak, stereotyp*, it is usually provided in SSJČ and SSČ with the abbreviation of the relevant field. In L 21 this technical meaning will be marked with the label *technical term* while showing the relevant field.

C) Each of the meanings is a term from a different field, e.g. *derivace, rovnice, rapsódie, steeplechase*. Terminologicality in L 21 will be signalled in a similar fashion.

D) Another type is the polysemic headword whose first meaning is non-technical and quite common, whereas the second, terminological, is specified, with the second meaning also covering the first non-technical meaning, e.g. *cukr* (sugar) 1. 'sladidlo' (sweetener) 2. 'chem. *cukry*' (sugars). We find this type of entry especially in the fields of chemistry and botany, e.g. *plíseň*, *houba*, *alkohol*, *sůl*. With both meanings, it will be stated within the note for the explanation of meaning that the second meaning also covers the first meaning.

7 Technical synonyms and antonyms

7.1 Synonyms

Standardised terms and non-standardised correct technical designations were understood in previous dictionaries as mutual synonyms. If a word was explained by a technical synonym which itself is not a technical term, this synonym was provided with the abbreviation marking the relevant field.

In L 21, both synonyms will be explained in the same way within the explanation of meaning and mutually inter-referenced. However, there are also several problematic cases for which a somewhat different approach of treatment is needed:

a) synonyms of the type hruška and hrušeň:
Whereas *hrušeň* (pear tree) only has the meaning 'fruit tree', *hruška* (pear) can be both a 'fruit tree' and the 'fruit of a fruit tree'.

In L 21, we suggest processing the entry *hrušeň* with one meaning ('tree whose fruit are pears'), which will have the submeaning ('botanical genus *Pirus*'); the Exemplification can then show e.g. *hrušeň obecná*, *hrušeň polnička*, etc. The entry *hruška* should then be treated as double-meaning, with the meaning 'tree' being synonymous with the headword *hrušeň* – however not with the above-mentioned submeaning of 'botanical genus' any more.

Another possibility is to treat botanical genus as an independent meaning (see the problem of polysemy above).

b) synonyms of the type *pampeliška* and *smetanka:*
Smetanka used to be the technical term, but the nomenclature has shifted and the official technical term is now only the designation *pampeliška*, which used to be understood and considered as a folk term. In L 21, *smetanka* will be displayed as a synonym in the entry for *pampeliška*, whereas with the entry for

smetanka, *pampeliška* will be shown as a synonym, and the given shift in the terminological usage of the expression will be explained in the Encyclopaedic Commentary in both entries. Since *pampeliška* will have a technical submeaning designating the botanical genus, it will in contrast to *smetanka* be provided with a field qualifier in this submeaning.

c) synonyms of the type *kaštan* and *jírovec, jasmín* and *pustoryl:*

The headwords *kaštan* and *jasmín* are non-technical designations, so they will have *neodborně pro* (common designation for) *jírovec*; *neodborně pro* (common designation for) *pustoryl*, etc. shown in the explanation of meaning.

7.2 Antonyms

With technical expressions, we will in some cases show, besides synonyms, also antonyms after the explanation of the meaning. However, this will only refer to the cases when the given headword forms a complementary (contradictory) pair with its antonym, and thus the shown antonym contributes to specifying the explanation of the meaning, e.g.: *konvexní – konkávní, krytosemenný – nahosemenný*.

References

1. Akademický slovník cizích slov. Academia, Praha (1995)
2. Duden Deutsches Universalwörterbuch. Duden Verlag, Mannheim (2001)
3. Filipec, J.: Teorie a praxe jednojazyčného slovníku výkladového. In: Manuál lexikografie. H&H, Jinočany (1995) 14–49
4. Frekvenční slovník češtiny. Nakladatelství Lidové noviny, Praha (2004)
5. Malá ilustrovaná encyklopedie. Levné knihy, Praha (1999)
6. Němec, I.: K problému slovníkové definice. NŘ 65 (1982) 113–118
7. New Oxford Dictionary of English. Oxford University Press, Oxford (2005)
8. Nová slova v češtině. Slovník neologizmů 1. Academia, Praha (1998)
9. Nová slova v češtině. Slovník neologizmů 2. Academia, Praha (2004)
10. Poštolková, B.: K specifičnosti významu termínů. SaS 41 (1980) 54–56
11. Příruční slovník jazyka českého. Státní pedagogické nakladatelství, Praha (1935–1957)
12. Slovník spisovné češtiny pro školu a veřejnost. Academia, Praha (2005)
13. Slovník spisovného jazyka českého. Nakladatelství Československé akademie věd, Praha (1960–1971)
14. Směrnice pro vypracovávání rukopisu slovníku spisovného jazyka českého. Praha (1957)
15. Zásady zpracování slovníku. In: SSČ. Academia, Praha (1978) 779–799

On Valency of Some Czech Verbs with Multi-Word Prepositions (Based on the Czech National Corpus)

Renáta Blatná

Ústav Českého národního korpusu
renata.blatna@ff.cuni.cz

1 The aim of this paper

Corpus linguistics gives us new possibilities to study esp. syntagmatic relations in language and within this area its main interest is the description of multi-word units, collocations. The multi-word units can be either autosemantic, such as the term *greenhouse effect*, as well as synsemantic ones, such as multi-word prepositions, conjunctions and particles, e. g. the preposition *with respect to*. The aim of this paper is therefore to study the valency of a special area of verbs which are used primarily in the scientific texts and are followed by multi-word prepositions.

2 The conception of valency

As this type of valency was not included in the dictionary Slovesa pro praxi (1997) and neither it is mentioned neither in Czech dictionaries nor in grammars of Czech language it is necessary to start from the lexicological conception of valency described by F. Čermák (1991), who defines valency as the categorial ability of a lexeme to bind one or more formal units. These units can be the parts of speech as well as their subcategories, such as case, preposition, conjunction infinitive, participle, comparative, degree form etc. Among prepositions one-word prepositions, such as *v, na* (in, on), and also multi-word prepositions, e. g. *v rámci* (in the area of), can appear.

3 The analysis of verbs and their co-occurence with multi-word prepositions

The starting point of the analysis was the discovery of the most frequent prepositions in the representative 100-million word corpus of the Czech written language SYN2000. The multi-word prepositions in the frequency zone 16000-5000 occurences were taken into account:

Multi-word preposition	**English equivalent**	**Frequency**
v rámci	within the bounds of	15760
vzhledem k	in view of	14980
spolu s	together with	14788
v případě	in the case of	14653
na základě	on the basis of	13788
do konce	till the end of	11630
v oblasti	in the area of	9540
na rozdíl od	as opposed to	9056
v průběhu	in the course of	8578
v souvislosti s	reffering to	8102
z hlediska	from the standpoint of	7280
v době	in the time/days of	7235
na konci	in the end of	7208
pokud jde o	as for	5757
společně s	together with	5505
ve srovnání s	in comparison with	5194

Table 1.

Then the collocability of these 16 multi-word prepositions was studied, esp. with the respect to the verbal collocates. Among the verbal collocates 30 verbs were found as the most typical for the valency with multi-word prepositions:

Verb	**English equivalent**	**Frequency**	**Occuring multi-word prepositions**
činit	make, cost	16734	*ve srovnání s, vzhledem k, spolu s*
dojít (k něčemu)	take place	34100	*v době, v průběhu, vzhledem k*
dosáhnout	reach	26745	*do konce, ve srovnání s*
dostat	get	69685	*spolu s, v rámci, v průběhu*
existovat	exist	29442	*v případě, v oblasti, v době*
hovořit	speak	16666	*v souvislosti s, na základě*
jednat	act	27624	*na základě, v rámci*
jít	go	124064	*na rozdíl od, v případě*
klesnout	fall	6385	*v průběhu, ve srovnání s*
objevit se	appear	26481	*na základě, v průběhu, v souvislosti s*
patřit	belong to	43371	*společně s, spolu s, z hlediska, vzhledem k*
platit	pay, hold	31528	*do konce, v případě*
pohybovat se	fluctuate, float	12285	*v oblasti, v rámci, vzhledem k*
pokračovat	continue	24383	*do konce, v průběhu, v rámci*
postupovat	proceed	6788	*v případě, na základě, v rámci*
posuzovat	view, judge	3248	*z hlediska, na základě, v rámci*
považovat	consider	33327	*z hlediska, vzhledem k*
pracovat	work	31820	*v oblasti, na základě, společně s, spolu s*

přijít	come	59553	*v době, v důsledku, na základě, v rámci, společně s, spolu s*
připravit	prepare	25970	*do konce, spolu s, společně s, v případě, v souvislosti s*
působit	cause	20890	*v rámci, v oblasti, společně s*
rozhodnout	decide	41460	*do konce, na základě, vzhledem k*
řešit	solve	12687	*v rámci, na základě, společně s, v souvislosti s, v případě*
spolupracovat	cooperate	6543	*v oblasti, v rámci*
stanovit	determine	11028	*na základě, v případě, vzhledem k*
uskutečnit (se)	be realized	13589	*v rámci, do konce, v době, na základě, v průběhu, na konci, v souvislosti s*
vyjít	appear, be published	18787	*do konce, na konci, spolu s*
vzniknout	emerge	21028	*na základě, v době, v důsledku, v souvislosti s, v rámci, v případě, vzhledem k*
zvýšit se	rise, climb	17796	*v důsledku, v průběhu*
žít	live	34817	*na konci, společně s, v době*

Table 2.

It appears that the most frequent multi-word prepositions are listed among the collocates of most of the verbs having a kind of intelectual meaning.

4 Functions of prepositions

All the prepositions – one-word and multi-word ones – have acoording to F. Čermák (1996), three functions (Czech examples and English similar parallels):

1. adverbal, i.e. V – S, e.g. *Vzpomene si na prázdniny* (he returned from holiday)
2. adnominal, i.e. S – S, e.g. *Prázdniny u moře* (holiday by the sea)
3. adverbial, i.e. PROP – S, e.g. *V létě si vzpomene na prázdniny u moře.* (In summer he returned from the holiday by the sea).

The following analysis will be therefore concentrated on the adverbal function of the multi-word prepositions, esp. on 10 verbs which combine primarily with multi-word prepositions of the adverbal function and the verbs combining with multi-word prepositions of adverbial function (i.e. with prepositions *do konce, v průběhu* = till the end of, in the course of) will be excluded, as well as the prepositions with causative meaning, e.g. *v případě* (in the case of) etc.

5 A more detailed analysis of 10 verbs with valency of multi-word prepositions

This detailed analysis will give a list of multi-word prepositions occuring on the first position after the verb (e.g. *posuzovat z hlediska něčeho*), i.e. not with other included words (i.e. *posuzovat něco z hlediska něčeho*), neither with reverse word order (e.g. *z hlediska něčeho posuzovat něco*, *z hlediska něčeho něco posuzovat*) etc. As this valency is usually added to the valency with one-word prepositions, the term additional valency will be used.

5.1 *činit* (make, cost, act), 16734

The verb činit is followed by 13 multi-word prepositions: the most frequent ones are *na základě, v rámci* and *v zájmu*. As the prepositions *na základě* and *v rámci* are on the list of the most frequent ones in this paragraph only the preposition *v zájmu* will be mentioned.

v zájmu	on behalf of	6

This valency follows the primary valency with an accusative or adverb: *činit* + acc. / adv. + *v zájmu* (gen). e.g. *činíme tak v zájmu lidstva* (we act like this on behalf of mankind).

5.2 *hovořit* (talk, speak), 16666

The verb *hovořit* is followed by 6 multi-word prepositions, the most frequent are the following ones:

v souvislosti s	in connection with, in context of	49
ve prospěch	for the benefit	38

The multi-word preposition *v souvislosti s* is an additional valency another to the verb hovořit *o něčem / někom* (to talk of sth./sb.), i.e. *hovořit* + *o* (loc) + *v souvislosti s* (instr), e.g. *hovořit v souvislosti s Hamletem o konjunkturalismu* (to talk of Hamlet in the context of conjuncturalism). The other multi-word preposition *ve prospěch* forms a primary valency (see above), e.g. *verdikt hovoří ve prospěch Pinocheta* (the verdict speaks for the benefit of Pinochet).

5.3 *jednat* (act), 27624

The most typical multi-word prepositions are the following:

v souladu s	in agreement with	52
v rozporu s	in conflict with	45

These multi-word prepositions are the opposites, cf. *jednat v souladu se / v rozporu se zákonem* (to act / to be in agreement with / in conflict with the principle). The valency of these prepositions is compulsory.

5.4 *patřit* (belong to), 43371

The verb *patřit* has the primary valency with dative or with the one-word preposition *k* and dative case form: *patřit* + dat or *patřit* + *k* + dat.

spolu s	together with	58

The most frequent multi-word preposition *spolu s* is the additional valency: *patřit* + (*k*) (dat) + *spolu s* (instr), e.g. *skupina The Offspring patří spolu s Green Day ke špičce tzv . neopunku* (the group Offspring belongs to the tip of so-called neopunk, together with Green Day).

5.5 *posuzovat* (view, judge), 3248

The verb *posuzovat* is combined only with accusative: *posuzovat* + acc.

z hlediska	from the standpoint of	47

The most frequent multi-word preposition *z hlediska* is the additional valency: *posuzovat* + acc + *z hlediska* (gen), e.g. *posuzujeme svět z hlediska našeho života* (we judge the world from the standpoint of our lives).

5.6 *považovat* (consider), 33327

The difference between the valency of the verbs *posuzovat* and *považovat* is that the verb *považovat* has another one-word preposition *za* in addition to the accusative case, hence: *považovat* + acc + *za* (acc). Moreover, this verb has 10 times more frequent than the previous one. The multi-word prepositions form the additional valency. The most frequent are the following ones:

z hlediska	in term of	24
vzhledem k	according to	18

This aspect meaning is clear in the examples, such as *senát považuji vzhledem k nákladům za zbytečný* (I consider senate according to the costs as useless), *zeleninu budeme z hlediska zdraví považovat za důležitou* (we will consider vegetables in term of health as being important).

5.7 *pracovat* (work), 31820

The verb *pracovat* is followed by 37 multi-word prepositions which are partly adverbial in their function. The most frequent multi-word preposition is quite specific and worth mentioning:

na principu	on the principle of	89

This valency is not additional to any case or other one-word preposition, i. e. *pracovat + na principu* (gen), e. g. *přístroj pracuje na principu analýzy stresu v lidském hlase* (the apparatus works on the principle of the analysis of the man voice's stress).

5.8 *řešit* (resolve), 12687

The verb řešit is followed by similar multi-word prepositions as the verb *jednat*, the most frequent preposition is *v souladu s*. Therefore the description will focus on another pair of opposite prepositions:

s ohledem na	in favour of, with regard to	7
bez ohledu na	regardless of	3

The valency of this verb has the structure *řešit* + acc (+ adv) and the valency of multi-word preposition can be added: *s ohledem / bez ohledu na* (acc). The verb is usually in passive voice, e.g. o*tevírání vozu je řešeno s ohledem na bezpečnost* (the car opening is resolved with regard to safety).

5.9 *stanovit* (determine), 11028

The verb *stanovit* is not very frequent and is followed only by 5 multi-word prepositions. The most frequent and the most typical at the same time is the preposition *ve výši*:

ve výši	amounting to	36

The primary valency of this verb is the accusative case: *stanovit* + acc and the additional valency of the multiword preposition *ve výši* (gen). The verb is again mostly in passive voice, e.g. *minimální mzda je stanovena ve výši* 3250,- Kč (the minimum wages is /determined/ amounting to 3250 crowns).

5.10 *zvýšit se* (rise, climb, increase), 17796

This verb has the primary valency with the multi-word prepositions with the meaning 'comparison', esp. the three most frequent ones:

ve srovnání s	in comparison with	30
na rozdíl od	in contrast to	13
v porovnání s	in comparison with	10

The valency structure is the following: *zvýšit se* + o (acc) + *ve srovnání s / v porovnání s* (instr) or *zvýšit se* + na rozdíl od (gen), e.g. *pojistné se zvýší ve srovnání s loňským rokem o 22 procent* (the insurance will rise in comparison with the last year for 22 percent).

Conclusions

In this study it was tested that the multi-word prepositions can be the part of the verbal valency:

- the most frequent multi-word prepositions appear in the context of the most verbs with intelectual meaning
- the analysis of ten intelectual verbs of the frequencies from 3000 to 40000 occurences in the 100-million word corpus SYN2000 showed that at least the most frequent multi-word prepositions following the verbs can be viewed as either the primary or the additional part of the verbal valency:

  činit + acc + *v zájmu* (gen)

  hovořit + *s* (instr) + *o* (loc) + *v souvislosti s* (instr)

  hovořit + *ve prospěch* (gen)

  jednat + *v souladu s* / *v rozporu s* (instr)

  patřit + dat / *k* (dat) + *spolu s* (instr)

  posuzovat + acc + *z hlediska* (gen)

  považovat + acc + *za* (acc) + *z hlediska* (gen) / *vzhledem k* (dat)

  pracovat + *na principu* (gen)

  řešit + acc + *s ohledem na* / *bez ohledu na* (acc)

  stanovit + acc + *ve výši* (gen)

  zvýšit se + *o* (acc) + *ve srovnání s* / *v porovnání s* (instr), *na rozdíl od* (gen)

Appendix: Analyzed verbs with lists of multi-word prepositions

činit

na základě	6
v rámci	6
ve srovnání s	6
v zájmu	6
v souladu s	4
pod vlivem	3
s cílem	3
v rozporu s	3
v oblasti	2
na poli	2
spolu s	2

hovořit

v souvislosti s	49
ve prospěch	38
v případě	8
z pozice	3
ve spojitosti s	3
v rámci	2

jednat

v souladu	52
v rozporu	45
na základě	30
v rámci	17
pod tlakem	13
na úrovni	10
v duchu	10
v prospěch	8
společně s	7
v jménu	6
v smyslu	5
s úmyslem	4
s vědomím	4
na půdě	2
z hlediska	2
v spolupráci	2
bez ohledu	2
v úmyslu	2
bez vědomí	2

patřit

spolu s	58
do rukou	13
v oboru	5
vzhledem k	8
co do	4

posuzovat

z hlediska	47
na základě	13
ve vztahu	5
závisle na	2
společně s	2
s přihlédnutím k	2
spolu s	2

považovat

vzhledem k	18
spolu s	6
v rámci	5
ve srovnánís	3
ve vztahu k	2

pracovat

na principu	89
v oboru	48
v oblasti	48
na základě	33
v rámci	33
ve prospěch	29
na poli	29
pod vedením	23
ve funkci	23
v režimu	23
na bázi	18
v zájmu	11
ve službách	11
na úseku	9
v podmínkách	9
s vědomím	7
po boku	6
v spojení	6
v souladu s	6
v rozsahu	5
na úrovni	5

bez ohledu na	5
na půdě	4
ve spolupráci s	4
pod patronací	4
pod tlakem	3
pod vlivem	3
ve shodě s	3
v měřítku	2
v pásmu	2
s pomocí	2
z důvodů	2
pod jménem	2
ve srovnání	2
v duchu	2
v intencích	2
v rozporu s	2

řešit

v souladu	15
společně s	7
s ohledem na	7
za účasti	4
bez ohledu na	3
v oblasti	2
z hlediska	2
v duchu	2

stanovit

ve výši	36
v souladu	5
v rozmezí	4
bez ohledu na	2
v poměru k	2

zvýšit se

ve srovnání	30
na rozdíl	13
v porovnání	10
v důsledku	9
na základě	5
v souvislosti	4
s účinností	3
na úkor	3
v oboru	2
v rozmezí	2

References

1. Čermák, F., Jazyk a jazykověda. Karolinum, Praha, 2001.
2. Čermák, F., K pojetí valence z hlediska lexikologického. In: Walencja czasownika a problemy leksykografii dwujezycznej, ed. D. Rytel-Kuc, Wydawnictwo polskiej akademii nauk, Wroclaw-Warszawawa-Krakow, 1991, s. 15–40.
3. Čermák, F., Systém, funkce, forma a sémantika českých předložek. Slovo a Slovesnost 57, 30–46, 1996 1997, Czech National Corpus: A Case in Many Contexts, International Journal of Corpus Linguistics Vol. 2, 181-197
4. 2000, Combination, Collocation and Multi-Word Units, In Proceedings of the Ninth Euralex International Congress 2000, eds. U. Heid, S. Evert, E. Lehmann, C. Rohrer, Inst. für maschinelle Sprachverarbeitung Universität Stuttgart, Stuttgart, 489–495
5. Čermák, F., Syntagmatika slovníku: typy lexikálních kombinací. In: Čeština – univerzália a specifika 3, ed. Z. Hladká a P. Karlík, Brno, 2001, 223-232.
6. Čermák, F., Statistické metody hledání frazémů a idiomů v korpusech, In: Studie z korpusové lingvistiky. Kolokace, Nakladatelství Lidové noviny, Praha, 2006, 94–106
7. Čermák, F., Holub, J., Syntagmatika a paradigmatika českého slova. Valence a kolokabilita, Praha, 2005.
8. Čermák et al., Slovník české frazeologie a idiomatiky. Výrazy neslovesné. Academia, Praha, 1988.
9. Kopřivová, M., Valence českých adjektiv. NLN, Praha, 2006
10. Kocek, J., Kopřivová, M., Kučera, K., Český národní korpus. Příručka uživatele. Ústav Českého národního korpusu, Praha, 2000.
11. Svozilová, N., Prouzová, H., Jirsová, A., Slovesa pro praxi. Valenční slovník nejčastějších sloves. Praha, Academia, 1997.

Systemic and Functional Features of the Ukrainian Nouns Category of Number

Tatyana Bobkova

Kiev, Ukraine

Abstract. The number is an inherent feature of the noun reflecting the philosophical category of quantity. The study of systemic and functional features shows that Ukrainian nouns have different statistical characteristics.

Keywords: *grammatical category, number, word, noun, singular, plural.*

The grammatical category of number is one of many manifestations of the general category of Quantity in the language. In general sense the number as a nominal declension category expresses the meaning of "a singular object and more than that" [1, p. 83]. The research of the category of number of the noun presupposes the study of the peculiarities of the origin and evolution of the number, its semantics, structural, functional and stylistic features as well as its relation to the philosophical category of Quantity [2, p. 184].

The research of the structural and functional characteristics of the category of number either dictionary-oriented (the dictionary being the reflection of the units of language) or text-oriented (the text being the product of speech). This research of the category of number of the Ukrainian nouns has been carried out on the official NATO papers with the volume of 20 000 word usage cases.

The micro vocabularies of nouns and their word forms have been compiled for the statistical research of the category of number features [4, p. 8]. The linguistic software ProLing Office 5.0 Complex designed for the Ukrainian and Russian texts processing have been applied in compiling the vocabularies of wordforms and words [3]. The most formal approach has been implemented in the wordforms frequency dictionary compilation. Under this approach a row of letters between two blanks or two punctuation marks is considered a separate wordform [4, p. 6].

As seen from the results of the wordforms frequency dictionary analysis the Ukrainian nouns are characterized by a high absolute frequency which makes up 8051 wordforms for 20 000 word usage cases, that means 40.26% from all wordforms of the text (go to the fragment of the dictionary in Table 1 which includes ten frequent forms of the Ukrainian nouns). The results of the official papers text analysis correlate with those of fiction text analysis in the Ukrainian: in all types of texts (with the exception of the drama) "among the most frequently used words nouns prevail" [4, p. 11].

Field1	Count-Field1	Field3
альянсу	250	альянс_
країн	138	країн_а
безпеки	132	безпек_а
року	127	рік_
членів	98	член_
сил	84	сил_а
співпраці	70	співпрац_я
миру	69	мир_
партнерства	69	партнерств_о
країни	48	країн_а
сили	48	сил_а
планування	47	плануванн_я

Table 1. number_noun_wordform

The complete list of wordforms includes 1881 noun wordform. The average frequency of nouns makes up 198.91 wordform per 500 text units. Every of the ten most frequent forms of Ukrainian nouns has the frequency of 47 and higher and the general absolute frequency of these words makes up 14.66% of 8051.

So the highest frequency (250) is that of the *альянс* wordform in the genitive case singular – *альянсу*, which makes up only 0.05% of the total number of all wordforms and 3.1% of their frequencies. Three noun wordforms: *країн, безпеки, року* are used more than a hundred times (138, 132, 127 respectively), which makes up 0.16% of the total number of all wordforms and 4.93% of their frequencies. The usage of 184 noun forms is rated at in tens (more than 10 and less than 100) and makes up 9.78% of the total number of all wordforms in the texts. From the list under analysis 960 forms are used more than once, and that is 51.03% of the total number of the wordforms. There are 921 form with the frequency of 1 which make 48.96% of the total number of the noun wordforms.

Thus the results of the Ukrainian texts of the official NATO papers prove the tendency of the increase in the number of text units with the decrease in the frequency of their use: “as a rule, the words with the frequency of 1 make up about a half of the total number of different words in the frequency dictionary” [4, p. 12].

An important problem one faces when applying the above described formal approach is lexical and grammatical homonymy. The homonymy phenomenon distorts the data of morphological forms frequency and that of grammatical categories as well in the texts under analysis. To withdraw the homonymy, a frequency micro dictionary of nouns has been compiled, the entries of the

dictionary are organized in a slot way and include every wordform of a certain noun used in the text.

The data analysis for the frequency dictionary shows that in the Ukrainian texts under analysis there are 932 nouns per 20 000 word usage cases. And according to the Table 2 data, the list of the ten most frequent words undergoes slight changes.

Field3	Count-Field3
альянс_	315
країн_а	264
рік_	178
безпек_а	168
член_	168
сил_а	161
партнерств_о	108
партнер_	97
співпрац_я	93
мир_	76
Європ_а	70

Table 2. number_noun_lems

Each of the ten most frequent Ukrainian nouns has the frequency of 70 and higher, and the total absolute frequency of these words constitutes 21.09% out of 8051. The highest frequency (315) is characteristic of the nouns *альянс* and *країна* (264), which makes up as little as 0.21% of all the nouns. Five of the ten most frequent nouns *рік, безпека, член, сила, партнерство* are used more than a hundred times in the texts which makes up 0.54% of the total number of nouns. 178 nouns under research have the frequency of more than 10 which is 19.1% of all words. 398 words from the analyzed list are used more than once and that is 42.7% of the total number of nouns. The words with the frequency of 1, and there are 349 of them in the frequency dictionary make up as much as 37.45 % of all the nouns.

It goes without saying that the meaning of the number of nouns cannot always be narrowed down to expressing singular and plural as "they have these meanings when it comes to the names of singular, discrete objects, i.e. countable objects" [1, p. 84]. In those cases when the number forms of nouns "do not have the function of quantity actualization" [1, p. 85] the binary structure of the category of number based on contrasting the singular and the plural undergoes changes. That is most proper to the Singularia and Pluralia Tantum nouns.

The results of the realization of the number forms in the texts enable to point out the following main groups of nouns:

1) Nouns that have both singular and plural forms.
2) Pluralia Tantum.
3) Singularia Tantum.

The peculiar character of the official NATO papers requires grouping the the nouns of the following kinds into separate groups:

4) Proper Names
5) Proper Names, Pluralia Tantum

The ULIS (УЛІС) bilingual electronic dictionary that is included into the ProLing Office software served the basis for figuring out the nouns that have both singular and plural forms from the Pluralia Tantum nouns (for example: *верх–верхи*, *курс–курси* and *інтерес–інтереси*, *сила–сили* respectively) [3].

Quantity distribution for the nouns of different groups for a hundred entries of the frequency micro dictionary is given in Table 3.

Hundred Number	**Nouns That Have Both Singular and Plural**	**Pluralia Tantum**	**Singularia Tantum**	**Proper Names**	**Proper Names, Pluralia Tantum**
1	64	3	13	19	1
2	72	4	17	7	
3	78		19	3	
4	72	1	14	13	
5	59	1	27	12	1
6	85	2	9	3	
7	88	1	11		
8	71		21	8	
9	66	2	22	10	
10	27		10	5	
Total	**682**	**14**	**163**	**80**	**3**

Table 3.

Just as it could be expected, the most numerous is the group of nouns that have both singular and plural forms – 73.18% of the total number of nouns. Next coming in quantity is the Singularia Tantum group nouns – 163 words, which makes up 17.66% of all the nouns. The great number of the nouns from the Proper Names group – 80 (8.67% out of the total number of nouns) can be

put down to the specific character of the official NATO papers text. The smallest group in number is Pluralia Tantum – 14 nouns constitute 1.52%.

As can be assumed from the official NATO papers text analysis the most frequent noun that has both singular and plural forms is the noun *альянс* (315), the most frequent Singulare Tantum noun is *безпека* (168) and the most frequent Plurale Tantum is *відносини* (29).

The greatest interest presents the most numerous group nouns that have both singular and plural forms. The analysis of the functional features of these nouns shows that the better part of them – 261 out of 682 (38.27%) are used in the texts of the papers only in singular. They are mainly the nouns denoting process, abstract notions and names of actions, for example: *бажання, бачення, бік, вдосконалення, вимір, вияв* etc. It should be noted that 68 – 9.97% of the total number of nouns that have both singular and plural forms are more frequently used in singular, for example: *акта* (a form in the Genitive case), *винятком* (a form in the Instrumental case), *війни* (a form in the Genitive case), *році* (a form in the Dative case). This can be explained by the high frequency of word-combination with government links in which the above mentioned nouns are in subordination.

In the group of nouns under analysis 131 words are used only in plural and that constitutes 19.2% out of the total number of the nouns that have both singular and plural forms. It should be noted that 85 – 12.46% of the words under analysis are more frequently used in plural, for example: *арсеналів* (a form in the Genitive case), *біженців* (a form in the Genitive case), *боєголовок* (a form in the Genitive case), *викликам* (a form in the Dative case), *держав* (a form in the Genitive case), *інтересам* (a form in the Dative case), etc. The results of the analysis show that in this subgroup the words denoting objects prevail. Thus in the group of the nouns under analysis those nouns used in the papers only in singular prevail, which is twice larger than the number of the nouns used in plural. Though the frequency of the noun plural form slightly (by 2.49%) exceeds the frequency of the noun singular form.

The results of systemic and functional features research of the Ukrainian nouns category of number based on the official papers texts have enabled the assumption that the realization of the category of number of nouns in the texts depends on the world picture created by its speakers. This explains both the high frequency of nouns that have both singular and plural and denote single, discrete objects and the low frequency of the Pluralia Tantum nouns.

References

1. Введение в сравнительную типологию английского, русского и украинского языков. – К.: Вища школа, 1977. 148 с.
2. Перебейнос В. И., Бобкова Т. В. Типология категории числа имени существительного (на материале русского, украинского и английского языков)// Труды и материалы III Международного конгресса исследователей русского языка „Русский язык: исторические судьбы и современность". – М., 2007. – С. 184–185.
3. ProLing Office 5.0, РУТА 5.0.
4. Перебийніс В. С., Муравицька М. П., Дарчук Н. П. Частотні словники та їх використання. – К.: Наукова думка, 1985. 204 с.

The Text Corpus and Dictionary Hierarchy

Natalia Darchuk and Viktor Sorokin

National Taras Shevchenko University of Kyiv
compling@uniling.kiev.ua

It is quite evident that for the last decades experts in computer text analysis need immensely precise functional characteristics of language units in texts of different types. For carrying out of theoretical and applied researches (i.e. machine translation, language analysis and synthesis, and text abstracting) as well as didactic studies the specialists in the sphere of cognitive linguistics, semasiology, functional grammar and word-formation do not have general classified data on principles of functioning of language units in speech. Due to this problem the necessity to develop the corpus linguistics is becoming more obvious. It will help the specialists to receive all necessary linguistic information and to apply it for further data processing in diverse philological studies.

When researching the corpus linguistics issues a group of professor-linguists from the Institute of Philology within the National Taras Shevchenko University of Kyiv decided to build text corpora that enable the specialists to receive all necessary information about single language units and their abilities. The text corpora by means of special tools for their processing are the source of building of electronic card catalogues, at the same time text processors enable not only to receive different information from the text corpus but also to construct electronic cards from lexical catalogues by set parameters for compilation of dictionaries of any sort.

The proposed concept of corpus text processing consists in *creation of principles of formalized description of language units of different levels*. Using these principles there were developed computer tools which provide extraction of specific information from texts and compilation of dictionaries of different kind such as frequency and grammar ones, dictionaries of morph structures, dictionary of roots and affixes, dictionary of roots family, dictionary of syntactic models of word-combinations and sentences. They serve for building of electronic card catalogues that can be used for different linguistic studies.

All above mentioned tasks can be solved successfully by using of the automated system for processing of Ukrainian texts which based on a line of linguistic lexicographical data bases. So it means that the structure of this system is based on module-type lexicographical ideology where system and structural relations of every level of language system is presented in single module such as morphological, morpheme, word-formative, syntactical and semantic ones.

The basic unit of each module is a word, for the following reason:

- the word is a central unit of language system which is defined by different structure types on its all levels;
- the word belongs to the units of high levels of language system of word-combination and sentences.

Each module is comprised of two blocks: 1) dictionary-block: the word inventory of each module is formed on basis of 11 volumes of Ukrainian Dictionary and which during the work on text corpus is updated with new words. For this case a special working dictionary is compiled where each word is provided with linguistic description, in particular, morph structure, derivate, explanation, quotation from text with relevant external description (totally the dictionary contains about 190 000 words); 2) block-analyzer which is considered a linguistic tool for completing of certain tasks in the text field (i.e. morpheme, morphological syntactical semantic ones).

The **Morphological module** is a central one because the other modules process text information through identification of it with the grammar-morphological code. After processing the text undergoes indexing where every word form is assigned with morphological information about part of speech and their lexical and grammar categories as follows.

- for gender, number, case;
- for verbs: tense, aspect, mood, voice, person, gender number, transitivity/intransitivity;
- for prepositions relationships with noun cases
- for conjunction: categories in respect to the function

Homonymy both lexical-grammatical and grammatical one is defined more exactly in this module through the elements of positional syntactical analysis.

Morphological module enables to compile dictionary-concordance for a certain word-form and a certain grammatical meaning.

The working of **syntactical module** is provided by the list of syntaxemes which is used for building of a dependency tree of a sentence. The dependency tree describes formal syntactical construction of simple and complex sentences only when the text is supplied with morphological information after the automated morphological analysis.

Applying of this module within automated syntactical analysis enables as follows:

- to research the functioning of morphological model of word-combination;
- to determine the structure of text, paragraph, and number of single and complex sentences (a paragraph is marked on the stage of the pre-

morphological analysis under the number of intervals between sentences, as well as the beginning and the end of a sentences).

- to build the list of syntactical-semantical frames as transitional stage to semantical text analysis. The frame as a basic unit of the list of syntactical-semantical frames is presented in the form of a pattern which is comprised of:
- preset and filled with certain units positions;
- empty positions that can be filled in compliance with three principles: selective compatibility; information capacity, and monosemy.

The **Semantical module** is based on thesaurus-dictionary where the words are systematized not on the lexical principle but using the principle of presentation of conceptual relationship – beginning from the concept till lexical means of its expression which in combination with certain content variations denote one meaning. The ideographic principle was used not occasionally as the following classification enables:

- to formalize description of structural relations of lexical system through synonymic, antonymic and polysemic relations;
- to determine the location of any lexical-semantical group and any lexeme within ideographic classification;
- to receive the formalized description of a word semantics through the sum of concepts.

The linguistic analyzer of this module builds the text thesaurus automatically by means of combination of general thesaurus with the list (incl. absolute frequency) compiled automatically from selected texts. Connection of electronic explanatory dictionary provides more precise and specific meanings in a text which is analyzed.

As a result of working of a semantical module linguists can receive the following information from the text:

- word semantics under the logical-conceptual synoptical scheme;
- peculiarities of usage of a polysemical word with certain meanings in different types of discourses,

as well as:

- classify words on the basis lexical-semantical groups;
- specify and describe different figurative meaning of words.

The morpheme module provides the automated segmentation of the input text into the morphs. The segmenting procedure is based on the list of word-form which are divided into morphs where the morph structure of word-forms represents structural und functional relations of morphs in a word. The presenting of any morph structure of a word-form in a linguistic model which specifies the boundaries and type of certain morph enables automatically to describe any morph structure through the program procedure, for example,

за студ и ти P2R6S7F9.

The morph structure which is built automatically by the program procedure provides with full detailed linguistic information about the morph, its structural relations with other morphs and is defined as a working unit of morpheme list.

The models of morph structures of words/word-forms enabled to create the automated system of linguistic analysis which performs following tasks automatically:

- to sort words into single-roots and single-affixes classes;
- to classify words upon morph-quantitative models;
- to compile lists of alphabetically arranged frequency, single-root and single-affix words;
- to accomplish morph segmentation of text word-forms and to receive lexicographic description of certain texts on the morph level.

The **Word-formative module** is based on word-formative list where all words are grouped into word-formative families. The words from word-formative list under support of morpheme list are grouped automatically into the samples with single-root words and then automatically classified within the frame of each single-root class according the quantitative-morph models.

Each sample consisting of single-root words builds a single field within the frame of word-formative module where the morph structures of words are classified according to quantitative-morph models. Using principles of derivation there were developed a technology of description of word-formative relations between motivating and motivated words that enable to build automatically the working model-hypothesis of word-formative family.

The classification of words from single-root sample requires checking which is carried out automatically in the three ways: 1) checking of correct grouping of words into the samples of single-root words, 2) determination of word-formative stem and word-formative formant in a derived word; 3) description of morphonological processes taking place at every stage of word-formation.

So, word-formative module enables to analyze word usage in texts considering following aspects:

- the means of formation and word-formative structure;
- word location in the word-formative family of words;
- classification of words by methods of formation, word-formative meanings;
- determination of neologisms and occasionalisms in texts;
- extension of electronic base word list with new words.

The technology of corpus text processing is the result of theoretical and practical methods in modern linguistics. This technology of construction of modules for text data processing becomes very efficient tool (it saves lots of time and human resources) for specialist-linguists of different specialization. It helps to carry out comprehensive linguistic studies quickly and at high quality level. The systematized in a series of dictionaries information is quite important for grammatical, stylistic, literature and semantic studies of Ukrainian texts of different discourses (from poetic till scientific ones) in its dynamics.

Collocations in Slovak (Based on the Slovak National Corpus)

Peter Ďurčo

University of St. Cyril and Methodius
Trnava, Slovakia
durco@vronk.net

Research on bound word combinations is one of the key tasks in the research of word stock. It focuses on the basic research dealing with word formation, lexicology, syntax and corpus linguistics. The results have direct usage in lexicography, translation theory and practise and didactics of (foreign) language teaching.

The research of collocations or set expressions has got a long tradition in Slovakia especially from a theoretical point of view (cf. Mlacek 2001, Jarošová 2000 a,b,c), in lexicography it was an accompanying phenomenon which has explained the meanings of the words. However, a brand new dictionary „Slovník súčasného slovenského jazyka“ depicts and describes intentionally collocation features of the words and uses the corpus data, the first analyses have shown that even the academic dictionary entails an area of collocations selectively and the compilation of a special dictionary of collocations is fully legitimate. Not only for a precise recording of the collocation paradigm and its forms, but it represents an independent type of recording the word stock. In addition, the research on collocability of the words is an important assumption for a contrastive description.

In communication, we do not use only individual words in order to create the sentences or texts which make sense, but to a high degree we use certain combinations of words or fixed combinations of words (collocations). They are important building units in every language. Combinatorical features of the words have been stored by the language users in an individual semantic memory and are activated spontaneously in the forms of word associations.

The combinatorial features of the language units, their cooccurrence potential, their distributive features and their collocation radii are specific in every language. They entail for every unit an arbitrary bound multivergent matrix which, in its complexity, includes hardly recognizable and unpredictable relations. It causes the biggest problems in acquiring and using foreign languages.

A topic on collocations has been nowadays one of the main topics of lexicography and nevertheless the most studied theoretical term in the corpus lexicography. Present enormous quantitative and qualitative development of the corpus linguistics, e.g. a notable quantitative increase in digitalized texts in corpora, their different ways of marking and the existence of strong statistical

tools for extracting and evaluating data, open completely new ways of describing linguistic material. As the most distinctive trend in the corpus linguistics it can be pointed out the need for the development of new methods and tools for recognizing significant cooccurrences of words, their collocability and compatibility.

The intricacy of the examined topic has been caused by difficulties in differentiating the sets of collocations in the scale from a simple, however common and textually high frequent neighbourhood of words without apparent or predictable bounds between them, that do not have to be significant from the point of frequency distribution, but they create a structural or semantic unity. We think, in spite of the fact that frequency and stability are connected phenomena, that instead of the conception of distinguishing a rank of the collocations at the scale, we should come out from the model of the radius structure of words and their collocations as a diffusion and confluent set of elements. There is a centre and a periphery, which can be changed according to the quantitative source field (the cooccurrence and frequency distribution of the elements) or to the qualitative source field (the semantic unity of the combination).

The project on Slovak collocations will be aimed at registration and description of not only multi-word lexemes and phrasemes, but also at registration of the so-called typical collocations, which have a wide collocability, they are frequentially differentiated and so limited in that way. Irregular systematic collocations (idioms, phrasemes), regular text-collocations (*zimná rekreácia*) and fixed text-systematic collocations (*krájať nadrobno*, *hovoriť úsečne*, *vystúpiť z auta*) will be described. The systematic terminological and proprial collocations (e.g. *difúzna množina*, *pravý uhol*; *Vysoké Tatry*) will be excluded from the description (cf. the classification of collocations by Čermák 2006a).

From the material point of view the research will be based on the Slovak National Corpus with 350 million text words at present time. The basic glossary of described words will be based on frequency criteria, it means that the most frequent words will be examined.

From a linguistic point of view, it is important the fact that collocations are not primarily bound and generalized to an abstract term “lexeme” or “lemma!”, but primarily to the paradigmatic forms of words, even in a different way according to certain meanings.

As a methodological base, the project will use methods and tools of the corpus linguistics, especially concordances, frequency distributions of the words and cooccurrence analyses. Due to the large amount of existing collocations and combinations of words in the corpus, we will use mathematical methods, which can recognize word cooccurrences using statistical models, integrated in searching tool of the corpus manager Bonito. Apart from evaluating the frequency distribution of the words (with the basic frame -2 -1 0 1 2) which recognize frequency of the words at a set interval, they will produce an index of

the mutual information (Mi-score), which measures the strength between two words according to the ratio of the occurrence probability between two words and occurence of each of the two words independently. This test is suitable for recognizing less frequent collocations. Then the contrast score (t-score) will be measured. It is based on the discrepancy between observed and assumed results. The test recognizes the degree of collocability of the elements according to the absolute occurrence of them in the corpus. These data represent the initial point for the creation of the collocation database in Slovak language and a would be a base for the contrastive description as well.

Despite the new language technological analysis scepticism still prevails regarding the possibility of seizing and of describing the data completely. This scepticism results particularly from two problems. Word combinations represent a difuse continuum of semantically differently strong connected elements. The borders between "free" and "firm" can not be specified clearly. According Hausmann collocations are typical, specific and characteristic combinations of words, which exhibit lexical selection restrictions (Hausmann 1985, 2004). The question is, what is typicall and what not? On the other hand by statistical approach the main problem is, that the frequency and stability of collocations are not directly connected correlations. Not all high frequent word combinations are also firm. One finds typical collocations in all ranks of the frequency distribution (cf. Čermák 2006a,b,c; Heyer et al 2006). The number of significant cooccurrences rises also with the absolute frequency of the word in the corpus.

Sorting of collocations

Startig point here is a binary combination between basis and collocator, where the basis is presented through a word form and collocators are the main parts of speech – nouns (N), verbs (V), adjectives (A) and adverb (D). All other longer structures are brought to a generalised binary structure. The combinatorial potentials of these elements are the basis for the creation of so called collocational templates which are basis for the patterns of collocations. By nouns the basic model for patterns is like this: N.* [N.* | V.* | A.* | D.*]

Example for noun as the basis:

Number	Case	Collocator	Example
SINGULAR			
PATTERN	NOMINATIVE		
Attr + Sub1Nom		ATTRIBUTE	akademický **rok**;
Sub1Nom + Sub2			bezúhonný **charakter**;
Verb + Sub1Nom			psí **čas**;
		NOUN	**charakter** človeka;
			čas a priestor;
			mrcha **čas**;
			rok basy;
			rok koňa;
		VERB	byť **charakter**;
			to je **charakter**!;
			charakter vytvára niekoho, niečo;
			blíži sa **čas** niečoho;
			čas je drahý;
			rok plynie;
			písal sa **rok** ...
	GENITIVE		
Attr + Sub1Gen		ATTRIBUTE	od toho **času**;
Sub2 + Sub1Gen			onoho **času**;
Verb + Sub1Gen			svojho **času**;
			do(jedného)**roka**;
			od budúceho **roka/-u**;
		NOUN	črta **charakteru**;
			centrum voľného **času**;
			postupom **času**;
			pre krátkosť **času**;
			do **roka** a do dňa;
			na prelome **roka**;
		VERB	je to vecou **charakteru**;
			dočkaj **času** (ako hus klasu);
			kráčať s duchom **času**;
			niet **času** (na niečo);
			tráviť veľa **času**;
			narodiť sa **roku** ...
	DATIVE		
Attr + Sub1Dat		ATTRIBUTE	oproti minulému **roku**;

Number	Case	Collocator	Example
Sub2 + Sub1Dat			oproti predchádzajúcemu **roku**;
Verb + Sub1Dat		NOUN	ku **koncu** roka;
		VERB	vyhovovať charakteru niečoho;
			zodpovedať **charakteru** niečoho;
			byť k **dispozícii** niekomu;
	ACCUSATIVE		
Attr + Sub1Aku		ATTRIBUTE	na dlhý **čas**;
Sub2 + Sub1Aku			náročný na **čas**;
Verb + Sub1Aku			po celý **čas**;
			v pravý **čas**;
			celý (boží) **rok**;
			raz za turecký **rok**;
			rok nato;
			takto **rok**;
		NOUN	rok čo **rok**;
			rok po **roku**;
		VERB	mať **charakter**;
			mať **charakter** niečoho (odporúčania / ...);
			mať rovnaký / podobný **charakter**;
			dožičiť **čas** niekomu;
			mrhať **čas**;
			dovŕšiť (…) **rok**;
			prežiť **rok**;
	LOCAL		
Attr + Sub1Lok		ATTRIBUTE	po (dlhom / istom / nejakom / pracovnom / určitom) **čase**;
Sub2 + Sub1Lok			v krátkom **čase**;
Verb + Sub1Lok			v ostatnom **čase**;
			v tom istom **roku**;
		NOUN	v **čase** i nečase
			v **čase** od deviatej do jednej;
			v **čase** niečoho (detstva / (najväčšej / rannej) dopravnej špičky …;
		VERB	narodiť sa v **roku** …
			po **čase** sa uvidí;
			pracovať po **čase**;

Number	Case	Collocator	Example
	INSTRUMENTAL		
Attr + Sub1Ins		ATTRIBUTE	svojim **charakterom** (svedčať / pripomínať / ...);
Sub2 + Sub1Ins			overený **časom**;
Verb + Sub1Ins			pred (nejakým) **časom**;
			každým **rokom**;
		NOUN	(zamračené a) **časom** dažď;
			v porovnaní s (uplynulým) **rokom**;
		VERB	byť daný **charakterom** niečoho;
			odlišovať sa **charakterom**;
			časom vybledne niečo;
			časom sa ukáže;
PLURAL PATTERN	NOMINATIVE		
Attr + Sub1Nom		ATTRIBUTE	rozdielne **charaktery**;
Sub1Nom + Sub2			staré (zlaté) **časy**;
			študentské **roky**;
Sub1Nom + Verb		NOUN	**charaktery** ľudí;
			charaktery postáv;
			časy zašlej slávy;
			roky dospievania;
		VERB	bývali **časy**;
			časy sa menia;
			tie **časy** sú (už dávno) preč
	GENITIVE		
Attr + Sub1Gen		ATTRIBUTE	do (tých / dnešných / detských) **čias**;
Sub2 + Sub1Gen			za starých (dobrých) **čias**;
Verb + Sub1Gen			od útlych **rokov**;
		NOUN	prepracovanosť **charakterov**;
			do skonania **časov**;
			pozostatky z **čias** niečoho;
		VERB	vrátiť sa do **čias** (niečoho / keď...);
			dožiť sa … **rokov**;
	DATIVE		
Attr + Sub1Dat		ATTRIBUTE	oproti minulým **rokom**;
Sub2 + Sub1Dat		NOUN	
Verb + Sub1Dat			

Number	Case	Collocator	Example
		VERB	
	ACCUSATIVE		
Attr + Sub1Aku		ATTRIBUTE	na tie **časy**;
Sub2 + Sub1Aku			toľké **časy**;
Verb + Sub1Aku			po dlhé **roky**;
		NOUN	**roky** a **roky**;
		VERB	blýska sa na **časy**;
			pamätať si **časy** (niečoho / keď...);
			prežiť **roky**;
			potrvá **roky**;
			cítiť **roky**;
	LOCAL		
Attr + Sub1Loc		ATTRIBUTE	v dávnych **časoch**;
Sub2 + Sub1Loc			v tých **časoch**;
Verb + Sub1Loc			v najlepších **rokoch**;
		NOUN	v **časoch** niečoho (existencie / krízy / …);
			(muž/žena/pán/pani / ...) niekto v **rokoch**;
		VERB	
	INSTRUMENTAL		
Attr + Sub1Ins		ATTRIBUTE	pred (…) **rokmi**
Sub2 + Sub1Ins		NOUN	
Verb + Sub1Ins		VERB	

Table 1.

For other templates of verbs, adjectives and adverbs with more examples see `http://www.vronk.net/wicol/index.php/Main_Page`.

References

1. Čermák, František: Kolokace v lingvistice. In: František Čermák – Michal Šulc (ed.): Kolokace. Studie z korpusové lingvistiky. Sv. 2. Praha, Nakladatelství Lidové noviny, Ústav Českého národního korpusu 2006a, 9-16.
2. Čermák, František: Statistické metody hledání frazémů a idiomů v korpusech. In: František Čermák – Michal Šulc (ed.): Kolokace. Studie z korpusové lingvistiky. Sv. 2. Praha, Nakladatelství Lidové noviny, Ústav Českého národního korpusu 2006b, 94-106.
3. Čermák, František: Statistical Methods for Searching Idioms in Text Corpora. In: Häcki-Buhofer, Annelies; Burger, Harald (Hrsg.): Phraseology in Motion 1. Methoden und Kritik. Baltmannsweiler (Schneider Verlag Hohengehren) 2006c (= Phraseologie und Parömiologie. 19), 33-42.
4. Hausmann, Franz J.: Kollokationen im Deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: Bergenholtz, Henning Mugdan, Joachim: Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch vom 28. bis 30.06.1984. Tübingen (Max Niemeyer) 1985 (= Lexicographica. Series Maior. 3), 118-129.
5. Hausmann, Franz J.: Was sind eigentlich Kollokationen?. In: Steyer, Kathrin: Wortverbindungen – mehr oder weniger fest. Jahrbuch 2003. Berlin, New York (de Gruyter) 2004, 309-334.
6. Heyer, Gerhard – Quasthoff, Uwe – Wittig, Thomas: Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse. W3L Verlag, Hardecke, Bochum 2006.
7. Jarošová, Alexandra: Spracovanie ustálených spojení vo výkladovom slovníku. Návrh. – In: Nová slovní zásoba ve výkladových slovnících. Sborník příspěvků z konference. Praha, 31. 10. – 1. 11. 2000. Red. O. Martincová – J. Světlá. Praha, Ústav pro jazyk český Akademie věd České republiky 2000a, s. 41 – 54.
8. Jarošová, Alexandra: Viacslovný termín a lexikalizované spojenie. – In: Človek a jeho jazyk. 1. Jazyk ako fenomén kultúry. Na počesť profesora Jána Horeckého. Red. K. Buzássyová. Bratislava, Veda 2000b, s. 481 – 493.
9. Jarošová, Alexandra: Lexikalizované spojenia v kontexte ustálených spojení. – In: Princípy jazyka a textu. Materiály z medzinárodnej vedeckej konferencie konanej 9. – 10. 3. 2000 na Katedre slovenského jazyka Filozofickej fakulty Univerzity Komenského. Zost. J. Dolník. Bratislava, Univerzita Komenského 2000c, s. 138 – 149, angl. res. s. 149, lit. s. 149 – 153
10. Mlacek, Jozef: Tvary a tváre frazém v slovenčine. Bratislava, Stimul 2001.

A Cross-linguistic Database of Children's Printed Words in Three Slavic Languages

Radovan Garabík[1], Markéta Caravolas[2], Brett Kessler[3], Eva Höflerová[4], Jackie Masterson[5], Marína Mikulajová[6], Marcin Szczerbiński[7], and Piotr Wierzchoń[8]

[1] L. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
[2] School of Psychology, University of Liverpool
[3] Psychology Department, Washington University in St. Louis
[4] Department of Czech Language and Literature with Didactics, University of Ostrava
[5] Institute of Education, University of London
[6] Faculty of Education, Comenius University, Bratislava
[7] Department of Human Communication Sciences, The University of Sheffield
[8] Institute of Linguistics, Adam Mickiewicz University, Poznań

Abstract. We describe a lexical database consisting of morphologically and phonetically tagged words that occur in the texts primarily used for language arts instruction in the Czech Republic, Poland and Slovakia in the initial period of primary education (up to grade 4 or 5). The database aims to parallel the contents and usage of the British English Children's Printed Word Database. It contains words from texts of the most widely used Czech, Polish and Slovak textbooks. The corpus is accessible via a simple WWW interface, allowing regular expression searches and boolean expression across word forms, lemmas, morphology tags and phonemic transcription, and providing useful statistics on the textwords included. We anticipate extensive usage of the database as a reference in the development of psychodiagnostic batteries for literacy impairments in the three languages, as well as for the creation of experimental materials in psycholinguistic research.

1 Motivations for the West Slavic database

Lexical databases that reflect language use across the developmental spectrum are critical tools for research on the development of spoken and written language skills because they allow researchers to select materials for their studies that are age- and grade-appropriate. A number of databases exist for adult language, but only a few have been developed based on child language. Available child-language corpora include the earlier American English sources *The American Heritage Word Frequency Book* (Carroll, 1971) and *The Educator's Word Frequency Guide* (Zeno, Ivenz, Millard & Duvvuri, 1995), and more recently, *Manulex*, a French database (Lété, Sprenger-Charolles & Colé, 2004) and the British English *Children's Printed Word Database* – CPWD (Masterson, Stuart, Dixon & Lovejoy, 2003). However, to our knowledge, no corpus of children's printed words has been published in any of the Slavic languages.

The data that can be generated from lexical databases have diverse applications in psycholinguistic research. For example they can produce statistics about lexical and sublexical variables such as frequency of specific units, word length in terms of letters or syllables, orthographic and phonological neighbourhoods, and grapheme–phoneme consistency, to name but a few. Accumulating evidence shows that text-based variables such as these affect learning to read and spell from an early age (e.g. Caravolas, Kessler, Hulme & Snowling, 2005; Treiman & Kessler, 2006; Pacton, Perruchet & Fayol, 2001). An emerging key issue in this research area concerns the relative influence of orthographic depth on the learning process: Does the predictability (transparency) of a specific writing system significantly influence the way children learn to read and write it? Direct cross-language comparisons based on corpus statistics will play a critical role in answering this question. However, a current limitation is that there are still few children's lexical databases in different languages, and those that do exist, rarely generate directly comparable statistics. This is because databases may be designed for different scientific purposes and thus do not always contain similar information from language to language. Moreover, linguistic features that are important in one language may be deemed to be of marginal importance and thus not warrant inclusion in another. Thus, a fundamental motivation for our project was to redress these shortcomings in the creation of a database that would allow direct cross-linguistic comparisons of a wide range of measures across Czech, Polish and Slovak. Cognizant of the prevalence of English-language research and of English-based models of language and literacy development, we based the West Slavic lexical database (Weslalex) on the existing English CPWD (Masterson et al., 2003). These design features will enable researchers of Slavic languages to investigate questions that could not be addressed without corpus data, and, they will facilitate meaningful comparisons to English measures, which so often provide the benchmark in developmental psycholinguistic research.

2 Types of corpus statistics provided in existing children's corpora

The existing American English corpora provide only word frequency information across (Carroll, 1971) and within (Zeno et al., 1995) primary school grades. The French Manulex (Lété et al., 2004) currently contains lemmatized and nonlemmatized grade-level word frequency lists, limited part of speech (POS) information, and letter frequencies. The more recent extension, Manulex-infra (Peereman, Lété & Sprenger-Charolles, in press), generates statistics at the sublexical level (syllable, grapheme-to-phoneme mappings, bigrams), and lexical level (lexical neighborhood, homophony and homography). The British English CPWD (Masterson et al., 2003) allows searches by grade and it offers a wide range of possible searches at the lexical and sublexical levels. These include searches of orthographic and phonological attributes such as neighbourhoods, component letters and phonemes, word length, and frequency. A feature that is currently missing from all of these corpora is a detailed morphosyntactic level of analysis.

Although impressive advances are now being made in several languages, no comprehensive children's database, that includes all of the above search possibilities, has yet been developed.

3 Features of the West Slavic database

The database that we are developing is modelled in part on the CPWD, and one of our key objectives is to make possible parallel cross-linguistic searches in any of our Slavic languages and this English language resource. In addition, however, a truly useful tool for psycholinguistic research in the inflected Slavic languages requires information not only at the lexical and sublexical (grapho-phonological) levels, but also at the morphophonological, grammatical and phrase levels.

Thus we include POS information derived from sentence-level analyses, and one of our search tools permits searching of multiword sequences. The integration of the lexical/sublexical database and of the sentence-level database is one of the critical challenges being addressed in our project.

4 A description of the pilot database materials

The database currently contains printed words in Czech (388 654 tokens, 64 411 distinct wordforms, 24 364 distinct lemmas), Polish (175 404 tokens, 34 067 distinct wordforms, 13 767 distinct lemmas), and Slovak (180 674 tokens, 30 060 distinct wordforms, 14 610 distinct lemmas)[1] from texts primarily used for language arts instruction in each country in grades up to 4 or 5. Based on surveys carried out in each of the three countries, we selected books and materials from those series that are currently the most widely used. Some intercultural differences necessarily emerged so that different numbers of books were sampled in each country. The simplest case proved to be Slovak where only one language arts series is approved by the Ministry of Education; thus we selected the designated readers and one Slovak language grammar book from each primary grade (1 to 4). The total number of Slovak books is therefore relatively small (9 books), but they represent an exhaustive sample of the materials children read as part of their language arts instruction. The Czech case was less straightforward because several Ministry-approved series exist; however, we chose the two series that predominate. Thus for each grade level (1 to 5) we chose one reader and one grammar workbook from each series for a total of 19 books (grade 1 did not have a grammar workbook). The Polish case was the most complex for two reasons. First, as in Czech, several series are Ministry-approved, thus necessitating the selection of a sub-sample. Second, the recently reformed primary education curriculum integrates teaching of different subject areas and thus no separate, dedicated language arts text books are currently in use. Instead, for

[1] The count contains the tokens without punctuation, digits and nonwords. Also the text annotated as instructions is excluded from the lowest grade – the reason being that these instructions are presumably not read directly by the children in this grade.

each grade level (0–3 equivalent[2] to 1–4 in the Czech and Slovak school system) children receive up to 20 booklets in which language as well as maths, science, etc. are covered in overlapping sequences. Consequently, we selected one widely used scheme and within this we selected 11 booklets (five from grade 0, two from grades 1–3 each), we prioritized those with a greater emphasis on language arts and reading where possible. We are confident that using these procedures we have sampled books in each language that are highly representative of the reading materials encountered by primary school children.

5 Text processing and annotation

All of the books were scanned and submitted to OCR. The OCR-ed text files were then proofread and annotated by proficient speakers of each language (typically students in language education or in psycholinguistics). The texts were then analysed for morphological categories and POS in each language, and were phonologically transcribed, obtaining texts with POS, morphological categories and phonemic transcription for each word.

The texts were manually annotated with XML to mark nonwords and metatext – instructions on using the text. Using XML tags provides the possibility of using already existing tools for XML validation, thus reducing the number of annotation mistakes.

6 Morphosyntactic annotation

Due to the highly inflected character and rich morphology of the languages in question, morphosyntactic analysis and lemmatization is not a trivial task. Given the number of texts present, it was impractical to annotate the words manually, and we had to use automatic tools. For Czech, the tagger described in (Hajič & Hladká, 1997) was used.

Since no sufficiently accurate Polish parser was freely available, we used the Waspell (Płotnicki, 2003) morphological analyzer program, which is based on the Ispell Polish dictionary.[3]

For Slovak, although another morphological tagset and tools for automatic text processing (Garabík, 2006) exists, we decided to use the same tagset as the Czech one, adapted for the Slovak language (Hajič, Hric & Kuboň, 2000). This will make eventual cross-linguistic comparison between Czech and Slovak texts easier.

6.1 Czech and Slovak tagsets

The Czech tagset describes 13 different morphological categories: part of speech, detailed part of speech,[4] gender, number, case, posessor's gender, posessor's number, person, tense, degree of comparison, negativeness, voice and register. There

[2] Children at the age of 6 enter a reception grade, which is referred to as grade 0.

[3] `http://ispell-pl.sourceforge.net/`

[4] As named by the authors.

are 12 different main part of speech categories, with particles forming their own category (as is the custom in Slavic linguistics). Punctuation also has its own tag (including sentence boundaries mark).

Special care has to be taken in interpreting the gender category – in addition to four common values (masculine animate, masculine inanimate, feminine and neuter), there are also additional possible values, corresponding to genders conflated due to inflectional syncretism. These are "feminine or neuter", "feminine singular only or neuter plural only", "masculine inanimate or feminine plural only", "masculine (either animate or inanimate)", and "not feminine".

The number category contains a special value for the old Slavic dual, present in the Czech language only in a few nouns (not present in the Slovak version of the tagset). The most notable deficiency of the tagset is the absence of a verb aspect category.

6.2 Polish tagset

The Waspell morphological analyzer analyses wordforms in isolation, without taking syntactic context into account. While it produces a detailed morphological description (similar to that obtained for the Czech and Slovak corpora) it results in a large proportion of alternative descriptions, due to inflectional syncretism. For example, each occurrence of the token *miał*, which can mean either 'coal dust' or 'he had', receives these three analyses:

- noun, masculine, singular, nominative
- noun, masculine, singular, accusative
- verb, perfective, indicative, past tense, 3^{rd} person masculine singular

Manual disambiguation of this output (by identifying ambiguous words in original texts) was not feasible at the level of full morphological description. Therefore, for the pilot version of the database we decided to specify only the part of speech. While part of speech could also be ambiguous (as in the example given above) this ambiguity affected only a small proportion (1–2%) of wordforms. Disambiguation was accomplished by listing all ambiguous wordforms in Waspell's output, and checking them against the original text.

7 Phonemic transcription

It was desirable to include a phonemic transcription of the written texts. This gives us access to various statistical analyses on the spoken, rather than the written, language level. There were, however, several open problems concerning the exact nature of the transcription. One possibility was to deploy the SAMPA transcription (Fourcin, Harland, Barry & Hazan, 1989), a substitute for the International Phonetic Alphabet which has the advantage of using only ASCII characters, which are easily entered and do not require any special software arrangements at the client's side, such as special fonts and keyboard layout. However, with modern computing systems, the technical advantages of SAMPA

diminish and some disadvantages come to the fore. The SAMPA transcription is usually specific for a given language, and transcriptions for different languages sometimes collide, complicating further comparisons (however, this is not the case with the Czech, Polish and Slovak transcriptions). Moreover, SAMPA for these languages uses a rich variety of non-alphanumeric symbols, interfering with regular expressions, thus complicating queries in the database. We therefore opted for an IPA transcription, using Unicode internally and presenting the output in UTF-8 encoding.

Pronunciation was rendered in terms of classical phonemic analysis, which means that sometimes salient phonetic and morphological information was disregarded. Diphthongs were treated as a sequence of two phonemes (vowel plus glide), affricates and long vowels were treated as single phonemes. When no confusion could result, typographically simple (preferably ASCII) IPA characters were chosen over more complex and precise ones: thus Czech *hezká* 'pretty' is transcribed as /heskaː/, not /ɦɛskaː/. Nonsyllabic components of diphthongs were given distinctive representations as glides (/j/ as in *kraj* /kraj/ 'region', or /u̯/ as in Czech and Slovak *auto* /au̯to/ 'car' – disyllabic sequences of vowel plus vowel are possible in the languages targeted.

Syllable boundaries were not marked, because they are usually inferrable from the phoneme sequence; at other times they can be controversial. Word stress was not transcribed because it is not phonemic.

There are different possibilities in representing sibilant affricates. The first and simplest possibility is to encode them as the sequence of constituent phonemes, /ts/, /dz/, /tʃ/, /dʒ/, /tɕ/, /dʑ/. This is clearly unacceptable for Slavic languages, because the affricates are phonologically different from diphonemic, non-affricate sequences of phonemes. Another possibility is the obsolete IPA ligature notation: /ʦ/, /ʣ/, /ʧ/, /ʤ/, /ʨ/, /ʥ/. This has the clear advantage of representing one (affricate) phoneme with one Unicode character, facilitating regular expression queries and statistical analysis. The last possibility to be considered is to use the official IPA transcription (with the tie bar above), /t͡s/, /d͡z/, /t͡ʃ/, /d͡ʒ/, /t͡ɕ/, /d͡ʑ/. This has the advantage of being official, but also two principal disadvantages. The first is that there is not yet quite uniform support of Unicode combining characters across various operating systems and font rendering engines commonly used. However, support is rapidly growing and is present in all the recent common computing environments, and, where lacking, an easy, free upgrade is almost always available. The second disadvantage is the fact that each affricate is now represented as a sequence of three Unicode characters. This makes the queries more difficult, e.g. to search for a word with a phoneme /t/, instead of writing a regular expression `.*t.*`, we have to make sure that the character following the character after /t/ is not a `U+0361 COMBINING DOUBLE INVERTED BREVE`. The above mentioned regular expression would look like `.*t(?!͡).*` – that is, /t/ not followed by two characters, the second of which is the tie bar, followed by any sequence of characters. This is complicated by difficulties of entering and editing text with a standalone combining character; however, these complications can be remedied by putting a special translation

level in the WWW interface, translating simpler (preferably noncombining), user entered characters into complicated IPA Unicode character sequences.

7.1 Czech phonemic transcription

A standard literary Czech pronunciation was chosen as a reference. In the absence of definitive, freely available pronunciators, Czech pronunciations were algorithmically inferred from their spelling. Proceeding from left to right through the word, each of approximately 94 context-sensitive rules were consulted, and the first rule that fit the context determined the pronunciation of the letter in question. For example, the rules for the letter *ě* decode it as follows:

1. /ɲe/ if immediately preceded by *m*
2. /e/ if immediately preceded by a dental stop letter, i.e., the regular expression `[dnt]`
3. /je/ otherwise

In turn, the rules for the dental stop letters each contain a clause mapping them to a palatal stop when followed by *ě*, among other letters (i.e., `[ěií]`). The contextual rules all reference the input orthography, not the output phonology, and there are no multi-step derivations for individual letters. A few rules make special provisions for consuming more than one letter at a time (*ch* mapping to /x/, *dz* mapping to /d͡z/, *dž* mapping to /d͡ʒ/, but such complications are rare in Czech, which has very few digraphs.

Most of the context-sensitive rules serve to handle voicing issues, such as when letters like *d*, normally voiced (/d/), are devoiced, as in *led* /let/ 'ice', or vice versa *četba* /t͡ʃedba/ 'reading'. Such mappings can easily be predicted from the immediate context. A slightly more complicated case is that devoicing occurs before final *me* but only in first-person plural imperative verb forms, as in *odpovězme* /otpovjesme/ 'let us answer'. Here the rule system can exploit the fact that the words entering the pronunciation module have already been tagged for morphosyntactics; analysis tags beginning `Vi-P---1` identify words as being first-person plural imperative verbs.

Unfortunately, however, the tagging system does not provide all the information needed to unambiguously apply all of the rules of Czech orthography. For example, in the word *odzátkovat* 'to uncork', the letters *dz* are to be interpreted as two individual phonemes, /dz/, not as the digraph pronounced /d͡z/. This could be inferred if the system knew the word has a transparent morpheme boundary between the *d* and the *z* (*od-* is a private prefix, *zátka* is 'cork'), but that information is not supplied. Furthermore, there are hundreds of words whose correct pronunciation could easily be inferred if the system only knew that they were of non-Slavic origin. For example, *n* is normally pronounced /ɲ/ before *i*, but not in Latinate words like *penicilin* /penit͡silin/ 'penicillin'. These and other issues are handled with an exception list. Virtually all of the exceptions are of the form `penicilin → p=p e=e n=n`, which means that if a word has been tagged as a form of the lexeme *penicilin*, then if the wordform in question begins *pen*, those

three letters should be assigned the phonemes /p/, /e/, and /n/, respectively, and the rest of the word should be decoded using the ordinary rules of Czech orthography.

Czech is sufficiently regular that this simple system is basically adequate, although it would be more satisfying if such "exceptional" pronunciations could be worked out from first principles (e.g., etymological information) instead of by stipulation. Catching exceptional pronunciations is fairly labour intensive, although the work was speeded significantly by checking a sorted list of the word types in the corpus against two Czech language works that include notes on some exceptional pronunciations: *Pravidla českého pravopisu* (2005) and *ABZ slovník cizích slov* (2006).

Further progress in correcting exceptional spellings will require careful attention from linguistically trained native speakers. Especially with lesser-known or more recently introduced foreign words, it is not always easy to guess how their pronunciation will be adapted to Czech phonology. Another continuing problem involves the pronunciation of unusual clusters such as double letters. Whether these are reduced to simpler pronunciations, such as single phonemes in the case of double letters, depends on a combination of factors such as the salience of any intervening morpheme boundary and a rather nebulous perception of which of multiple variant pronunciations is currently considered too colloquial or too stilted to be considered standard. In the absence of authoritative reference books, the active assistance of a skilled orthoepist is required.

Pronunciations are internally stored in a way that explicitly aligns them to their spelling. For example, the pronunciation of *hezká* 'pretty' is stored in the form `h=h e=e z=s k=k á=aː`, with letters to the left of each equals sign and phoneme representations to the right. This representation facilitates the compilation of statistics on spelling consistency and enables the searcher to easily request, for example, all words where /z/ is spelt *s*. This notation accommodates the relatively rare instances where a sound is spelt with two letters, as in *Čech* 'Czech man' `č=t͡ʃ e=e ch=x` (technically *ch* is considered a single letter in Czech) and *Anglie* 'England' `a=a n=n g=g l=l i=ij e=e`.

7.2 Polish phonemic transcription

The Polish pronunciator (which is still under development) is modeled very closely on the Czech pronunciator described above: it algorithmically derives pronunciation of single words from their spelling.

Building the pronunciator required choosing one model of "standard Polish pronunciation" among several available alternatives. Some particularly important phenomena include the varying pronunciation of word-final *ę* (pronounced as nasal vowel /ẽ/ in careful, conservative speech, but typically is oral /e/); the distinction between dental nasal [n] and velar nasal [ŋ], which is phonemic in the Warsaw dialect of Polish, but phonetic in the Poznań-Kraków dialect, where [ŋ] is only a positional variant of /n/ (Strutyński, 1997); and the distinction between some palatalized consonants (e.g. [pʲ], [kʲ]), and their non-palatalized

equivalents ([p], [k], etc.) which is again either phonemic or merely phonetic, depending on the region.

Since Polish has several digraphs, (e.g. *rz*) where Czech and Slovak have accented letters (e.g. *ř*), this opens up more possibilities for parsing ambiguities; for example, *rz* represents two separate sounds in *marznąć* /marznoɲt͡ɕ/ 'to freeze', rather than its usual /ʒ/. In practice, however, we have found that very few exceptions arise if the program always treats digraphs as such; such exceptions can be handled by explicitly listing exceptional lemmas. As was the case in Czech and Slovak, the identification of such exceptional pronunciations requires checking the output of the analyzer by linguistically trained native speakers.

7.3 Slovak phonemic transcription

For the automatic Slovak transcription, we used the system described by (Ivanecký, 2003). The software transcribes the text into SAMPA (Ivanecký & Nábělková, 2002); we created translation tables into IPA as used by our database. The software can take into account consonant assimilation across word boundaries; however, we decided not to use this possibility for the pilot version, since at least initially, most queries will concern only isolated words, and because we will thus maintain compatibility with Czech and Polish transcriptions. However, devoicing at the ends of words was still applied. One problem concerns the transcription of the vowel *ä* and the syllables *le*, *li* and *lí* – all of which have two possible pronunciations, /æ/ vs. /e/; /ʎe/, /ʎi/ and /ʎiː/ vs. /le/, /li/ and /liː/. While the former pronunciations have for all practical purposes disappeared from standard Slovak, they are still considered formally correct, and as such are taught in elementary schools. Thus, using /e/ and /l/ for the transcription would accurately reflect standard spoken language, while using /æ/ and /ʎ/ would reflect the prescribed school usage. We have chosen the official prescribed variant, since teachers are more likely to pronounce the phonemes in the "correct" way, especially when teaching the letter–sound correspondences.

The accuracy of the transcribed phonetic assimilation heavily depends on the detection of intra-word morpheme boundaries, which in turn requires a good morphological dictionary, which is not currently available. Another problem is a (relatively) huge number of orthographic exceptions in Slovak, especially in loanwords. These mostly involve *e*, *i* and *í*, which routinely mark preceding *t*, *d* or *n* as palatal in native words but not in non-Slavic loans. Currently, the software contains just a small dictionary of orthographic exceptions, and therefore transcription of most of the loanwords marks palatal consonants incorrectly.

8 Query interface

We made available two independent query mechanisms, accessible from the project's page at `http://spell.psychology.wustl.edu/weslalex`. One indexes the data and presents the query results in a keyword-in-context (KWIC) interface; the other is similar to the CPWD functionality.

For the first tool, the files were converted into so-called vertical files, where each word in the text is represented by a separate line. Each line contains several tab-separated fields: the word's spelling, lemma, morphological tag, and pronunciation in IPA. This format is easily parseable by simple computer programs and therefore is well suited for custom statistical analyses beyond the capabilities of the systems described here.

The vertical files were indexed using the Manatee corpus manager system (Rychlý, 2000). It is possible to query the corpus using specialized multiplatform client software *Bonito*, offering a rich set of query possibilities and statistical analysis. It is possible to query individual token attributes matching given regular expressions, to search for sequences of arbitrary tokens, to apply negative and positive filters depending on context to search results, to sort the concordances by different criteria, and to obtain a frequency analysis of any token: raw frequency, collocates, and measures of mutual independence (raw MI scores and frequency-adjusted /t/ scores). A simplified user interface has been built to accommodate the need for quick access to queries, without the need for complex statistical analysis. This interface is accessible through a simple WWW interface, and affords the possibility of the same type of queries as Bonito. This interface contains a simple on-screen virtual IPA keyboard, to facilitate queries in the phonemic transcription.

The other query tool is intended to provide for the Czech corpus a functionality similar to that of the CPWD. It is an HTTP-based service that lets users query the corpus by desired lexical characteristics. Users can specify which specific texts they wish to search, whether they wish to include metatextual instructions, what lexical properties they wish to search by, and what properties they wish to see displayed for each retrieved word type. In addition, token counts are always provided for each word type, as well as certain types of statistics for each displayed property.

The basic properties defined for each word are its case-sensitive spelling; its lemma, lemma properties, and morphosyntactic analysis, as supplied by the (Hajič & Hladká, 1997) program; and its letter–phoneme alignment. For convenience, several secondary properties are derived when users refer to them. These include an uppercase version of the spelling, which is useful when users wish to ignore case distinctions; the pronunciation; various fields for different parts of the morphosyntactic analysis, such as part of speech, gender, and so forth; and spelling and pronunciation lengths.

An unusual characteristic of the retrieval tool is that the definition of a word type is not built in, but is defined by the end users based on what properties they ask to have displayed. For example, if users request case-sensitive spelling, lemma information, and full morphosyntactic analysis, the definition of word type will be very specific. A word like *růže* 'rose(s)' could appear as several different word types depending on the specific case and number it represents (nominative singular, genitive singular, nominative plural, etc.), and a capitalized version would represent many more types. On the other hand, if users request only the

uppercased version and do not ask for full morphosyntactic analysis, all uses of *růže* or *Růže* will be subsumed under only one word type.

The system also displays aggregated statistical information depending on what fields the users ask for. If users ask to see a spelling field, the tool will give the range and mean of the word lengths and will also tell how often each letter appeared. If the letter-phoneme alignment field is requested, spelling correspondence statistics are provided in each direction: e.g., how many times the letter *z* represented each of its possible pronunciations (/z/ and /s/), and how many times a phoneme, such as /z/, is spelled with each of its possible spellings (*z* and *s*). For the categorical fields, such as gender, a tally is provided telling how many times each individual level, such as feminine or masculine, appeared. All counts are given both by word types and by word tokens.

By default, information is presented about all the words in the corpus. It is also possible to limit the search to words that have particular properties. Typical boolean search queries are accepted. Most fields contain character strings and can be searched via arbitrary regular expressions. For example, the following query looks for nouns that end in *-o* in the nominative or accusative case but which are (contrary to the usual pattern) not neuter: `spell_uc = ".+O" and pos = "N" and (case = "1" or case = "4") and not gender = "N"`

Output is sorted in Czech lexicographic order and is presented as a list of tab-separated values, to facilitate importation into statistical programs or spreadsheets. The default character encoding is Unicode UTF-8, with spellings presented in normal Czech orthography. However, because it is not always convenient for users to input such characters, ASCII sequences are also understood, such as *c<* for *č* and *u0* for *ů*. For IPA characters, several different synonyms are often accepted: /t͡ʃ/ can be input as /t_s</ or /c</, or, for those with Czech keyboards, /t_š/ or /č/. Display uses the full Unicode character set unless the user specifically requests a more limited set; choices are ASCII and two character sets often used for Central European computing: Windows-1250 and ISO/IEC 8859-2. These are provided primarily to accommodate legacy software.

Although the tool is currently very usable, several enhancements are being planned, foremost among these being the inclusion of material from the Polish and Slovak corpora. We also plan several additions to bring the interface and capabilities more in line with those of the CPWD. This involves adding new derived properties, such as counts of orthographic neighbours (words that differ by only one letter). Perhaps even more important will be the addition of simplified search forms to make the corpus easy to search by people who have no prior experience with boolean search techniques and who have not read the help files.

9 Applications

The database was conceived by developmental psycholinguists and linguists for primary use in various types of developmental research. One important application of children's printed word statistics is in the study of the impact of 'exposure to print' and of implicit learning mechanisms on the development of reading and

spelling skills (e.g., Cassar & Treiman, 1997; Steffler, 2004). For example, in research on spelling development in a language like Slovak, it is important to understand the effectiveness of explicit instruction on children's learning of rules such as that governing the spelling of palatalized consonants (e.g., /c/ is spelt *t* before *i*, *í*, *e* but as *ť* in other environments), and to what extent this learning is influenced simply by the frequency of exposure to constructions that reflect this rule; there are many more words with *t*, such as *teta* 'aunt' than with *ť*, such as *ťava* 'camel'. Until now, only adult-language corpora were available to address this question. However, it is not always appropriate to rely on adult-language statistics. To illustrate, for a high frequency word in child language such as *teta* (in Slovak /ceta/), which is a model word used to teach the spelling rule for the palatalized consonant /c/ in second grade, the adult frequency according to the Slovak National Corpus is 23 occurrences per million (according to Mistrík, 1969 it is 3 occurrences per million). In contrast, the real frequency of *teta* in children's printed materials is much higher — in our Slavic database, it is 175 occurrences per million. Moreover, this word is acquired early, around 2.6 years of age (based on age of acquisition norms derived from adult Slovak speakers' ratings (Mikulajová, in preparation)), and thus is highly familiar to children by the time they are in second grade. Consequently, children might learn to spell /ceta/ correctly as *teta* and not *ťeta* through exposure to print during activities like shared reading of fairy tales and bedtime stories, well before the rule is explicitly taught.

The database is already being put to use in a broader project aimed at the creation of the first diagnostic spelling test in the Slovak language (Caravolas, Mikulajová & Vencelová, in preparation). Word lists for this standardized test battery were selected according to several criteria, one of which was word frequency. The database allowed us to check the frequency of words and sublexical units that appear in the reading materials that the majority of Slovak children are exposed to in grades 1–4. Thus, the role of explicit (via classroom instruction of rules and patterns) and implicit learning of spelling rules (from exposure to print) could be compared.

In another study (Caravolas, Mikulajová & Vencelová, 2007) the database was used to estimate frequencies of different types of Slovak graphemes (with and without diacritics) and syllables to investigate the role of sound–letter consistency and letter–sound complexity in learning canonical and contextually conditioned letter spellings in Slovak.

We envisage many further uses of our database, not only for a wide variety of single-language and cross-linguistic investigations of literacy development, but also in studies of child language development, bilingualism and biliteracy (at least in the languages included in the West Slavic database and English). Importantly, we expect the database to have applications beyond the field of psycholinguistic research. For example, teachers will be able to search for appropriate word lists for teaching and remediation, and writers of children's materials will have easy access to the vocabulary that children encounter in schoolbook reading at different ages. Not least, the database can be used for advancing the work of

computational linguists currently working on Czech, Polish and Slovak adult corpora, none of which currently contain sublexical or phonological information. Thus, we see our contribution of a child-language corpus as being of potential general interest to educators, linguists and psycholinguists alike.

Acknowledgement

The present project was supported by grant SG-40461 from the British Academy awarded to Markéta Caravolas, Jackie Masterson and Marcin Szczerbiński. The authors gratefully acknowledge the assistance of Vladimír Petkevič and Pavel Machač in POS tagging and phonological analysis of the Czech corpus, Małgorzata Kurek and Mária Podhradská for proofreading and mark-up of the Polish and Slovak texts.

References

ABZ slovník cizích slov (2006). Retrieved from `http://slovnik-cizich-slov.abz.cz/web.php/o-slovniku`.

Caravolas, M., Kessler, B., Hulme, C., & Snowling, M. (2005). Effects of orthographic consistency, word frequency, and letter knowledge on children's vowel spelling development. *Journal of Experimental Child Psychology*, *92*, 307–321.

Caravolas, M., Mikulajová, M., & Vencelová, L. (2007). *Effects of sound-letter consistency and letter-spelling complexity in learning canonical and contextually conditioned letter spellings in Slovak*. Poster presented at Society for the Scientific Study of Reading. Prague, Czech Republic.

Caravolas, M., Mikulajová, M., & Vencelová, L. (In preparation). Súbor testov na hodnotenie pravopisných schopností.

Carroll, J. B. (1971). *The American Heritage Word Frequency Book*. Houghton Mifflin.

Cassar, M. & Treiman, R. (1997). The beginnings of orthographic knowledge: Children's knowledge of double letter in words. *Journal of Educational Psychology*, *89*(4), 631–644.

Fourcin, A. J., Harland, G., Barry, W., & Hazan, V. (1989). *Speech input and output assessment: multilingual methods and standards*. New York, NY, USA: Halsted Press.

Garabík, R. (2006). Slovak morphology analyzer based on Levenshtein edit operations. In *Proceedings of the WIKT'06 conference*, (pp. 2–5).

Hajič, J., Hric, J., & Kuboň, V. (2000). Machine translation of very close languages. In *Proceedings of the sixth conference on Applied natural language processing*, (pp. 7–12)., San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hajič, J. & Hladká, B. (1997). Probabilistic and rule-based tagger of an inflective language: a comparison. In *Proceedings of the fifth conference on Applied natural language processing*, (pp. 111–118)., San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ivanecký, J. (2003). *Automatická transkripcia a segmentácia reči*. PhD thesis, Faculty of Electrical Engineering and Informatics, Technical University, Košice.

Ivanecký, J. & Nábělková, M. (2002). Fonetická transkripcia SAMPA a slovenčina. *Jazykovedný časopis*, *53*(2), 81–95.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 156–166.

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2003). Children's Printed Word Database. ESRC research project.

Mikulajová, M. (In preparation). Age of Acquisition norms derived from adult Slovak speakers' ratings.

Mistrík, J. (1969). *Frekvencia slov v slovenčine*. Vydavateľstvo slovenskej akadémie vied.

Pacton, S., Perruchet, P., & Fayol, M. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*(3), 401–426.

Peereman, R., Lété, B., & Sprenger-Charolles, L. (in press). Manulex-Infra: Distributional characteristics of grapheme-phoneme mappings, infra-lexical and lexical units in child-directed written material. Behavior Research Methods.

Pravidla českého pravopisu (2005). Academia. Kolektiv Ústavu pro jazyk český.

Płotnicki, Z. (2003). Słownik morfologiczny języka polskiego na licencji LGPL. Master's thesis, Poznań University of Technology.

Rychlý, P. (2000). *Korpusové manažery a jejich efektivní implementace*. PhD thesis, Faculty of Informatics, Masaryk University, Brno.

Steffler, D. (2004). An investigation of grade 5 children's knowledge of the doubling rule in spelling. *Journal of Research in Reading*, *27*, 248–264.

Strutyński, J. (1997). *Gramatyka polska*. Kraków: Wydawnictwo Tomasz Strutyński.

Treiman, R. & Kessler, B. (2006). Spelling as statistical learning: Using consonantal context to spell vowels. *Journal of Educational Psychology*, *98*, 642–652.

Zeno, S. M., Ivenz, S. H., Millard, R. T., & Duvvuri, R. (1995). *The Educator's Word Frequency Guide*. Brewster, NY, USA: Touchstone Applied Science Associates.

Effective Methods of Building Slovak-Czech Dictionary

Marek Grác

Fakulta informatiky
Masarykova Univerzita
Botanická 68a, 60200 Brno, Česká republika

Abstract. Machine translation from Czech to Slovak is still in its early stage. Bilingual dictionaries have big impact on quality of translation. As Czech and Slovak are very close languages, existing dictionaries cover only translation pairs for words which are not easy to infer. Proposed method described in this paper attempts to extend existing dictionaries by those easily inferable translation pairs. Our semi-automatic approach requires mostly 'cheap' resources: linguistic rules based on differences between words in Czech and Slovak, list of lemmata for each of the language and finally a person (non-expert) skilled in both languages to verify translation pairs. Proposed method tries to find candidate translations for given word and select the most similar lemma from other language as translation. Preliminary results show that this approach greatly improves effectivity of building Czech-Slovak dictionary.

1 Introduction and Motivation

Machine translation counts among those problems expected to have been already solved using computers. Development in this area passes through periods of both prosperity and diminished research investments. In recent years, unrealistic expectations seems to have faded. Investments are logically focused on translations among languages with commercial potential: English–Japanese, Arabic or Spanish. Minor languages including Czech and Slovak, lack both resources and commercial potential. Therefore, translation among those is generally neglected.

One of the basic pieces needed for creation of machine translation tool is a translation dictionary. In case of close languages, we can usually do with a simple dictionary, that only contains translation pairs and smaller, more robust dictionary, for more complex cases (e.g. context-dependent word translation). In this paper, we will focus on a method for efficient creating a dictionary with translation pairs only.

2 Slovak and Czech Languages

Czech and Slovak are two closely related languages which together form the Czecho-Slovak sub-group of West Slavonic languages [1]. Despite their similarity and nearly universal mutual intelligibility, the literary languages are clearly

differentiated. The effect is that among other things they are each based on a different dialect due to their separate standardisation. Although Czechs and Slovak lived together in one state for relatively short period of 68 years (1918–1938 and 1945–1992) compared to their literary tradition their linguistic relations have been very close since the Middle Age. Most of this time, however, the relationship was somewhat asymmetrical.

Similarities among these languages allow us to simplify the groundwork required for a translation dictionary of words and phrases. In most cases, it is sufficient to add the important suffixes for creation of inflected word forms. These are extractible from a morphological database. Word, that overlap in only a subset of meanings need to be found and processed manually. These words form the basis for differential dictionaries, which have been – and still are – published as books, and it is therefore possible to rely on them. An other language resources for Czech and Slovak are single-language corpuses (e.g. ČNK [2] and SNK [3]), morphological databases and lists of lemmata from other sources (e.g. spell-checkers).

3 Rule-based method

In the creation of the dictionary, we chose to use lists of words from freely accessible sources used for spell-checking ([4] and [5]). Since these linguistic resources don't even have lexical categories assigned they can be easily extended. Same is true for the dictionary. For reference, we use PC Translator[1] dictionary, which we have available in digital form.

Due to similarities among the two languages, we have decided to devise rules to rewrite letters in a word (in the direction Slovak–Czech). In the process, two sets of rules have been created. First set K, by a native Czech speaker, contains 50 simple rules, the second set G, written by a native Slovak speaker, contains 15 simple rules. The rules in both sets are written using the same formal language regular expressions. The differences among the sets K and G are not only in the number of rules but also in approach. On average, set K generates four times as many candidates for Czech words as the set G. Translation candidates created this way are looked up in Czech list of words, and if found, we assume the translation pair valid. This way errors are introduced into the system, specifically with words that have different meaning, but after application of the rules are spelt the same (e.g. *(sk) kel → (cz) kapusta, (sk) kapusta → (cz) zelí*). In addition, we have included another set, B, which does no transformations at all (the words have to be spelt the same way in both languages). The sets have been tested over words present in the reference dictionary. Results are presented in table 1.

[1] The dictionary has been modified to only contain lemmata, which reduced the number of translation pairs to roughly 80 thousand pairs.

Set of Rules	Recall	Precision
B	18.17%	99.36%
G	37.91%	98.98%
K	52.65%	97.07%

Table 1. Results of method based on set of rules

4 Method based on edit distance

For creation of candidate translation pairs, it is possible to also use edit distance of two words. There are several metrics that could be used. We have chosen Levenshtein distance [6], which computes the number of changes (addition, deletion and replacement of a single character) that are necessary to get the second word (e.g. *l(kitten, sitting) = 3, since kitten → sitten → sittin → sitting*). Another important property of this metric is that can be used with words of different length. After computation on the reference dictionary, we have found that 75% of the translation pairs have Levenshtein distance in the range from zero to three. Edit distance between the words in the translation pair is generally not related to frequency of the word, and differences among distribution of words with different distance among 1000, 5000, 10000 and 20000 most frequent words is according to our research just above statistical error.

Apart from Levenshtein distance, we have tried also a modified version of it [7], that allows us to set prices for different changes (i.e. change of 'á' to 'a' is cheaper than a change from 'g' to 'a'). Due to large number of calls to the distance function, this more expensive function would have caused the whole computation to run for several weeks. Due to this, together with relatively low gain we have decided not to use this method.

Finding translation pairs is done in two independent steps. First, finding candidates, since in a given distance, there may be more than one single word. Then, selecting the right candidate pair, out of the words, that are in the same distance.

We have implemented the search itself by picking the words with minimal distance to each Slovak word. This operation is fairly computationally intensive, since for each pair the computation has to start over.

Selection of translation pair was facilitated using three basic methods: FirstMatch, AnyMatch, JustOne. First of the them, FirstMatch takes for the valid pair the one that is first in alphabetic order, and it has been used as a baseline for comparison. The AnyMatch method accepted the translation pair to be valid if at least one of the pairs were present in the dictionary. Since this method never removes valid results, we have used it as an upper bound. The JustOne method lies in picking the translation pair only if there is exactly one candidate pair. The results for different distance thresholds are presented in table 2.

5 Combined methods

The described approaches have different combinations of coverage and precision. We wanted to find a method with high precision ($\geq 90\%$) and suitable recall.

We started to use method with best precision and use others approaches afterwards. First, we produced a translation candidates using a set of rules. If one of the translation candidate existed in target language we used it to create a translation pair. Otherwise we had count the edit distance between each translation candidate and each word in target language. Words that have edit distance below threshold (maximal acceptable edit distance) had been used as our new translation candidates and we have used elimination methods on them, e.g. JustOne to produce a translation pair.

It is also possible to use method based on edit distance more than once but we have to change metrics for edit distance. We have decided to use a modified Levenshtein method that is computationally intensive but we had use it only to count edit distance between source word and its translation candidates.

Selected results are presented in table 3. It is interesting that the difference in recall between rule sets G and K rapidly decrease when they are used in combined methods.

Differences in coverage between rule sets G and K have diminished from 18 to 2 percentage point.

6 Conclusions

This project has shown, that it is possible to create a Czech-Slovak dictionary even without extensive resources. Since in dictionaries, high precision is expected, the presented method is not suitable for fully automatic dictionary creation. The precision achieved, however, gives hope that the method will be useful, accompanied with existing differential dictionary, which should contain all kinds of problematic words. This approach is unlikely to be useful for languages with deeper differences, but there is possibility to apply it for e.g. creation of dialectologic dictionaries.

Acknowledgments

This work has been partly supported by the Academy of Sciences of Czech Republic under the project T100300419, by the Ministry of Education of CR within

	Edit Distance 1		Edit Distance 2		Edit Distance 3	
Method	Recall	Precision	Recall	Precision	Recall	Precision
AnyMatch	40.35%	93.04%	68.64%	83.75%	90.35%	79.34%
FirstMatch	40.35%	84.27%	68.64%	71.63%	90.35%	62.23%
JustOne	33.62%	94.46%	50.04%	88.89%	57.63%	86.47%

Table 2. Results of the method based on the edit distance

Sequence of Actions	Recall	Precision
rules G, distance ≤ 3 – AnyMatch	97.89%	85.28%
rules K, distance ≤ 3 – AnyMatch	98.83%	85.22%
rules K, distance ≤ 1 – JustOne, AnyMatch	99.12%	85.12%
rules K, distance ≤ 3 – JustOne, First	99.12%	77.36%
rules K, distance ≤ 1 – JustOne, ML ≤ 1 – JustOne	74.00%	93.68%

Table 3. Results of combined methods

the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project 201/05/2781.

References

1. Nábělková, M.: Closely-related languages in contact: Czech, slovak, czechoslovak. International Journal of the Sociology of Language (2007) 53–73
2. Kocek, J., Kopřivová, M., Kučera, K., eds.: Český národní korpus – úvod a příručka uživatele (Czech National Corpus – Introduction and Users Guide). FF UK – ÚČNK (2000)
3. ústav Ľ. Štúra SAV, J., ed.: Slovenský národný korpus, Bratislava (2007)
4. Kolář, P.: Czech dictionary for ispell. http://www.kai.vslib.cz/∼kolar/rpms.html (2006)
5. Podobný, Z.: Slovak dictionary for ispell. http://sk-spell.sk.cx (2006)
6. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10** (1966) 707–710
7. Štancel, R.: Methods for evaluation of machine translation (in slovak). Master's thesis, FI MU, Brno (2007)

Administration Framework for the DEB Dictionary Server

Aleš Horák and Adam Rambousek

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
http://nlp.fi.muni.cz/projects/deb2

Abstract. This paper presents a new implementation of administration framework for the DEB II dictionary writing system. We present the details and examples of the user management part as well as graphical scenarios for dictionary service setup, adaptation and automatic generation of user application based on the dictionary XML schema.
Keywords: *dictionary server; dictionaries administration framework; DEB*

1 Introduction

The DEB development platform is being developed at the NLP Centre at FI MU Brno for more than two years. The primary impulse for such system came from the need of fast implementation of new dictionary applications for various kinds of dictionaries. In order to cope with the required versatility, the XML format of the dictionary entry was a natural choice for the new system. Other useful features of the platform, that resulted from a complex analysis, are further described in the Section 2.

Of course, there were other systems designed on similar criteria as the DEB platform, such as the Papillon project [1] or the TshwaneLex [2] system, but they did not offer all the required qualities. The systems that are up-to-date and actively developed are usually distributed on a commercial basis, and the freely available ones come from older and terminated projects.

The DEB system in its current version is being actively used by several hundreds of users world-wide. In the following text, we will concentrate on the description of the new administration framework of the platform, which allows to handle the most frequent management tasks.

2 Overall description of the DEB dictionary writing system

The most important property of the system is the *client-server* nature of all DEB applications. This provides the ability of distributed authoring teams to work fluently on one common data source. The actual development of applications within the DEB II platform can be divided into the server part (the server

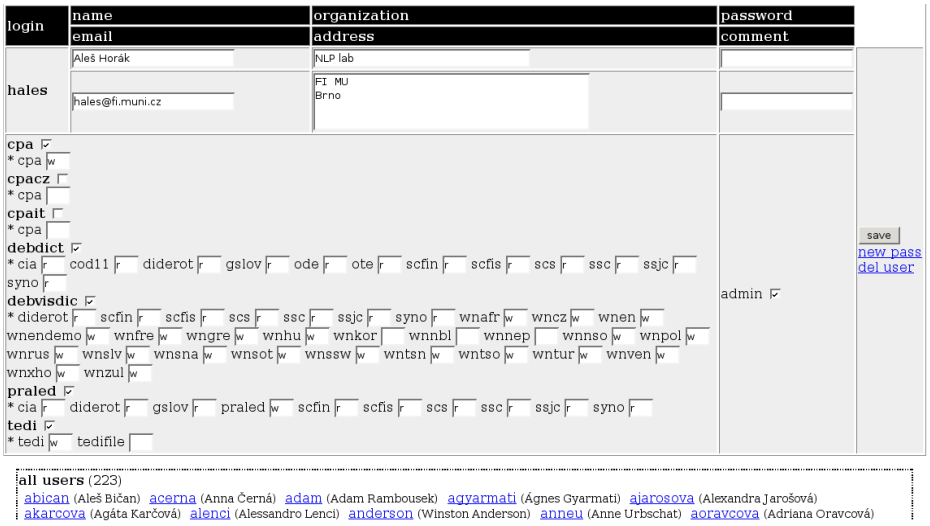


Fig. 1. User access rights settings.

side functionality) and the client part (graphical interfaces with only basic functionality). The server part is built from small parts, called *servlets*, which allow a modular composition of all services. The client applications communicate with servlets using the standard web protocol HTTP.

Since the data on the server is stored in XML, the actual data storage backend is provided by Berkeley DB XML [3], which is an open source native XML database providing XPath and XQuery access into a set of document containers.

The user interface, that forms the most important part of a client application, usually consists of a set of flexible forms that dynamically cooperate with the server parts. According to this requirement, DEB II has adopted the concepts of the Mozilla Development Platform [4]. Firefox Web browser is one of the many applications created using this platform. The Mozilla Cross Platform Engine provides a clear separation between application logic and definition, presentation and language-specific texts.

2.1 Current client applications

Current development of the DEB II platform includes implementation of several real-life dictionary applications such as DEBVisDic [5], PRALED [6], DEB CPA [7], Cornetto and others.

3 The DEB administration framework

Initially, DEB server was developed with just command-line management of dictionaries and administration of user passwords for authentication. The configuration was realized by structured text files and data processing scripts.

```
<user>
 <login>adam</login>
 <name>Adam Rambousek</name>
 <email>xrambous@fi.muni.cz</email>
 <org>Faculty of Informatics</org>
 <addr>Botanicka 68a, Brno</addr>
 <pass>3Ja8ivX12OB0U</pass>
 <services><service code="debdict">
  <dict code="scs" perm="r"/>
  <dict code="scfis" perm="r"/>
  <dict code="cia" perm="r"/>
  <dict code="scfin" perm="r"/>
  <dict code="diderot" perm="r"/>
 </service></services>
</user>
```

Fig. 2. XML entry for the user from the Figure 1

After DEBVisDic has spread to more users world-wide and has been used for building several national Wordnets (Polish, Hungarian, Slovenian or Afrikaans), a more sophisticated administration interface for DEBVisDic users and dictionaries was created. Later on, this interface was transformed to more general and complex dictionary management application for the whole DEB II server.

The DEB II server packages are currently being deployed on several servers in different organizations and often more than one user need to administer a single DEB server without having a direct server access. Thus, the administration interface must be accessible remotely and without any special tools. The best choice for this task is a web-based interface, where the user needs just a web browser.

The interface should support easy administration of all the server areas. Of course, the main area of a dictionary management server is the dictionary management. Each dictionary is described with several basic attributes, like its name and code, the filename of its storage in the DB XML database, its dictionary type, the XML schema or indexed elements or XSLT templates for output displaying. Also, some projects may need extra specific settings – e.g. the DEBVisDic clients need to store information about the inter-dictionary links. After the dictionary is set up, the interface has to support import and export of XML data into and from the DB XML format.

3.1 The implementation

The server administration interface is based on the same postulates as the other DEB II server dictionaries and modules. The Berkeley DB XML database provides a storage backend for the administration meta-data. The server-side scripts are developed in Ruby programming language.

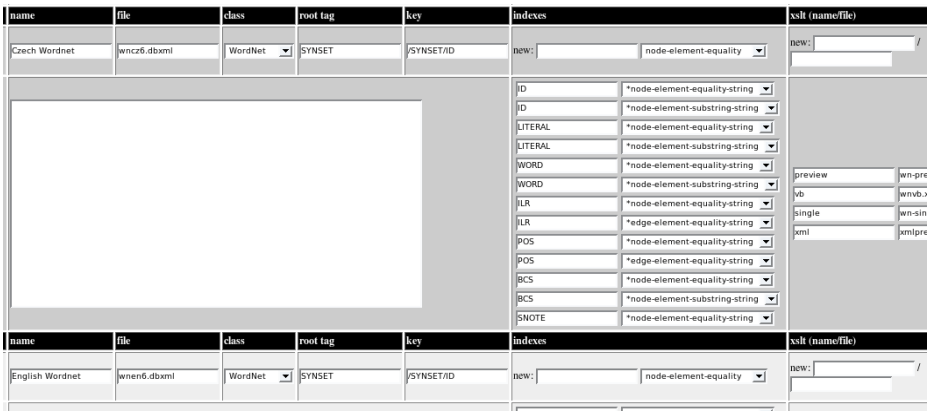


Fig. 3. Dictionary management showing basic information and indexed elements for the Czech Wordnet dictionary.

All the data about users, dictionaries, permissions and other control data are stored in the DB XML database in the XML format. Each dictionary module of the DEB II server uses a common interface to access data from this administration database.

The administration module provides several services – user authentication, access rights control, entry locking and journaling of dictionary changes.

The administration interface is a web-based application where the web pages are generated using an HTTP template which allows easy design and content modification and then served to the users by a light-weight web server – WEBrick [8]. The users are authenticated using standard HTTP authentication mechanism. The administration module extends the standard interface for passwords stored in a file and loads user's login and password from the XML database. Each change in user accounts or access rights is propagated to all DEB services in the real-time.

User Access Rights When the administrator sets up the server dictionaries, these can be grouped to "services." A service is one individual part of the DEB server, usually used for one particular project. For example, DEBVisDic or DEBDict are separate services, but they share the same base libraries and management database. Several services can access the same dictionaries, each providing different view on the data.

The user accounts are shared between all the services. Thanks to the database sharing between services, each user needs just one account for all the services he or she may use. The administrator can restrict access to selected services and for each service, more detailed access permissions can be set for each dictionary (read-only, read-write, update, see the Figures 1 and 2). The actual usage of the dictionary access permissions depends solely on the service. This means, one

service can ignore permissions at all and another service can use complex access rights.

Apart from access rights, the user account management provides all the needed functions – it allows to create, modify and delete user accounts. Each user can log-in to the administration interface and change his or her password. In case the user forgets a password, he or sher can ask for a new random password.

The Dictionary Management For each dictionary, the administrator has to define several attributes (see the Figure 3). The minimal set of attributes contains a unique dictionary code, a database filename and a dictionary class (the implementation class), the other attributes are more or less optional. The meaning of the dictionary attributes is:

- The dictionary name is displayed to users by the client application.
- The definition of the XML entry root tag and its key element are needed for XML import and for searching (in case, the application does not have its own, more complex search method).
- Indexes speed up search operations, so each element or attribute that is used in user queries should be indexed.
- The XSLT templates transform XML data to another form suitable for presentation or machine processing.

Extra dictionary attributes are required for the DEBVisDic dictionaries:

- Each DEBVisDic dictionary is linked to the client software by the client package code.
- The DEBVisDic Dictionaries can reference to each other using "equivalence tags."
- In the next field, the administrator can enter dictionaries that should be reloaded after an edit action in the client (usually in another dictionary).
- And the last option specifies related dictionaries – for example, several national Wordnets linked with ILI (Inter-Lingual Index). It is possible to display the same entry in different languages or to copy entries between languages.

Import and Export The import function takes an XML file and stores the data into the DB XML database. The XML file has to be uploaded to the server (it is possible to upload it through web interface). All entries must share the same root tag (specified in the dictionary management), entries with different root tags are ignored. The administrator can choose if he or she wants to delete all the entries from database before the import or just add the new entries. The import utilizes two methods for XML reading. The first method loads the whole XML file into memory and uses an XML parser on the big document. This method is accurate, unfortunately it has exponential time complexity, so it can take hours for large XML files (over 10 MB). The second method uses regular expressions to read entries one by one from the XML file and then each single

entry is parsed. Entries are stored in the database with value of the specified key tag as a unique key. The administrator is informed about the import progress on the web page – a number of processed entries, a total number of entries, an estimated time till the end and last ten entry keys are displayed.

The administration module also supports export from database to plain XML file, the output files may be compressed to save disk space. The export also has an option to save the file in the form of a Ruby language script that will setup the database and import initial data. This is needed for the administration database itself. The output files are saved in a specified directory on the server and the administrator is informed about the export progress. Once the export ends, the administrator is offered a link to download the file through the web interface. The same function is used also for daily database backup.

Locking and Sequences of Identifiers The administration interface offers entry locking management to other DEB server modules. If multiple users can edit the database at the same time (which is one of the basic advantages of the client-server architecture), it is crucial to provide exclusive write locking of entries so that two users are not able to edit the same entry at a time. Decisions about entry locking depends on each application design.

An application then sends the request to the administration module which updates the lock database. The administration module provides several functions – besides simple lock and unlock functions, it can tell which user has locked a given entry, return the list of locks for selected user and/or dictionary or group several locks together if they are related. The administrator has access to the list of all locks and he or she can also delete chosen locks if the application did not release them correctly.

Newly created entries should have a unique identifier. If the application does not generate its own identifiers, the administration module can provide such service. It is possible to set an identifier pattern for each dictionary – this pattern looks like `CZE-[id]` and [id] will be replaced with sequentially increased number. The administrator can also affect the number used.

The Installation Packages The administration interface supports automated creation of Firefox Extension installation packages (XPI). If the administrator specifies a Relax NG schema for the dictionary, it is possible to transform this schema to an application design description in the XUL description language and the supporting code in JavaScript. The application created in this way supports basic forms – single and multiple text fields, select-boxes of specific values or relational links to other dictionaries. It can serve as basis for custom modifications. Of course, the application is able to connect to server, load data from server and save a modified entry back. We are currently working on more complex support for creation of new packages, mainly for the DEBVisDic client packages.

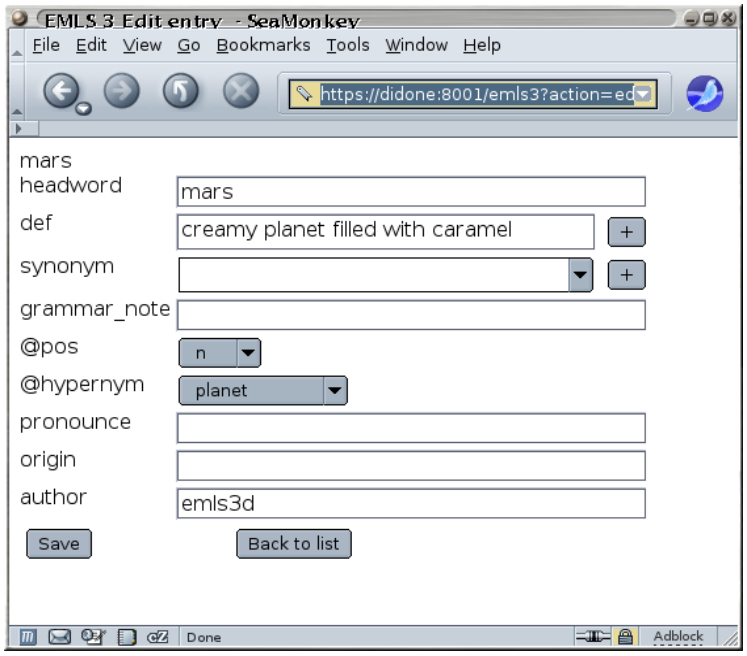


Fig. 4. Sample automatically build web service

Web-service generation For certain environments that either do not allow users to install new software packages or where the deployment of the software would be too time consuming, the DEB II server is able to generate simple web-service (see the Figure 4). The same as for XPI package generation, this function uses the dictionary Relax NG schema. To work with the dictionary, user needs only a web browser based on Gecko engine. All parts of the generated web-service are easily customisable with XSLT templates.

4 Automatic generation of a dictionary application

4.1 New dictionary definition

As a first step, the administrator needs to provide basic information about the dictionary. It does not matter if there is already an existing dictionary full of data, or whether the dictionary is going to be built from scratch. The administrator must specify an entry root element, where to find the unique key, several indexes and an XML schema of the entry.

As an example, we will describe the procedure of a new demonstration dictionary preparation from scratch. We will name the root element `entry` and have the unique key identifier in the element `/entry/headword`. The corresponding Relax NG schema is given in the Figure 5.

```
<element name="entry">
  <element name="headword">
    <attribute name="pos">
      <text/>
    </attribute>
    <text/>
  </element>
  <oneOrMore>
    <element name="sense">
      <text/>
    </element>
  </oneOrMore>
</element>
```

Fig. 5. The Relax NG schema of an example dictionary

This schema describes entry with one `headword` element, with `pos` attribute, and one or more `sense` elements. Of course, Relax NG supports description of much more complex XML structures.

4.2 Preparation of an installation package

The preparation of a new basic client application package requires selection of a dictionary and running the package generation. The administration module checks the Relax NG schema and finds all elements or attributes that contain `text` child element. All such elements and attributes are transformed to XUL textbox fields with the respective name as a label describing the field. If an element can occur multiple times in the entry (like `sense` in our Example), buttons for adding and removing the textbox are added to the application form, too.

The created JavaScript supports loading and saving documents and also searching for documents. The application thus enables querying each indexed field specified in the dictionary management interface. For example, users can easily find all nouns.

All the created application files are then packaged into the Firefox extension installation package (XPI). Users can download this package for installation or individual files for editing.

For the new client, there are also two basic preview templates (in XSLT) saved on the server side. One provides basic entry preview displaying all the data and the second displays raw XML data.

4.3 Application customization

Thanks to the design of applications based on the Mozilla development platform, these applications are easily customizable.

Any change in the layout and design of the form is done by editing the XUL (XML user interface Language) files accompanied with standard CSS stylesheets. The application logic (i.e. procedures implemented in JavaScript) stays the same for a new layout. Combination of XUL and CSS languages is very powerful and supports long list of features that are commonly used in desktop applications. For example, we can change PoS textbox field into a drop-down list, see the Figure 6.

a)
```
<textbox id="entry.headword.@pos"/>
```

b)
```
<menulist id="entry.headword.@pos">
  <menupopup>
    <menuitem label="noun"/>
    <menuitem label="verb"/>
    <menuitem label="adjective"/>
    <menuitem label="adverb"/>
  </menupopup>
</menulist>
```

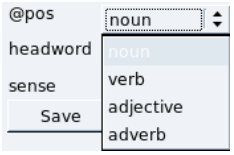


Fig. 6. Change of a textbox field to a drop-down list.

As we can see, the field labels contain element names only. This allows the application designer to change them to something human-readable. The actual texts are stored in a DTD (Document Type Definition) file as XML entities, so they can be adjusted to any texts in one place. Moreover, this mechanism is also used for localization of the application, see the Figure 7. It is possible to include several DTD files for different languages into installation package and (automatically) switch between them.

```
application.xul:          <label value="&entry.headword;"/>
en-US/application.dtd: <!ENTITY entry.headword "headword">
cs-CZ/application.dtd: <!ENTITY entry.headword "heslo">
```

Fig. 7. A field label and the respective entity in the localized DTD files.

After all the application source files are modified to meet the designer's requirements, he or she can upload them using the administration interface and let it build a new version of the installation package.

The application designer can also supplement the dictionary editor with more preview templates or modify the existing ones for different data presentation. When adding a new template, the template name must be added to the dictionary description in the database management interface. The modified templates are again uploaded to the server using the administration interface.

5 Conclusion

The presented DEB platform has already reached a deployment phase suitable for nearly ten full-featured dictionary writing applications used by more than 200 users in Czech Republic, Slovakia, the Netherlands, Poland, Hungary, Slovenia, South Africa and other countries.

The new features described in this paper provide complex functions that are shared by all parts of the DEB server such as user access rights handling or first step generation of a new dictionary application.

Acknowledgments. This work has been partly supported by the Academy of Sciences of Czech Republic under the projects T100300414 and T100300419, by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009.

References

1. Boitet, C., Mangeot-Lerebours, M., Sérasset, G.: The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons. In Wilcock, G., Ide, N., Romary, L., eds.: Proc. of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop, Taipei, Taiwan (2002) 93–96
2. Joffe, D., de Schryver, G.M.: TshwaneLex – professional off-the-shelf lexicography software. In: Third International Workshop on Dictionary Writing Systems: Program and List of Accepted Abstracts, Brno, Czech Republic, Masaryk University, Faculty of Informatics (2004)
3. Chaudhri, A.B., Rashid, A., Zicari, R., eds.: XML Data Management: Native XML and XML-Enabled Database Systems. Addison Wesley Professional (2003)
4. Feldt, K.: Programming Firefox: Building Rich Internet Applications with Xul. O'Reilly (2007)
5. Horák, A., Pala, K., Rambousek, A., Povolný, M.: First Version of New Client-Server Wordnet Browsing and Editing Tool. In: Proceedings of the Third International WordNet Conference - GWC 2006, Jeju, South Korea, Masaryk University, Brno (2006) 325–328
6. Pala, K., Horák, A.: From WEB Pages to Dictionary: a Languge-independent Dictionary Writing System. In: Proceedings of the 12th EURALEX International Congress, Turin, Italy (2006)
7. Horák, A., Pala, K., Rambousek, A., Rychlý, P.: New clients for dictionary writing on the DEB platform. In: DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems, Italy, Lexical Computing Ltd., U.K. (2006) 17–23
8. Santoso, Y.: Gnome's Guide to WEBrick (2004) (`http://microjet.ath.cx/WebWiki/WEBrick.html`).

Precision of Statistical Syllable Segmentation as a Function of Training Data Quality

Jozef Ivanecký[1] and Daniela Majchráková[2]

[1] European Media Laboratory GmbH, Heidelberg, Germany
Jozef.Ivanecky@eml-d.de
[2] Jazykovedný ústav Ľ. Štúra, Slovenská akadémia vied,
Bratislava, Slovakia
danam@korpus.juls.savba.sk

Abstract. The statistical approach for syllabic segmentation in Slovak seems to be easier to implement and gives better or at least similar results when compared with a rule-based system. The performance strongly depends on the quality as well as quantity of the training data. The proper test set is also very important. The paper describes our efforts to achieve the optimal error rate. We give a theoretical overview on training and testing techniques as well as a description of the real experiments with different selections of training and test data sets. The results lead to the conclusion that in case of limited training data, the selection of the data is particularly important.

1 Introduction

Although the automatic determination of the syllabic boundaries does not have many practical applications, it is necessary for automatic transcription in case the methods are based on the production rules instead of statistical approaches [2]. In Slovak language the pronunciation of *de, te ne, le* and *di, ti, ni, li* changes on syllabic boundaries. Therefore the detection of distinct boundaries is necessary for the correct determination of the pronunciation of *de, te, ne, le* and *di, ti, ni, li*.

Determining syllabic boundaries is complicated by the fact that the syllable definition is ambiguous. For our experiments we applied Pauliny's syllable definition which is primarily based on phonological principles [6]. Considering this fact, that syllable definition is not strictly defined, the specification of exact rules for syllabic segmentation is also difficult. Rule-based systems require sets of simple rules. Applying these rules we did not achieve better results than an 80%–85% success rate. Better results can be achieved only by using more complicated rules, but this correlates with an increase in exceptions.

The other problem is that in some words more than one syllable segmentation is possible which is also issued from the absence of strict rules. For example the word *bystrý* can be segmented as *by-strý, bys-trý* and *byst-rý*. In each case the syllabic segmentation is correct and the number of syllables is fixed.

In our experiment we had to reflect two facts:

- There are no exact rules for the determinination of syllabic boundaries in Slovak.
- In some cases, several different, though correct segmentations are possible.

Our first approach combined the rule-based syllabic segmentation together with a new approach which applies language model theory to the syllabic segmentation. Since we believe that a merely statistical approach is promising, we focused only on the improvement of the statistical model.

In the statistical approach we applied well known methods used in language modeling to the syllabic segmentation. In the case of language models the basic unit is a word. In our case these are syllables. Each word is first split into all possible sequences of syllables. For each syllable sequence $\mathbf{S}$, where

$$\mathbf{S} = s_1, s_2, \ldots, s_n \qquad s_i \in \xi \tag{1}$$

and ξ is the set of all possible syllables. Based on the Bayesian criterion, one can define the likelihood for each given sequence of syllables $\mathbf{S}$ as

$$P(\mathbf{S}) = \prod_{i=1}^{n} P(s_i \mid s_1, \ldots, s_{i-1}) \tag{2}$$

where $P(s_i \mid s_1, \ldots, s_{i-1})$ is the likelihood, that the syllable s_i follows after syllables $s_1, \ldots, s_{i-1}$. If we consider just two anterior syllables, the equation (2) can be rewritten as

$$P(\mathbf{S}) = \prod_{i=1}^{n} P(s_i \mid s_{i-2}, s_{i-1}) \tag{3}$$

To estimate the likelihoods $P(s_i \mid s_{i-2}, s_{i-1})$ we used the "syllabic corpus" created from the training set. The selection of the training set is discussed in Section 2. For the estimation we used the counts of syllabic sequences. As we mentioned above, we consider just two anterior syllables and thus (3) can be written as

$$P(s_3 \mid s_1, s_2) = f(s_3 \mid s_1, s_2) \doteq \frac{C(s_1, s_2, s_3)}{C(s_1, s_2)} \tag{4}$$

where $f(\ \mid\ )$ is the occurrence count function.

For two reasons equation (4) is not suitable for the likelihood estimation of a given syllable:

- for Slovak language monosyllabic and bisyllabic words are common
- not all sequences of syllables s_1, s_2, s_3 may occure in the training set

Based on the previous equation it is necessary to consider the likelihood $P(s_3 \mid s_1, s_2)$ as an interpolation of the count occurance for the sequence of three, two and one syllable:

$$P(s_3 \mid s_1, s_2) = \lambda_3 f(s_3 \mid s_1, s_2) + \lambda_2 f(s_3 \mid s_2) + \lambda_1 f(s_3) \tag{5}$$

where $\lambda_1 + \lambda_2 + \lambda_2 = 1$. The detailed description of the training as well as the testing process can be found in [3].

For further improvements of this technique we focus now on the selection of the training and test data sets. During the last years we found out, that training data selection is particularly important. Since data for the training has to be manually checked before one can use it, we were looking for the right pre-selection to minimize the human work.

The remainder of the paper is organized as follows: In Section 2 we give a brief overview on the data source for the training as well as testing set. In Section 3 we focus on the experiment design, the experiments and provide a brief summary in Section 4.

2 Data selection

Experimental results in [3] showed that the amount of training data is important. Besides the importance of the amount of the data, we believe that the quality of the data is also important. Since our ability to manually check huge amounts of data is limited, we decided to use first the 11 000 most frequent words from the Slovak National Corpus [1]. For the testing we also used the last 500 words from the corpus.

The Slovak National Corpus in version prim-3.0 has approximately 340 million tokens and contains Slovak language text(mainly journalistic) of many language styles. We extracted a list of the most frequent tokens from prim-3.0. As the list contained not only words but also many tokens like punctuation, one-letter words and abbreviations, we first had to modify the list to get a proper sample set for the experiment. Then we changed all characters to lower case and eliminated all duplicate words. The sample of the first 11 000 words from the corpus is representative of the most frequently used words and their case-forms in Slovak language. It also shows that the most frequent words in Slovak language are usually the shortest (mainly pronouns, conjunctions) and many frequent words appear in our list in all case-forms. Therefore the variability of the first 500 words is not very high.

The frequency sorting of the 11 000 words can be useful if we want to achieve the right segmentation of the most frequent words. Unfortunately it does not have to necesserily imply a good syllable coverage. The amount of syllables in the entire corpus is the same. By using just a fraction of the corpus, we wanted to know if it is better to use a corpus sorted by occurrence frequency or by randomly selected words.

Thus we created two lists from the 11 000 words. In the first list the words were sorted as they were in the corpus. The second one contained the same words but in random order. For each list we counted the number of different syllables after each 100 words. The number of unique syllables in our experiment was 2670. The results are shown in Figure 1.

From Figure 1. it is clear that randomly selected data has a better syllabic coverage. This is important information, as the number of words incorrectly

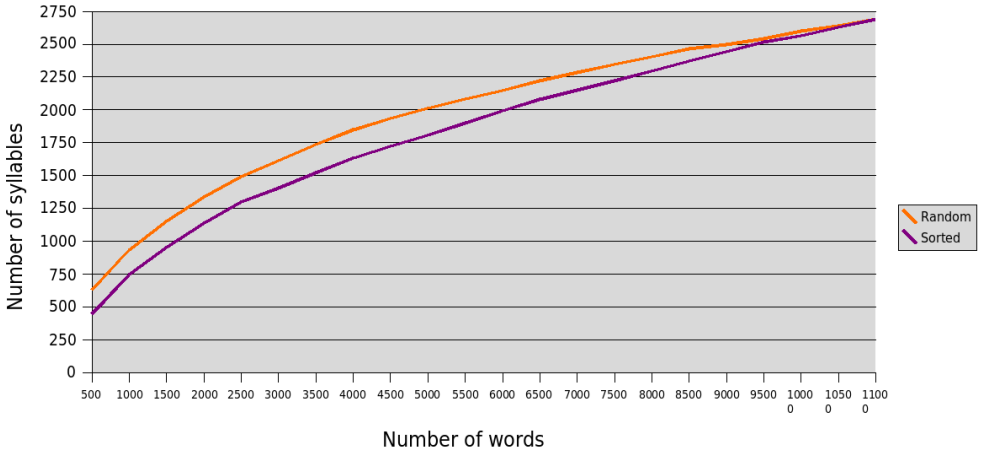


Fig. 1. Number of unique syllables as function of word amounts for a sorted and a random corpus.

segmented due to missing syllables in the training set is relatively high. Results in Section 3 demonstrate this conclusion.

3 Experiments

For the training data selection we designed several scenarios. We always used 10 000 words for the training and 1000 words for testing. In addition we also used the last 500 words from the corpus for testing. We tested all sets against the old system described in [3]. From 11 000 words the first 10 000 were used for the training and last 1000 for the testing. The same was done with the random list of 11 000 words. Therefore we ended up with 3 different systems for the syllabic segmentation as well as 3 different data sets for testing.

During the testing we generated the best sequence of syllables for all the words in the test set. We then compared the best syllable with the segmentation from the data preparation. If they did not match, we manually checked if it was another correct segmentation, an incorrect segmentation or an incorrect segmentation due to a missing syllable in the training set.

In contrast to the test scenario described in [3] we examined just the first (the best) segmentation and did not look at any other generated syllable sequences. The reason for this approach was the fact that if the tool for the segmentation is used in automatized process, it is not possible to determine if only the first or also the second segmentation is correct, or if the first one is not correct and the second one is. We manually checked all incorrect results to ensure proper classification. The initial results are in following table.

	1000 Sorted	1000 Random	Last 500
Sorted training set	17.30 %	7.20 %	30.00 %
Random training set	8.00 %	13.80 %	27.00 %
Old training set	19.20 %	14.60 %	23.20 %

From the table it is clear that results for the last 500 words have a much lower variance than the results for the first two test sets. The explanation is very simple, and also explains why there are 2 results which are much better then the rest. The combination of the sorted training and random test as well as the random training and sorted test should not be part of the test set. In these two particular combinations the test set contains part of the training data. This is the reason why these two results need to be excluded from the results table. These two results point us indeed to the syllable coverage in the test set. This is the main reason why two excluded results are much better then the others. In the following table the number of syllables from the test set is not covered by the training set as well as the percentage of words influenced by missing syllable. The sorted training set contained 2424 unique syllables, the random training set 2459 and finally the old training set 3009 syllables. It is necessary to point out that in the old training set foreign words were not excluded.

	1000 Sorted	1000 Random	Last 500
Sorted training set	116 (10.1%)	9 (0.6%)	139 (20.8%)
Random training set	8 (0.5%)	81 (5.8%)	133 (19.9%)
Old training set	127 (11.7%)	107 (8.3%)	68 (11.8%)

As we can see, the best syllabic coverage is for the "excluded" combinations. In both cases just less then 1% of syllables from the test set are not covered by the training set. From the syllable coverage table it is also clear that there is some relation between word occurrence frequency and syllabic occurrence frequency. 20% of syllables from the last 500 words are not covered by the training set. On average it is 2 times more than for words from the beginning of the corpus.

When we excluded all words containing syllables from the test set which are not covered by the training set, the results of the syllabic segmentation were as follows:

	1000 Sorted	1000 Random	Last 500
Sorted training set	8.00 %	6.63 %	11.61 %
Random training set	7.53 %	8.49 %	8.75 %
Old training set	8.49 %	6.87 %	12.92 %

We can see that the result variance for each test set is lower than in the first results table. But unallowed combinations (sorted–random, random–sorted) are also better than allowed combinations here. From the achieved results we can derive the following conclusions:

– Less frequent syllables appear more often in less frequent words than in common words. The last 500 words test continually showed worse accuracy and had the biggest number of syllables not covered.

- Sorted training data give better performance for more frequent words, but the words from the end of the corpus are significantly worse.
- The amount of training data is still not sufficient. With the current training set almost every second error is caused by syllables not covered by the training data.

To improve the performance of the system the most important thing seems to be to increase the amount of training data. The random selection of the data seems to be more suitable, but this advantage may disappear when the amount of data used for the training is increased several times.

4 Summary

In this paper we described the influence of training data selection for statistical syllabic segmentation to the overall performance of the syllabic segmentation process. We showed that to have better coverage for words from the entire corpus it is better to use random selection of the training words rather than words selected by their occurrence frequency. More important than data selection seems to still be the amount of data. Our experiments confirmed initial assumptions that to achieve acceptable results, the amount of training data has to be increased a few times.

Acknowledgement The authors wish to thank the Slovak National Corpus team for access to the Slovak National Corpus database.

References

1. Garabík, R. Gianitsová, L., Horák, A., Šimková, M., Šmotlák, M.: *Slovak National Corpus*. In: Proceedings of the conference TSD 2004. Brno, Czech Republic: Springer-Verlag, 2004
2. Ivanecký, J.: *Analysis of the Rule Based Phonetic Transcription Technique Applied to the Slovak Language*, Slovko 2005, Bratislava, 2005
3. Ivanecký, J.: *Štatistický prístup pri určovaní slabičných hraníc*, Slovko 2003, Bratislava, 2003
4. Jelinek F.: *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, Massachusetts, London 1998, 283 s.
5. Kráľ A.: *Pravidlá slovenskej výslovnosti*, Slovenské pedagogické nakladateľstvo, Bratislava 1983, 632 s.
6. Pauliny E.: *Fonológia spisovnej slovenčiny*, Slovenské pedagogické nakladateľstvo, Bratislava 1968
7. Rabiner L., Juang B.-H.:*Fundamental of Speech Recognition*, Prentice Hall, New Jersey, 1993, 507 s.

Program Concorder and Jaroslav Seifert's Individual Dictionary

Ladislav Janovec and Martin Wagenknecht

KČJL UK PedF
Prague, Czech Republic
ladis.janovec@seznam.cz

Abstract. Our contribution describes the process of building the frequency dictionary of Jaroslav Seifert´s Poetry. This project has been running since 1997 with a couple of little breaks. We have used a software called Concorder, which is applied on Apple Macintosh computers. This software was created by specialists from McGail University – David Rande and Tatyana Patera who built some frequency dictionaries of Russian authors (A. Achmatova, Y. Brodskij). For non-commercial usage, Concorder was afforded to scholars of Czech Language Department (UK PedF) who, with the help of Czech language students, began to build three dictionaries – Seifert´s one is the broadest one. Three researchers (S. Machová, L. Haasová, L. Janovec) have already presented and published some results of the project on several linguistic conferences. The software enables us to create a primary catalogue of word forms and their frequencies from the text of a poetry set. Then there are some different ways to create a lemmatized catalogue, eventually, add other characteristics of items to suit the linguist´s interests and intention. Concorder can unit particular catalogues into one whole. In addition, it is possible to extract some special catalogues, such as eg. catalogue of nouns, adjectives, the most frequent units etc.

1 About Concorder and the conception of the dictionary

This contribution is connected with a long-term project of Jaroslav Seifert's individual dictionary of poetry which has been created on The Czech Language Department of Faculty of Education, Charles' University. Works on the project started in 1997 (as a grant project *Slovník básnického díla Jaroslava Seiferta*), the researchers were three Czech Language students and S. Machová, a senior lecturer of the department, as the leader and project coordinator.

Individual dictionaries represent the possibility of studying and analyzing the author's language from the point of view of his word-stock used in his texts. These dictionaries help to observe and compare the author´s lexicon independent of his texts, to acquire some linguistical coeficients (such as frequency, word length, index of word repetition) or, in some cases, they can help to establish the author of an anonymous text. However, the most important

disadvantage of individual dictionaries is that semantic information composed on the text level is lost very often.[1]

Before computerizing linguistics those dictionaries used to be created manually (eg. SJP, LS, Vinogradov 1965[2]), but with the development of computer technologies the authors of dictionaries can make their work faster and more effective (eg. Patera 1995, 1997).

No large complete individual dictionaries of Czech authors have been published[3] – the most important base for the individual dictionary of Vítězslav Nezval was created in 1970s by J. Levý, but after his death it disappeared. Czech indvidual lexicography is represented only with some small dictionaries such as Kvítková 2001. As we know, only one individual dictionary of a Czech writer, created in Germany, was published abroad – it was a dictionary of Otokar Březina (Holman 1993)[4].

In 1996 University of Montreal afforded the Czech Language Deparment (Faculty of Education, Charles' University, Prague) a concordance software called Concorder for incommercial usage. This software is applied on Macintosh and it helps building individual dictionaries. The first, base version of this software was defined for English dicitonaries. Then the authors[5] created some modification defined for other languages including Slavonic languages[6].

The researchers of Czech Language Department prepared and defanded three projects of qunatitative characteristics of writers language – *Index of the Lexems and Forms in the manuscripts of So-called Dalimil's Chronicle* (N. Kvítková), *The Individual Dictionary of Jaroslav Seifert's Poetry* (S. Machová) and *The Language of František Nepil* (R. Brabcová), but the third was not planned to be crowned with an individual dictionary. The second one was already published (Kvítková 2001) and work on the Seifert's dictionary will have been finished in two years.

1 For example, the dictionaries could not take into consideration metaphorical usage given in wide context

2 More about some dictionaries, their conceptions and characteristics see Janovec 1999a.

3 There exist some studies concerning authors‘ individual language characteristics from the quantitative point of view (such as Těšitelová 1948, Kučera 1992, 1994, Fiala 1996, Šafaříková 2002)

4 Authors know that the unpublished text of the dictionary of Marie Pujmanová's trilogy (Lidé na křižovatce, Hra s ohněm, Život proti smrti), deposited on Czech Language Department of St. Peterburg University, exists as well.

5 The authors of software are D. W. Rand, Centre des Recherchers Mathématiques, Montréal, and T. Patera, Department of Russian and Slavic Studies, McGill University, Montréal).

6 The first individual dictionaries of Slavonic writers built on Concorder were Patera 1995, 1997).

Working on mentioned project the team had to solve some problems connected with information which should be note for every item and weigh possibilities which the software offers its users.

First, the transformation of the text into electronically readable text form had to be decided. Linguists have three ways to make and obtain these data – using electronic type (rarely given by publishers), scanning text and make a useful format using other software or (in the case of small corporas) typing text manually. At the beginning of work we decided to type the texts by ourselves, because we didn't have possibility to scan texts. The advantage of manual typing was finding soon some multiverbal expressions which are analysed as a whole (eg. names, idioms, phrasemes or terms, such as *retranslační věž, šípková růže, Figarova svatba*). The disadvantage of manual typing was a high number of errors in texts influenced by mistyping. The advantage of scanning text in next phases of this project was saving time, but the control of the final text had to be very careful as well because of the automatical misreading of some words and the multiverbal expressions were much harder to find.

Having the electronical text version, readable in Concorder, a researcher created automatically a primary catalogue (catalogue of word forms) which consists of four standardised columns (six other columns could be added if needed). Two columns are filled by Concorder – the second one which contains word forms and the third one which contains frequency of word forms. The linguist must fill other columns – the first one is established for lemmas (the representative forms of lexems), the fourth for parts of speech.

When one sees that it is necessary to make some changes in the text during lemmatization of the primary catalogue, he/she must create a new catalogue from the modified text, but information from the old catalogue can be transferred into the new one.

Concorder offers the means of automatical lemmatization of a new catalogue under some conditions. First, the fully lemmatized catalogue must already exist. Second, information of parts of speech must be reduced and then lemmas for word forms which are common in both catalogues can be transfered to a new created catalogue. This function of Concorder is not effective – it is not possible to transfer the information about parts of speech – all in all, it must be filled manually. The transfer of lemmas took pretty much time (when we tried to lemmatize automatically two collections of poems of middle extent – 2300 items and 3030 items, the lemmatization take twenty five minutes). As only the common word forms in both texts could be lemmatized, the successfulness was very low – 6-9%, which could be caused by the richness of a poet's language – there are not so many word forms repeating in poems.

Next problem was connected with homonymy. If there were some homonyms in the catalogue, Concorder lemmatized them by chance and didn't split them (about splitting see below).

When the primary catalogue is lemmatized, the linguists create a united catalogue (three columns), where the word forms are united under common lemma and their frequencies are summed up. The united catalogue can be organised by frequency (frequency order) or by alphabet (alphabetical order) and it can be a source for sorting subcatalogues of parts of speech.

2 Building the dictionary

As it was said, when building a dictionary, it is necessary to make a primary catalogue from the text of chosen poetry book first. The catalogue consists of four columns, and there are word forms sorted alphabetically (the second column) and by their frequency (third column). We fill the first column (lemmas) and the fourth one (parts of speech). A part of speech is marked with an abbreviation of its term – *n* (nouns), *adj* (adjectives), *pron* (pronouns), *num* (numerals), *v* (verbs), *adv* (adverbs), *prep* (prepositions), *conj* (conujnctions), *part* (particles), *i* (interjections). With the abbreviation of nouns we write a gender difference too (*nM* – masculines, *nF* – feminines, *nN* – neutral). If the noun is a proper noun, we marked this information as an abbreviaton */vl*, eg. *Figaro = nM/vl*. When a noun is a geographic name, we sign it as *g* following */vl*, e.g. *Kralupy = nF/vlg*. For foreign or unclear expressions we use abbreviation *cv* (meaning „cizí výraz").

For the form of lemmas, most often, we use their neutral literary representative forms of lexems, following the forms in *SSJČ*, because it is the representative dictionary of Czech, containing many poetisms and rare words (often excerpted from Czech poetry). The current literary form for a lemma we also choose in case when Seifert used old literary form or non-standard word form, so *neřek* is lemmatized as *říci*, *zúřit* is lemmatized as *zuřit*, but there is not waste of information – it is already fixed in the column of word forms.

The infinitive forms of verbs are linked to this problem, we had to decide between their endings *–t* X *–ti* and *–ct* X *–ci*. We use the ending *–t* in dictionary (even if the poet uses *–ti*) not to enlarge the variant lemmas. For infinitives, such as *péci, moci*, we use the ending *–ci*, because we find the progressive *–ct* too informal for Seifert's poetics.

We differ perfective (completive) and imperfective verbs, each has its own lemma (*plakat* X *plakávat*). We do not pay attention to negative verbs (except modal verbs, where we can find some semantic specifications – *nemoci* X *moci*, *nemuset* X *muset*), which we include under the positive lemma (*nemilovat* and *milovat* both have the lemma *milovat*, but *nemoci* X *moci* are two lemmas).

Adjectives and pronouns which differ in gender we include under the masculine lemma form (which is basic), so all of the pronouns *ten, ta, to, ti* have lemma *ten*. This principle follows the conception of old frequency Czech

language dictionary (Jelínek – Bečka – Těšitelová 1961[7]). Adjective and adverb gradation does not project into lemmas (comparatives and superlatives are included under the positive adjective form – *nejbližší – blízký, časněji – časně*).

It is difficult to differ adverbs and particles in some cases. If we have different opinions, we lemmatize it as adverbs. In the case that the word form seems to be the second temporal adverbial (eg. *Už v sedm hodin* apod.) we prefer to lemmatize the word form *už* as *už2 particle*).

At the very beginning we wanted to differ lexems from the semantic point of view (homonymy and polysemy), but during work we found it very difficult – so we differ only homonymy of parts of speech (eg. *místo* can be *nN* or *prep*). For this distribution Concoder offers a means called splitting which helps to index different kinds of items dipending on our needs.

It is necessary to dispose of a homonymy list with indices which is updated very often by team members and they must follow it not to mix up different parts of speech, eg. verb *být* has three indices – *být 1* is a part of verbonominal predicate, *být 2* is autosemantic *být* and *být 3* is an auxiliary verb, or *mrtvý 1* is a adj, *mrtvý 2* is a nM.

As we wrote above complex of words and proprial nouns, such as *retranslační věž, šípková růže, Tycho de Brahe, Figarova svatba* etc., and idioms we find as a one lexical unit. The part of speech we classify from the syntactic point of view (the syntactic base of expressions represents also the part of speech – *šípková růže – nF, Figarova svatba – nF* etc.).

3 Some frequency characteristics (based on the lexical material of collection *Koncert na ostrově*)

Number of all word forms	5475
Number of different word forms	2630
Number of lemmas	1845
Index of repeating words	2.9674

7 The new frequency dictionary was published in 2005 so in the beginning of this project we could not take it into account.

	1	**2 (%)**	**3**	**4 (%)**	**5**
n	1526	27,87	763	41,36	2,0000
adj	446	8,15	269	14,58	1,6580
pron	581	10,61	35	1,90	16,6000
num	40	0,73	20	1,08	2,0000
v	1048	19,14	499	27,05	2,1002
adv	611	11,16	171	9,27	3,5731
prep	537	9,81	24	1,30	22,3750
conj	504	9,21	26	1,41	19,3846
part	167	3,05	27	1,46	6,1852
i	10	0,18	7	0,38	1,4286
cv	5	0,09	4	0,22	1,2500

Table 1.

1 – number of all word forms of particular part of speech
2 – ratio of used word forms of particular part of speech
3 – number of lemmas of particular part of speech
4 – ratio of lemmas of particular part of speech
5 – index of repeating

The most frequent lexems in the collection:

a1 (205), ten (138), být1 (127), na (117), on (77). v (74), já (70), a2 (69), být2 (67), když (67), už1 (66), z (57), který (53), jen1 (40), s (40)

The most frequent lexems in the collection according to parts of speech:

Nouns. oko (22), smrt (19), píseň (18), láska (17), rok (14), ruka (13), chvíle (12), okno (12), žena (12), čas (11), člověk (11), svět (11), tma (11), mrtvý2 (10), život (10)

Adjectives. krásný (12), starý (11), černý (10), jiný (8), celý (7), rád (7), růžový (6), velký (6), bílý (5), malý (5), modrý (5), mrtvý1 (5), prázdný (5), smutný (5), šťastný (5), těžký (5), vlastní (5), zlý (5)

Pronouns. ten (138), on (77), já (70), který (53), svůj (36), jeho (35), co1 (34), všechen (28), my (17), nic (13), můj (10), vy (9), kdo (8), někdo (8), ten (7),

Numerals. dva (7), jeden (4), oba (3), první (3), sedmdesátý (3), tolik (3), tři (3), dvacet (2)

Verbs. být1 (127), být2 (67), mít (38), říci (17), jít (16), padat (14), chtít (13), vědět (12), moci (11), čekat (9), tancovat (9), dát (8), dívat se (7), plakat (7), muset (6), patřit (6), spát (6)

Adverbs. už1 (66), jen1 (40), ještě (37), pak1 (23), tak1 (20), jak1 (17), tu (17), až2 (13), také1 (13), kde (11), tam (11), již (10), kdy (10), tenkrát (9), jenom1 (8), možná (8), dávno (7), někdy (7)

Prepositions. na (117), v (74), z (57), s (40), do (39), o (37), po (28), k (24), pod (15), za (15), ve (13), nad (12), před (12), bez (11), od (8), u (8), ze (8)

Conjunctions. a1 (205), když (67), i1 (33), ale1 (31), jako2 (31), však2 (30), že1 (24), aby1 (23), než (12), kdyby1 (11), až1 (7), jak2 (4), nebo1 (4), neboť (4), protože (4)

Particles. a2 (69), i2 (14), ať2 (11), snad (10), ani2 (6), ne (6), už2 (6), však1 (6), ano (5), ale2 (3), asi (3), jen2 (3), jenom2 (3), kdyby2 (3), alespoň (2), aspoň (2), tak2 (2), tedy1 (2), třeba (2), vždyť 2 (2)

Interjections. ach (3), proboha (2), amen (1), basama s fousama (1), dobrý večer (1), k čertu (1), sbohem (1)

4 Some frequency characteristics (based on the lexical material of collection *Samá láska*)

Number of all word forms	6463
Number of different word forms	2888
Number of lemmas	1908
Index of repeating words	3,3873

	1	**2** (%)	**3**	**4** (%)	**5**
n	1926	29,80	774	40,57	2,4884
adj	659	10,20	324	16,98	2,0340
pron	693	10,72	39	2,04	17,7692
num	55	0,85	26	1,36	2,1154
v	1130	17,48	458	24,00	2,4672
adv	514	7,95	480	25,16	1,0708
prep	727	11,25	29	1,52	25,0690
conj	642	9,93	35	1,83	18,3429
part	84	1,30	26	1,36	3,2308
i	32	0,50	16	0,84	2,0000
cv	1	0,02	1	0,05	1,0000

Table 2.

1 – number of all word forms of particular part of speech

2 – ratio of used word forms of particular part of speech

3 – number of lemmas of particular part of speech

4 – ratio of lemmas of particular part of speech

5 – index of repeating

The most frequent lexems in the collection:

a1 (315), na (158), být1 (155), v (140), on (79), svůj (63), ten (58), když (56), do (55), to (53), k (51), já (50), z (49), že1 (48), být2 (43), ty (43)

The most frequent lexems in the collection according to parts of speech:

Nouns. láska (54), ruka (35), hvězda (27), svět (27), člověk (26), oko (24), krejčí (22), noc (22), píseň (21), ulice (20), žena (18), den (17), dítě (17), okno (17), sen (16), srdce (16)

Adjectives. krásný (29), rád (18), celý (17), bílý (16), černý (15), smutný (14), nový (10), jiný (8), modrý (8), plný (8), veliký (8), malý (7), těžký (7), tichý (7), dlouhý (6), dobrý (6), kamenný (6), sladký (6), zelený (6), železný (6)

Pronouns. on (79), svůj (63), ten (58), to (53), já (50), ty (43), který (36), můj (35), jeho (34), jenž (34), my (30), každý (18), všecko (17), tvůj (13), co1 (12) vy (12)

Numerals. tisíc (8), jeden (6), jedenáct (4), tolik (4), dvanáctý (3), oba (3), čtyři (2), druhý (2), dva (2), dvě (2), pátý (2), stokráte (2), třináct (2)

Verbs. být1 (155), být2 (43), mít (35), chtít (22), dát (14), vědět (14), jít (13), muset (13), pět (13), stát (13), chodit (12), milovat (12), zpívat (12), moci (11), plakat (11)

Adverbs. tak1 (26), tam (24), už1 (23), kde (19), jak1 (16), dnes (13), jen1 (13), večer (13), ještě1 (12), již (10), vždycky (8), zase (8), jednou (7), také1 (7), jistě (6), někde (6), nikdy (6), pak1 (6), uprostřed (6), však1 (6)

Prepositions. na (158), v (140), do (55), k (51), z (49), s (34) za (30), nad (29), po (26), pro (25), o (23), v (22), ze (18), mezi (12), u (11)

Conjunctions. a1 (315), když (56), že1 (48), jako2 (42), i1 (18), až1 (17), aby1 (14), však2 (14), neboť (13), ale1 (11), kdyby1 (11), než (9), vždyť1 (9), jak2 (8), ani1 (6), nežli (6)

Particles. a2 (14), až2 (9), proč (8), ani2 (7), ale2 (5), ať2 (5), snad (5), nechť2 (4), asi (3), přece2 (3), aspoň (2), dokonce (2), jen2 (2), ještě2 (2), možná2 (2)

Interjections. brnk (5), brnky (5), ó (4), bože (2), břink (2), sbohem (2), tralá (2), tralalalá (2), ach (1), haló (1), ježíši kriste (1), nuž (1), tarara (1), tram (1), viď (1), viďte (1)

Foreign expressions. Pére Lachaise (1)

5 Sample of concordances based on the lexical material of collection *Kamenný most*)

Every concordance starts with a number which identifies the sequence of a word in texts.

MRTVÝ 2 3 n
5.175 do podloubí, spi sladce každý, kdo už spíš! Jen **mrtví** bdí tu, mrtví střehou tmu času širou, bezebřehou, dvě pěsti k nebi

5.175 do podloubí, spi sladce každý, kdo už spíš! Jen mrtví bdí tu, **mrtví** střehou tmu času širou, bezebřehou, dvě pěsti k nebi

5.164 prach jak pel, jejž z květů nadnáší vítr ve svůj let, by mohl **mrtvý** oplodniti i živé za tisíce let. Čas ve svém

POVÍDALI ŽE MU HRÁLI 1 v
2 než usednou na ledolamu. Nad věžemi se hvězdy smály, ach,
povídali~že~mu~hráli, copak se hvězda smáti může? Pod jezem

6 Finishing project

Because of finishing lemmatization of primary catalogues of all Seifert's poetry we are preparing a finish product. It consists of a fused catalogue which contains all particular lemmatized catalogues. After it we are going to present the result as a CD and a paper version.

References

1. Brabcová, Radoslava (1997): O jazyku Františka Nepila. In: Filologické studie XX, Karolinum, Praha, p. 20-28.
2. Fiala, Jiří (1996): *Erbenova Kytice v počítači.* Česká literatura, 44, p. 656-657.
3. Haasová, Lenka (1998): *Frekvenční slovník básnického díla Jaroslava Seiferta.* In: Sutdentská jazykovědná konference 1997, Ostravská univerzita, Ostrava, p. 73-77.
4. Haasová, Lenka (1999a): *Autorský slovník Jaroslava Seiferta.* Diploma thesis. UK PedF, Praha.
5. Haasová, Lenka (1999b): *Pokus o stanovení sémantických tříd v díle Jaroslava Seiferta.* In: Varia VIII, Slovenská jazykovedná spoločnosť při SAV, Bratislava, p. 75-81.
6. Holman, Petr (1993): *Frequenzwörtebuch des lyrischen Werkes von Otokar Březina. Teil I, II.* Köln – Wiema – Wien.
7. Janovec, Ladislav (1998): *Využití programu Concorder pro tvorbu autorských slovníků.* In: Varia VII, Slovenská jazykovedná spoločnosť při SAV, Bratislava, p. 63-70.
8. Janovec, Ladislav (1999a): *Frekvenční slovník pozdní básnické sbírky Jaroslava Seiferta.* Diploma thesis. UK PedF, Praha.
9. Janovec, Ladislav (1999b): *Z autorského slovníku Jaroslava Seiferta – sémantická charakteristika a použití zájmen já a my v některých sbírkách.* In: Varia VIII, Slovenská jazykovedná spoločnosť při SAV, Bratislava, p. 82-91.

10. Janovec, Ladislav (2002): *Sémantická charakteristika barev a jejich použití v díle Jaroslava Seiferta – pokus o lingvistickou interpretaci.* In: Varia IX, Slovenská jazykovedná spoločnosť při SAV, Bratislava, p. 37-42.
11. Janovec, Ladislav (2003): *Autorský slovník Jaroslava Seiferta – současný stav, změny koncepce a budoucnost projektu.* In: Varia X, Slovenská jazykovedná spoločnosť při SAV, Bratislava, p. 197-201.
12. Jelínek, J. – Bečka, J. V. – Těšitelová, M.: *Frekvence slov, slovních druhů a tvarů v českém jazyce.* SPN, Praha, 1961.
13. Kvítková, Naděžda (2001): *Staročeský text z kvantitativního hlediska.* Univerzita Karlova, Praha.
14. LS: Častotnyj slovar´jazyka M. Ju. Lermontova. In: Lermontovskaja encyklopedia. Moskva, 1981, p. 717-774.
15. Machová, Svatava (1997a): *Autorský slovník básnického díla Jaroslava Seiferta I. část.* In: Filologické studie XX, Karolinum, Praha, p. 29-34.
16. Machová, Svatava (1997b): *Slovník básnického díla Jaroslava Seiferta.* Final report of 1st phase of project, UK PedF, Praha.
17. Machová, Svatava (2000): *Slovník básnického díla Jaroslava Seiferta.* Final report of 2nd phase of project, UK PedF, Praha.
18. Machová, Svatava (2002): *Slovník básnického díla Jaroslava Seiferta.* Final report of 3rd phase of project, UK PedF, Praha.
19. Patera, T. (1995): *A concordance to the poetry of Anna Achmatova.* Ann Arbor.
20. Patera, T. (1997): *A concordance to the poetry of Joseph Brodsky.* Ann Arbor.
21. Podráská, Eva (1999): O jazyku Františka Nepila II. In: Varia VIII, Slovenská jazykovedná spoločnosť při SAV, Bratislava, p. 70-75.
22. Rand, David W.: *Concorder. Concordance software for the Macintosh. User's Manual.* Centre de Recherches Mathématiques, Montréal.
23. SJP: Slovar' jazyka Puškina, I-IV. Moskva, 1956-1967.
24. SSJČ: *Slovník spisovného jazyka českého.* Academia, Praha, 1960–1971.
25. Šafaříková, Linda (2002): *Cambridžský rukopis Dalimilovy kroniky z kvantitativního hlediska.* Diploma thesis. UK PedF, Praha.
26. Těšitelová, Marie (1948): *Frekvence slov a tvarů ve spise „Život a dílo skladatele Foltýna" od Karla Čapka.* Naše řeč, 32, p. 126-130.
27. Těšitelová, Marie (1968): *O básnickém jazyce z hlediska statistického.* Slovo a slovesnost, 29, p. 362-368.
28. Těšitelová, Marie (1987): *Kvantitativní lingvistika.* SPN, Praha.
29. Těšitelová, Marie (2000): *K současné české próze z hlediska frekvence slov.* Naše řeč, 83, p. 1-9.
30. Vinogradov, V. L. (1965): *Slovar'-spravočnik „Slova o polku Igoreve".* Nauka, Moskva.
31. Wagenknecht, M. (2007): *Plnovýznamové sloveso být v básnické sbírce Jaroslava Seiferta Osm dní.* A paper presented on a student conference, Wroclaw (in print).

Collocations in Russian: Analysis of Association Measures

Maria Khokhlova[1,2)]

[1)] Department of Mathematical Linguistics
Philological Faculty, St. Petersburg State University
St. Petersburg, Russia
[2)] Institute for Linguistic Studies
St. Petersburg, Russia
vertikal-maria@yandex.ru

Abstract. The notion of collocation is quite ambiguous. A concise survey of different approaches to it (British contextualism, lexicographical approach, approach of the "Meaning-Text" theory) is proposed in the paper. The paper evaluates results of retrieving collocations from a corpus of Russian texts. It also discusses the issue of presentation of information about collocations in modern Russian dictionaries.

1 Introduction

The methods for collocation extraction proposed in most works have not been evaluated so far whether they can be applicable to Russian, and if yes, to what degree. Also there's a question what types of set phrases they allow to retrieve. The explanatory dictionaries do not always consecutively reflect the information about set phrases. The boundary between free and set phrases is quite ambiguous.

According to some scientists [1] the property of stability (for phrases) is inherent to all word combinations. A threshold of stability should be chosen to range them, above which a word combination can be called a set phrase.

The term "collocation" has come to use in Russian linguistics, after Western linguistics, to designate set phrases. Although the term itself appeared long ago [2], it is not generally recognized by Russian scholars. Such language units have various names in different works; cf. "set verbal-noun expressions" [3], "analytic lexical collocations" [4] etc. The majority of authors understand under collocation a statistically set phrase. Collocations can be put between free phrases and idioms on a scale of phrases.

At first the notion of collocation was introduced by the founder of London School of Structural Linguistics and the representative of British contextualism J. R. Firth [5]. The word meaning, in Firth's opinion, is closely connected with its ability to collocability. Collocation is a tendency of a word to a certain environment. So, he stated the hypothesis according to which it is possible for a word to be attributed to a group by its neighbourhood. The parts of collocation

occupy certain positions and, thus, are characterized by mutual expectancy of appearance. Collocations can be viewed as forms of meaning [5].

It is possible to allocate also the lexicographic approach to studying the phenomenon of collocation. While in British contextualism collocation is defined on the basis of statistical assumptions about the probability of co-occurrence of two (or more) lexemes, and especially frequent combinations of lexical units are considered as collocations, the lexicographic approach considers collocation as semantic-syntactic unit or a combination of lexically defined elements of grammatical structures.

2 The notion of "collocation" in Russian linguistics

The monograph [6] has proved to be the first work in Russian linguistics, completely devoted to the research of the concept of collocation on a material of Russian. One of the key properties of collocation is "the impossibility of prediction of such combinations on the basis of meanings of their components" [6: 13].

Another classification of collocations is given in [4]. Under the term "collocation" Teliya understands the combination characterized by a nominative regularity, i.e. due to the bound component it has the ability to designate the senses possessing the content of common category, "typical of aspectual and temporal meanings and also of meanings correlating with semantic cases of deep structure (in the sense of Fillmore [7])" [4]. In Teliya's opinion, it is this principle that underlies lexical functions of the "Meaning – Text " theory. For example, *byt' ne v nastrojenii* = "to be in bad mood" (cf. *byt' v dome* = "to be at home"), *luch nadezhdy* = "a ray of hope" (cf. *luch sveta* = "a ray of the sun"), *kormilo vlasti* = "at the helm" (cf. *kormilo korablya* = "helm of a ship") etc.

In the "Meaning – Text" theory collocations are considered as a subclass of more extensive class of set phrases, or phrasemes. "The idiom is an expression consisting of several lexemes which meaning cannot be completely deduced by general rules of the given language from meanings of its constituent lexemes, from morphological characteristics (if those are available) assigned to them semantically and from their syntactic configuration" [8: 215].

According to Meľchuk and Teliya, collocations can be understood as word-combinations in which one of the elements is viewed a semantic dominant, and another is chosen depending on it in order to express the sense of all combination (the same approach is adhered by M. Hausmann, A. Cowie[1], S. Kahane and A. Polguère). The dependent word, thus, can be interpreted only in combination with the dominant. The similar standpoint we find in [6].

1 A. Cowie calls such combinations *restricted collocations*.

3 The analysis of retrieving collocations in Russian

Nowadays there are several ways in linguistics to calculate the degree of coherence of parts of the collocation. They are based on comparison of frequencies for word pairs received on a material of a real corpus with independent (relative) frequencies. Statistically significant deviations of real frequencies from hypothetical probabilities (for more details see [9]) are searched.

Statistical methods for data treatment are widely used in corpus linguistics. There are different measures based on calculation of a degree of nearness of words in a text, namely, MI (mutual information), t-score, Log-Likelihood, z-score, chi-square.

The object of research in the given work is collocations of Russian, their presentation in dictionaries of modern Russian.

The aim was to carry out a number of experiments in order to find a suitable measure of association for different classes of set phrases; to define opportunities of statistical methods as a whole and several measures in particular; to find ways of a combination of statistical and semantic-syntactical methods in collocation retrieving.

We have leaded a series of experiments with the purpose of comparison the efficiency of statistical methods.

During experiment the following ideas were tested:

- to what degree the proposed methods can be applicable to Russian;
- whether the given methods allow to reveal other classes of set phrases.

We have chosen collocations of 19 nouns that don't have homonyms as material for our research. The nouns have been selected by their sufficient high frequency (see the electronic frequency dictionary of Russian by A. Sharoff [10]): *власть* "power", *внимание* "attention", *возможность* "opportunity", *война* "war", *вопрос* "question", *дождь* "rain", *жизнь* "life", *закон* "law", *любовь* "love", *место* "place", *мнение* "opinion", *мысль* "thought", *ночь* "night", *ответ* "answer", *помощь* "help", *радость* "joy", *слово* "word", *случай* "case", *смысл* "sense".

The research has been lead on the corpus of Russian newspapers created at the University of Leeds (Great Britain) [2] under the guidance of S. Sharoff. This corpus includes nearby 78 million words from several major Russian newspapers (for example, "Izvestia"), its part-of-speech tagging was done using the program Mystem [3].

In a search mode one can choose one or several statistical measures (MI, t-score, log-likelihood), set a span in words, and also it is possible to set a part of speech of a collocate.

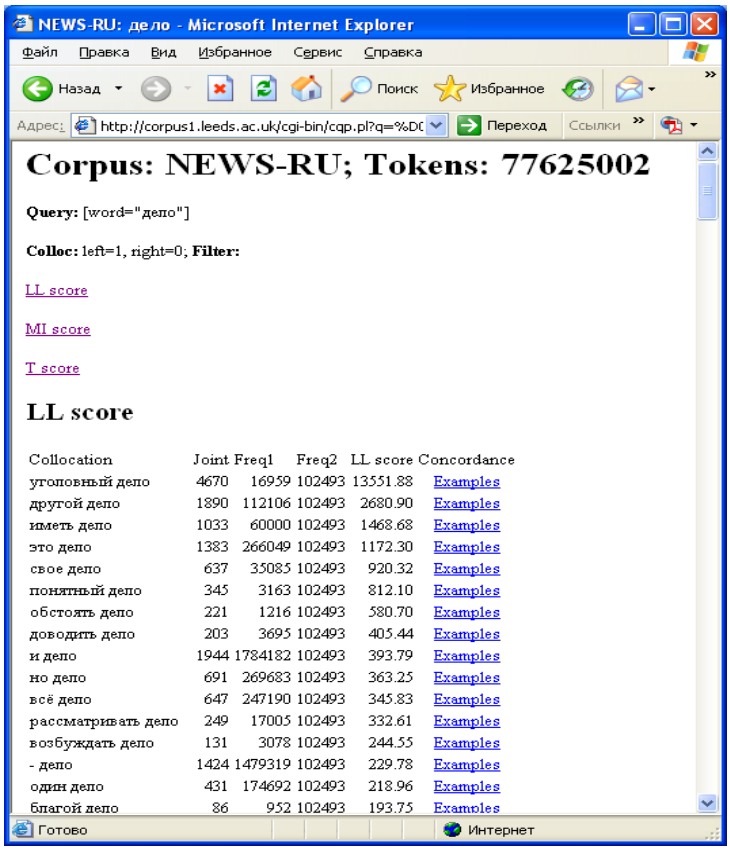
2 http://corpus.leeds.ac.uk/ruscorpora.html

3 http://corpora.narod.ru/mystem

For each word we examined bigrams as an example of collocations, i.e. combinations of a given word with a word which is on its right or on its left. Thus, for each noun the following information was given out: 1) its left bigrams; 2) its right bigrams.

It is necessary to mention beforehand two moments. First, each element of the corpus which stands before or after a blank including punctuation marks is considered a token. Secondly, the corpus manager CQP uses lemmas while processing data, thus, results of search are presented by combinations of lemmas.

The result of the query is represented by a list of collocations organized in the form of one, two or three tables (depending on the quantity of the chosen measures) with six data columns (see Fig. 1):



Collocation	Joint	Freq1	Freq2	LL score	Concordance
уголовный дело	4670	16959	102493	13551.88	Examples
другой дело	1890	112106	102493	2680.90	Examples
иметь дело	1033	60000	102493	1468.68	Examples
это дело	1383	266049	102493	1172.30	Examples
свое дело	637	35085	102493	920.32	Examples
понятный дело	345	3163	102493	812.10	Examples
обстоять дело	221	1216	102493	580.70	Examples
доводить дело	203	3695	102493	405.44	Examples
и дело	1944	1784182	102493	393.79	Examples
но дело	691	269683	102493	363.25	Examples
всё дело	647	247190	102493	345.83	Examples
рассматривать дело	249	17005	102493	332.61	Examples
возбуждать дело	131	3078	102493	244.55	Examples
- дело	1424	1479319	102493	229.78	Examples
один дело	431	174692	102493	218.96	Examples
благой дело	86	952	102493	193.75	Examples

Fig. 1. Example of the output of the query on the word дело “business”

The first column shows the collocation (represented by lemmas) itself. The joint frequency of occurrence of bigram's components, the frequency of the first word and the frequency of the second word stand in the second, third and fourth columns accordingly. The data in all tables are sorted on decrease of value of a corresponding measure. The query results for each noun are brought to one table, and then we compared them to the entries for these nouns in the Dictionary of Collocations [11], in the explanatory dictionaries of Russian (the Dictionary of Modern Russian [12]; the Big Academy Dictionary of Russian [13], the Dictionary of Russian [14]) and in the Dictionary of Synonyms and Similar Expressions [15].

3.1 Results for Log-Likelihood

For LL measure the following results were received. 1763 bigrams were found in total. Among them there were:

47 bigrams are fixed in two or more dictionaries;

79 bigrams are fixed only in [11];

48 bigrams are fixed only in [14];

20 bigrams are fixed only in [15];

11 bigrams are fixed in [13];

6 bigrams are fixed only in [12].

Also there were 15 combinations with punctuation marks.

Values of LL proved to be the largest for the collocations found in two or more dictionaries.

3.2 Results for MI

1755 bigrams were found in total. Among them there were:

68 bigrams fixed in two or more dictionaries;

73 bigrams fixed only in [11];

27 bigrams are fixed only in [14];

13 bigrams are fixed only in [15];

9 bigrams are fixed in [13];

25 bigrams are fixed only in [12].

Also there were 11 combinations with punctuation marks.

Bigrams, extracted by MI and t-score also correlate with data of dictionaries.

Values of the MI measure are the largest for the collocations found only in [14], and also found in two or more dictionaries. After examination of the list of results we found out, that only two combinations were retrieved (and both

were not fixed in the dictionary of collocations) within a range from 0 to 1 (according to the value of MI). It allows us making a conclusion that the combination is statistically insignificant if the MI appears in the given interval. Thus the hypothesis that was applied to other languages can be extrapolated to Russian.

3.3 Results for t-score

1755 bigrams were found in total. Among them there were:

71 bigrams fixed in two or more dictionaries;

73 bigrams fixed only in [11];

22 bigrams are fixed only in [14];

14 bigrams are fixed only in [15];

8 bigrams are fixed in [13];

23 bigrams are fixed only in [12].

Also there were 20 combinations with punctuation marks.

The combinations that have large values of t-score prove to be rather frequent while, unlike the previous measures, one of their parts is a preposition or a pronoun. And also there were more bigrams (in comparison with other measures) in which a punctuation mark is one of their parts.

We confirmed the hypothesis that t-score allows to retrieve collocations which have very frequent words, and also punctuation marks as their constituents. Thus, as well as for other languages, it is true for Russian that words with the largest value of t-score are frequent and can be combined with a large number of words. The right context reveals more combinations with punctuation marks than the left one.

The analysis of the data received shows that the majority of collocations (phrasemes), fixed in dictionaries, stand in the top part of the list, i.e. their parts co-occur very often.

The combinations which had not been fixed in the dictionaries before were also retrieved during the experiment. The analysis of these combinations that show both high and low values of measures of association (one or several), reveals, that bigrams which stand on the top of the list of collocations (sorted on decrease), with some degree of probability prove to be set phrases and, hence, can be included in the dictionary. The overwhelming majority of collocations that stand in the bottom part of the list prove to be free phrases.

Also it is possible to note the combinations recognized by us as collocations, but not listed in dictionaries. In case of large value of a measure for such combinations one can say to a certain degree that they belong to a class of set

phrases: for example, *центр внимания* "the focus of attention", *укромное место* "secluded corner", *покончить жизнь* "to commit suicide", *драконовский закон* "draconian law", *щекотливый вопрос* "ticklish question" etc.

4 Conclusion and further work

The results of this work (and the data about word collocability in general based on statistical measures), first of all, can be applied to a lexicographic practice. The statistical collocations which are extracted by measures of association, and not fixed in a dictionary, can be added to the existing dictionaries after careful analysis. Application of corpus methods to the analysis of lexical collocability will allow to create, finally, the dictionary of a new type, namely an integrated dictionary of set phrases, or the dictionary of collocations.

It is obvious, that the automatic text analysis (for example, by means of the above described statistical tools) is only an initial stage for retrieving collocations. Then the received results must be manually processed within the framework of traditional linguistics and compared to the data from dictionaries (first of all, explanatory dictionaries and dictionaries of set phrases). One should take into account also structural formulas which underlie collocations. Combined with statistical approaches, in our opinion, they can give quite good results. Programs which allow for stop-words and punctuation marks must also be used. It is syntactic tree banks that may solve the task in question. It is possible to combine statistical tools with structural (syntactic) models of phrasemes and collocations, thus, uniting two approaches.

Acknowledgements

I am grateful to my supervisor Victor Zakharov (Saint-Petersburg State University) for his support, and encouragement in this work. Also I'd like to express my gratitude to Irina Azarova for inspiring discussions on this topic.

References

1. Meľchuk I.A. O terminakh "ustojchivost" i "idiomatichnost". Voprosy jazykoznanija. M., 1960, № 4. P. 73–80.
2. Akhmanova O.S. Slovar' lingvisticheskikh terminov. M., 1966.
3. Deribas V.M. Ustojchivye glagoľno-imennye slovosochetanija russkogo jazyka. M., 1983.
4. Teliya V.N. Russkaja frazeologija: semanticheskij, pragmaticheskij i lingvokuľtorologicheskij aspekty. M., 1996.
5. Firth J.R. Papers in Linguistics 1934–1951. London, 1957.
6. Borisova E.G. Kollokatsii. Chto eto takoe i kak ikh izuchat'. M., 1995.

7. Fillmore C.J. 1968 The case for case. E. Bach, R.T. Harms eds. Universals in linguistic theory. – L. etc.: Holt, Rinehart and Winston, 1968. 1-88.
8. Iordanskaja L.N., Meľchuk I.A. Smysl i sochetaemost' v slovare. M., 2007.
9. Stubbs M. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. Functions of Language, 1, 1995. P. 23–55.
10. Sharoff S. Chastotnyj slovar' slovar' russkogo jazyka. 2002.
URL: `http://www.artint.ru/projects/frqlist.asp`
11. Borisova E.G. Slovo v texte. Slovar' kollokatsij (ustojchivykh slovosochetanij) russkogo jazyka s anglo-russkim slovarem kljuchevykh slov. M., 1995.
12. Slovar' sovremennogo russkogo literaturnogo iazyka: vol, 1-17. Moscow, Russia (1948-1965) (BAS-17).
13. Boľshoi akademicheskii slovar' russkogo iazyka: vol. 1-6. Saint-Petersburg, Russia (2004-2007) (to be continued) (BAS-25).
14. Slovar' russkogo iazyka: vol. 1-4. Moscow, Russia (1957-1961) (MAS).
15. Abramov N. Slovar' russkikh sinonimov i skhodnykh po smyslu vyrazhenij. M., 2006.

The Role of Word Frequency Vocabularies in the Research of Psychology and Philosophy Terminological Systems

Oksana S. Kozak

Kiev, Ukraine

Abstract. The paper covers on the ways to use word frequency vocabularies for the data analysis of Psychology and Philosophy terminological systems which have been specified upon and carried out through pointing out, defining and statistically studying the conceptual sets on the language material of scientific papers which were published in the 19th and 21st centuries.

Keywords: *sublanguage, functional semantics and style category (FSSC), lexical and grammatical paradigm, conceptual set, set unit, centre (core), the periphery, conceptual diagram.*

Language structure as well as its functioning in speech, the relation of language and ideation, language and society are subordinated to statistical laws, thus the methods of statistical analysis should be referred to when studying the above mentioned phenomena [1, 7] as these methods are particularly efficient when carrying out a comparative study/analysis. As suggested by F. P. Filin, every word reveals the story of the whole world [2, 226]. In the case of the given research, the combination "the whole world" is equipollent to the term "conceptual sphere", and "the story it reveals" stands for the concept transformations. The latter are reflected in the changes of lexical and grammatical text paradigm. In other words, the concept transformations correlate with the dynamics of functional semantics and style categories (FSSC) of the corresponding sublanguages. This dynamics is characterized by the system of markers, the markers in this research being the most frequent conceptual sets.

The study of the sublanguages evolution is topical nowadays. Not only it suggests various ways of complying, modifying and updating terminological (but not limited to these) vocabularies for various scientific fields, but also can serve the basis for determining general tendencies and trends in the development of certain science branches. Terminology is one of the few spheres of lexis units of which belong to almost every functional semantics and style category of Psychology and Philosophy sublanguages, terms of various kinds being the constituents of the category centre (core).

There is no denying the fact that each of the two sublanguages operates by the set of lexical and grammatical units proper to it. This set is to be replenished and perfected as "at this very stage of the English language development there is a strong tendency for language economy and the breach in

the 'one symbol – one meaning' language automatism" [3, 27]. Moreover, the absence of a trenchant division between terms and common use lexis is obvious resulting from the constant process of terms transition into common use vocabulary accompanied by the use of trivial vocabulary for forming new branch terms [4; 12].

I aim to find out the possibility of using word frequency vocabularies of articles on Philosophy and Psychology when studying Psychology and Philosophy terminological systems, with the idea of building conceptual diagrams of terminological systems for the corresponding sublanguages at the next stage of my research. The peculiarity of this research lies in the attempt to frame out the comparison for the articles representing two different periods.

It is necessary that the definition for the 'conceptual set' term used in this study should be given. 'Conceptual set' stands for the complex of lexical and grammatical units that serve as the embodiment for one and the same concept in the texture of text.

The set of terms for a certain sublanguage is constantly replenished and perfected which causes the necessity to terminally involve new language resources to name to the new objects, phenomena or processes discovered. Thus, the task of my research is to define the most frequent 'conceptual sets' and to compare the terminological vocabularies of the Philosophy and Psychology sublanguages for the two periods.

The 'conceptual set' that has been defined and studied within this research, fits into the plane of lexical and grammatical abstraction and correlates with the concept field through the paradigm of lexical and grammatical units. In their form the elements of the 'conceptual set' are common root lexemes which stand for one and the same concept and represent the scope of declension range for the notional parts of speech (noun, adjective, pronoun, verb and its non-finite forms).

The total number of word usage in the papers under analysis made up 49 280 cases. Scientific papers on and Philosophy and Psychology which were published in the 2nd part of the 19th century and the beginning of the 21st century served the material for this research. Every of the two periods is represented by four articles each bearing thematic and size approximation to one from the other period. The target material for the vocabularies is the selection of the most frequently used nouns, verbs and adjectives. Each terminological concept includes a core unit and periphery units.

Thematically related texts for the two periods have been analysed. The analysis resulted in defining the most frequent 'conceptual sets' for the corpora of Philosophy and Psychology papers compared. The total number of sets may indicate the theme breadth of the articles under study. For the first period (the 2nd part of the 19th century) the total set number in the Philosophy corpus is

167, for the second period (the beginning of the 21st century) – 184 sets. The total set number in the Psychology corpus is 182 and 201 sets respectively. Relative difference in the set number for Philosophy is 9.2 % and 9.5 % is the figure for the Psychology sublanguage.

Absolute and relative set frequency for usage cases has been calculated, which was then followed by the analysis of set frequencies for the two periods according to:

a) **frequency dispersion** (the presence of the set units in the texts of one of the periods under research; the presence of the set units in the texts of the two periods with significant/nonsignificant frequency difference);

b) **frequency range** (diapason):

- low (0.01 – 0.09%);
- lower than the medium (0.10 – 0.19%);
- medium (0.20 – 0.29%);
- higher than the medium (0.30 – 0.39%);
- high (0.40% and more);

c) **frequency homogeneity** (statistically homogeneous/ nonhomogeneous frequencies);

d) **set composition** (the number of units, set filling/lexical and grammatical unit status).

Frequency dispersion analysis has provided the possibility to find out that the presence of the set units in the texts of one of the periods under research and their absence in the texts of the other period is the indicator of increase/ decrease in evincing interest for a certain philosophical/psychological issue as well as of the invention of new terminological units (i.e. units «affective», «allele», «biochemical», «depression», «genes», «interpretation», «rivalry» belong to the set core but that is proper to the papers of the second period only)/ the transition of some terms into common use vocabulary; it can also be the indicator of low texts homogeneity for a certain set.

Frequency range (diapason) analysis enabled not just arriving at the conclusion concerning the core/periphery scheme of the conceptual set and outlining the framework of the conceptual sets for the two periods, but also prompted the idea of building conceptual diagrams of terminological systems for the corresponding sublanguages[8] in accordance with the core/periphery scheme.

Frequency homogeneity analysis made it possible to estimate frequency homogeneity/absence of homogeneity for the units of every conceptual set concerning the two periods.

Set composition analysis (the number of units, set filling) enabled to retrace the dynamic changes of lexical and grammatical paradigm for a certain conceptual set (the set centre shift from one unit form/invariant onto another;

increase/decrease in the relevant frequency of conceptual sets; the change in lexical and grammatical unit status.

Here below the full scheme analysis for the 'individual' conceptual set is given.

This is the table of comparison for the units frequency of the "individual" conceptual set.

№ п/п	Одиниця	К-сть слововживань (n=25)
1	**individual**	17
2	individualism	1
3	individuality	2
4	individuals	4
5	individual's	1

Table 1. Correlation between the units quantity for the 'Individual' Conceptual Set. Period 1. m=0.36%

№ п/п	Одиниця	К-сть слововживань (n=43)
1	individual	3
2	**individuals**	27
3	individual's	1
4	individuals'	12

Table 2. Correlation between the units quantity for the 'Individual' Conceptual Set. Period 2. m=0.5%

PS 1 and PS 2 stand for the first and the second periods of papers on Psychology, m and n – relative and absolute frequencies respectively for a certain set, the unit in bold stands for the core of a conceptual set of a certain period.

According to the frequency dispersion this conceptual set is a prevailing one with a nonsignificant frequency difference. According to the frequency range (diapason) the set belongs to different groups for the two different periods: higher than the medium frequency (0.30 – 0.39%) group (for the first period); high frequency (0.40% and more) group. Thus according to the frequency homogeneity 'individual' is a set with statistically nonhomogeneous frequencies. According to the set composition this set is characterized by a significant difference in units number for the two periods (the relative difference making up 0.14%) which has resulted in set filling/lexical and grammatical unit status difference. As can be assumed from *Table 1* there is a core shift in the set

(individual 17 → individuals 27), the number of units for the two periods is almost about the same, though there is a great difference in the set units representation. I have arrived at the conclusion that the 'individual' conceptual set has a significantly different weight in the papers of the two periods. From the set data analysis I can assume that while in the the 2nd part of the 19th century the impact a considerable accent was put on the individual's inner world, at the beginning of the 21st century the interest has been shifted onto the way the individual is/is not able to integrate with the surroundings.

The results of the research prove the significance of referring to word frequency vocabularies while conceptually diagramming terminological systems of philosophy and psychology sublanguages with the aim of their comparison.

References

1. *Перебийніс В. І.* Статистичні методи для лінгвістів: Навчальний посібник. – Вінниця: Нова книга, 2001. – 168 с.
2. *Филин Ф. П.* Очерки по теории языкознания. – М, 1982. – 358 с.
3. *Ивина Л. В.* Лингво-когнитивные основы анализа отраслевых терминосистем (на примере англоязычной терминологии венчурного финансирования): Учебно-методическое пособие. – М.: Академический Проект, 2003. – 304 с.
4. *Головин Б. Н., Кобрин Р. Ю.* Лингвистические основы учения о терминах.– М.: Высшая школа, 1987. – 104 с.
5. *Козак О. С.* Мовленнєва системність та стилістичні категорії підмов філософії та психології // Проблеми семантики, прагматики та когнітивної лінгвістики: Зб. наук. пр. Вип. 10 / Київ. нац. ун-т ім. Т. Шевченка; Відп. ред. Н. М. Корбозерова. – К., 2006. – С. 139-143.

Variation of Czech Lexicon as Reflected by Corpora Comparison

Michal Křen

Institute of the Czech National Corpus
Charles University, Prague
michal.kren@ff.cuni.cz

Abstract. SYN2000 and SYN2005 are both 100-million representatively balanced corpora of contemporary written Czech that cover two consecutive time periods. It is therefore desirable to take advantage of this and to compare word frequencies in both corpora in order to discover lexicon development tendencies. However, both corpora differ in many other aspects that make direct comparison questionable. The paper describes research based on normalised corpus frequencies devised in order to enable the lexical comparison. Statistical significance measures are used for evaluation of frequency differences of the individual items. It is shown that careful interpretation of the observed results is necessary, because the differences can have various causes including corpus composition issues and non-random nature of language. True examples of lexical variation are found to be rare and hardly distinguishable.

1 Achieving comparability

The Czech National Corpus is an ongoing wide-scale project aiming to provide the research community with large variety of Czech corpora (Čermák 1997 and 1998). Perhaps the most widely used are 100-million monolingual synchronic written corpora SYN2000 and SYN2005. They are disjunctive, i.e. none of the texts was included into both of them. Both corpora are representatively balanced and cover two consecutive time periods: while SYN2000 contains texts from the 1990s, SYN2005 concentrates on texts from the first half of the 2000s. Despite their similar concept, the corpora differ in many other aspects that may not be apparent. Major difference is the notion of representativeness: SYN2000 contains 15% of fiction, 25% of professional literature and 60% of newspapers and magazines, while SYN2005 contains 40% of fiction, 27% of professional literature and 33% of newspapers and magazines. The proportions were in both cases based on sociological research that emphasised text reception (reading) rather than production (writing). The difference between their results can be explained by a turn-away from newspapers at the end of the 1990s as well as by different research methodology. Other differences between the two corpora include mainly various processing issues: improved tokenization (dividing the texts into sequence of tokens), segmentation (sentence

boundary recognition) and mainly lemmatisation with morphological tagging (Hajič 2004 and Spoustová 2007). The differences between the two corpora should thus be considered as an inevitable improvement, although they make lexical frequencies directly incomparable: even significant frequency difference may not reflect any lexicon development at all.

The processing differences were overcome simply and effectively by reprocessing SYN2000 with the same set of tools that were used for processing SYN2005. This has been done only internally, because the corpora are claimed to be reference entities, i.e. they are never altered once published. However, retokenization of the corpus changed its size to 96.23 million tokens while the size of SYN2005 remained 100 million tokens, thus this newly emerged difference had to be taken into consideration. The influence of the modified sampling criteria was minimised by compiling comparative frequency lists (CFLs) for both corpora in two versions: for word forms and lemmas. Among other data, the CFLs provide for every item (word form or lemma) overall normalised frequency in a 100-million comparative corpus. The comparative corpus is a virtual construct, normalised counterpart of the respective real corpus, where all three main registers are equally represented, i.e. with one third share each. The average frequency of every item in each of the shares is normalised to be the same as in the corresponding register of the respective real corpus. As a consequence, the normalised frequencies are directly comparable between the corpora. Moreover, they are regular frequencies, although in a virtually non-existing comparative corpus. It is thus possible to handle them the same way as regular corpus frequencies, e.g. their total sum over every item in the corpus is 100 million. The CFLs are publicly available on our web pages together with usage examples and other practical notes, their detailed description can be found also in (Křen 2006).

2 Aims and pitfalls of the comparison

Perhaps it should be stressed beforehand that the paper does not aim to compare both corpora as a whole and to quantify the difference between them. The principal aim is to find out whether there are significant differences in usage of individual Czech word forms or lemmas reflected in the two corpora that could be discovered by the means of the CFLs. The paper can also be considered as an attempt to explore both the advantages and limitations of the CFLs as a publicly available resource, so that it would be possible for anybody interested to take up and perhaps extend this evaluation.

According to Rayson and Garside (2000) there are number of issues that should be considered when comparing corpora or frequency lists based on them: the representativeness of the corpora, their homogeneity and applicability of used statistical tests (e.g. their suitability for corpora of different size). They stress that word frequencies tend to differ across any two texts just because of the non-random nature of language. Moreover, the differences do not balance

out as the texts (or corpora) grow larger, so that there always are rather small but statistically significant differences in frequencies of high frequency words. This is also the main reason why they are overestimated by some measures, although this is statistically well-grounded result. We should be aware that the measures in general are likely to highlight differences in word frequencies that are salient statistically, although this salience itself need not show any difference in usage. However, this is often not recognisable merely from frequency data the measures are based on. Statistical measures should be therefore used as a useful prerequisite for ranking the candidates that should be finally examined by the researcher. This is also the approach adopted here: statistical measures described below were used for wordlists ranking, the individual items at the top of the lists were then inspected manually, examining the results and their possible causes at the same time.

3 Statistical measures

Kilgarriff (1996) surveys different statistical approaches used in order to find words that are characteristic for particular text. This problem can be also viewed as finding the most significant differences in word frequency between the text and large representative corpus. He summarises various approaches, evaluates them and indicates circumstances in which they are applicable. However, this paper focuses on evaluation of individual lexical items rather than the statistical measures. Kilgarriff's survey was therefore used as a source of suitable techniques to choose from, not even attempting to test those that were disapproved (e.g. MI-score). Moreover, it should be mentioned that measures requiring any additional information to the CFLs were not even considered, although more information might have improved the performance. For instance, this is the case of the Mann-Whitney ranks test mentioned by Kilgarriff as a suitable option. For each word, he used it on ranks of word frequencies in 2000 same-sized samples from both compared corpora. Because this information cannot be inferred from the lists, the test was not used in this comparison.

Pearson's χ^2 test is one of the statistical measures most frequently used in similar cases. It was used also by Johansson and Hofland (1982) in order to find significant differences between British and American English by comparing wordlists generated from two comparable corpora, LOB and Brown. However, most of the frequent function words were marked as having significantly different frequency. Kilgarriff (1996) shows that similar result can be obtained also when comparing corpora of the same language type. Therefore, the differences found by LOB-Brown comparison cannot be interpreted as differences between British and American English. Oakes (1998) explains that for all but purely random populations χ^2 tends to increase with frequency. Because

language is not random, frequency differences for frequent words are almost always significant. In other words, χ^2 answers question whether the two corpora were drawn randomly from larger population (the null hypothesis), but we already know that they were not. Thus the answer provided by χ^2 in fact is whether there is enough evidence to claim already known fact on given significance level.

Let a be normalised frequency of given item w in SYN2000 and b be its normalised frequency in SYN2005. The size of the comparative corpus is 100,000,000 for both corpora. Observed frequencies are then given by the following contingency table:

	SYN2000	**SYN2005**
w	a	b
non w	100,000,000 – a	100,000,000 – b

Table 1.

Expected frequencies are as follows:

	SYN2000	**SYN2005**
w	$(a + b) / 2$	$(a + b) / 2$
non w	100,000,000 – $(a + b)/2$	100,000,000 – $(a + b)/2$

Table 2.

χ^2 is then computed in a standard way as a sum over each cell of the table. Since we are using 2×2 contingency table, Yates's correction is applied:

$$\chi^2 = \sum \frac{(O - E - 0.5)^2}{E}$$

There can be various ways how to overcome some of the objections against χ^2, one of them is proposed by Kilgarriff and Salkie (1996). Their CBDF (chi by degrees of freedom) measure improves plain χ^2 by dividing its value by number of degrees of freedom for wordlist-based corpora comparison. In their case, the number of degrees of freedom is equal to the wordlist size minus one. However, our goal is to compare word frequencies individually in order to find out whether the observed difference is or is not significant, the number of degrees of freedom being always 1 in this case.

To sum up the major limitations of χ^2: it overestimates high frequency items, it should not be used for comparison of different sized corpora, and it should not be used in cases when expected frequencies are less than 5. However, using normalised frequencies provided by the CFLs ensures that the expected frequencies are always at least 10 (less frequent items are not included in the lists) and both corpora are normalised to the same size, so the only problematic feature of χ^2 is its overestimation of high frequency items. Kilgarriff and Rose

(1998) show that χ^2 performed the best out of five measures they evaluated, thus it can be very useful despite its limitations. Therefore, it was decided to adapt χ^2 to the task of finding significant differences in usage of individual words between two corpora based on the CFLs. The measure is called CBF (chi by frequency) and its value is given by dividing the χ^2 value by the square root of the expected frequency. The square root was determined empirically as a suitable curve between linear relation on one hand and logarithmic one on the other. Dividing directly by the value of expected frequency showed to disadvantage frequent words too much, while logarithm of the expected frequency proved to be too low to improve the original χ^2 value noticeably. CBF is given by the following formula:

$$CBF = \frac{\chi^2}{\sqrt{(a+b)}}$$

Another obvious option is to use different statistics instead of attempting to improve χ^2. Dunning (1993) proposes log-likelihood (hereafter LL; also known as G^2), an asymptotic hypothesis test similar to χ^2. He shows that it is more appropriate than χ^2 if observed frequency is rather small and sample size relatively large. Since this is the case of our lexical comparison, it was decided to evaluate also LL in addition to χ^2 and CBF. LL is given by the following formula:

$$LL = 2 \sum O \ln \frac{O}{E}$$

Two remarks should be added at this point. First, we are aware of the fact that CBF is merely an ad hoc solution devised for this task and not showing any statistical significance. However, the goal is "only" to rank the differences, so this should not be considered a drawback of the measure. Second, there is a serious theoretical problem related to the different nature of differences between low vs. high frequency words. For instance, let us consider the following table that shows three lemmas together with their normalised frequencies for both corpora:

lemma	**SYN2000**	**SYN2005**
esemeska (*SMS message*)	0	217
euro (the currency)	1128	9530
kraj (*county*)	8920	24434

Table 3.

It is not clear which of the frequency differences we should rank as the most significant if we did not know the language? What is the desired result of such comparison? How do we weigh the significance? We are convinced that there

can be no universally accepted ideal measure because the answers will surely be task-dependent and often individual. This should limit our expectations concerning the results the measures are able to provide us with. It also emphasises the importance of human intuition and common sense that should guide their interpretation.

4 Evaluation

This chapter evaluates the results provided by LL, χ^2 and CBF measures, the top 20 items for each of them being shown in the tables below. For the sake of clarity, the tables do not give exact values of these measures, but instead only rank assigned by them to the individual items. All word frequencies used in the evaluation are normalised frequencies described in the first chapter. Proper names were not taken into consideration because their frequencies largely depend on selection of particular texts and they are not related to language usage in general. Finally, all the items mentioned from now on are lemmas. Although the evaluation was carried out for word forms as well, the results were very similar and therefore it was decided to demonstrate them on lemmas only.

lemma	**SYN2000 frequency**	**SYN2005 frequency**	**LL rank**	**χ^2 rank**	**CBF rank**
euro (the currency)	1128	9530	1	2	1
kraj (*county*)	8920	24434	2	1	11
zvěř (*wildlife animals*)	9941	1989	3	3	3
pan (*Mr.*)	78821	52879	4	4	320
b (abbreviation)	29461	14574	5	5	59
fax (*fax*)	6492	1075	6	6	5
prag (geologic period)	3226	62	7	15	2
internetový (*internet* – adj.)	1449	6882	8	12	12
on (*he*)	1031606	1122317	9	7	29006
cm (centimeter)	17533	7863	10	9	76
plyn (*gas*)	18932	8961	11	11	93
strana (*side* or *party*)	138041	108141	12	10	2313
se (reflexive -*self*)	2658917	2799660	13	8	33700
m (abbreviation)	27041	15090	14	13	182
foto (*photo*)	7990	2249	15	14	23
šaman (*medicine man*)	341	3767	16	20	6
myslivecký (*huntsman* – adj.)	4100	475	17	18	8
logistický (*logistic*)	356	3745	18	21	7
cz (part of internet address)	1343	6036	19	16	19
xxxx (corpus cleanup failure)	2404	12	20	28	4

Table 4. The most significant frequency differences according to LL.

lemma	SYN2000 frequency	SYN2005 frequency	LL rank	χ^2 rank	CBF rank
kraj (*county*)	8920	24434	2	1	11
euro (the currency)	1128	9530	1	2	1
zvěř (*wildlife animals*)	9941	1989	3	3	3
pan (*Mr.*)	78821	52879	4	4	320
b (abbreviation)	29461	14574	5	5	59
fax (*fax*)	6492	1075	6	6	5
on (*he*)	1031606	1122317	9	7	29006
se (reflexive -*self*)	2658917	2799660	13	8	33700
cm (centimeter)	17533	7863	10	9	76
strana (*side* or *party*)	138041	108141	12	10	2313
plyn (*gas*)	18932	8961	11	11	93
internetový (*internet* – adj.)	1449	6882	8	12	12
m (abbreviation)	27041	15090	14	13	182
foto (*photo*)	7990	2249	15	14	23
prag (geologic period)	3226	62	7	15	2
cz (part of internet address)	1343	6036	19	16	19
krajský (*county* – adj.)	2986	8846	21	17	42
myslivecký (*huntsman* – adj.)	4100	475	17	18	8
můj (*my*)	239701	204127	22	19	12540
šaman (*medicine man*)	341	3767	16	20	6

Table 5. The most significant frequency differences according to χ^2.

lemma	SYN2000 frequency	SYN2005 frequency	LL rank	χ^2 rank	CBF rank
euro (the currency)	1128	9530	1	2	1
prag (geologic period)	3226	62	7	15	2
zvěř (*wildlife animals*)	9941	1989	3	3	3
xxxx (corpus cleanup failure)	2404	12	20	28	4
fax (*fax*)	6492	1075	6	6	5
šaman (*medicine man*)	341	3767	16	20	6
logistický (*logistic*)	356	3745	18	21	7
myslivecký (*huntsman* – adj.)	4100	475	17	18	8
honitba (*hunting ground*)	2902	205	25	29	9
myslivost (*hunting* – subst.)	2860	220	27	32	10

kraj (*county*)	8920	24434	2	1	11
internetový (*internet* – adj.)	1449	6882	8	12	12
plynovod (*gas pipeline*)	2990	290	29	34	13
bosenský (*bosnian*)	3154	360	31	35	14
hořák (*burner*)	3257	387	30	33	15
česko (*Czechia*)	650	4152	23	25	16
hn (newspaper abbreviation)	28	1426	57	85	17
souvrství (*strata*)	1831	103	49	59	18
cz (part of internet address)	1343	6036	19	16	19
podzol (*podsole*)	1520	55	59	83	20

Table 6. The most significant frequency differences according to CBF.

Generally, LL and χ^2 give ranking very similar to each other, while CBF is more remarkably distinct. The basic feature of CBF is that it prefers low frequency items with greater frequency differences between the corpora to high frequency items with smaller differences. The latter are statistically more significant and therefore preferred by the other two measures. However, since CBF is based on χ^2 it cannot be expected to find anything really new and surprising that χ^2 itself would not rank noticeably high anyway, although CBF ranks such items higher in accordance with human intuition. This can be demonstrated on the last three columns of Table 3: none of the ranks is greater than 100, which is certainly not true for Table 1 and Table 2. In other words, given the χ^2-ranked list, CBF tends to push highly frequent items down significantly rather than to pull low frequency items up accordingly. For instance, frequencies of lemmas *xxxx* or *hn* are quite high in both corpora, their ratio being ca. 200 and 50 respectively (cf. Table 3). Although the occurrences of both lemmas are concentrated in a few texts in both cases, any conclusion based only on the frequency information should no doubt conclude that the observed difference is significant. CBF ranks both lemmas slightly higher than the other two measures. On the contrary, frequencies of lemmas *se* and *on* are both extremely high, but the observed difference is rather small and can be explained by the non-random nature of language. CBF ranks them significantly lower which can be considered more appropriate. Therefore, CBF can be more helpful for automatic ranking rather than for preprocessing followed by manual examination of the results. χ^2 and LL are more useful in this respect, because the minor differences of highly frequent items can be easily left out if found unimportant. On the other hand, they may reflect gradual usage change and this can be analyzed only by means of professional analysis. However, CBF can be seen as a suitable base for further improvement of the properties of χ^2. Frequent general language words are already suppressed and additional distribution information would presumably improve its performance even more, since most of the words it prefers could be easily detected as domain-specific. As for the comparison between χ^2 and LL, LL seems to perform slightly better, as it

does not overestimate highly frequent items that much is χ^2 and is thus closer to CBF in this respect.

The words provided by the tables can be roughly divided into four main groups. Probably the most remarkable is the number of domain-specific words that can be found especially in Table 3 and the fact that they come from only a few domains, namely hunting (*zvěř*, *myslivecký*, *honitba*), gas industry (*plyn*, *plynovod*, *hořák*) and geology (*prag*, *souvrství*, *podzol*). Their common feature is that the frequency differences heavily depend on corpus composition and therefore could have been avoided by employing some kind of additional distribution information. The second group consists of period-specific words that are topics of public discourse or reflect technical development: *euro*, *kraj*, *internetový*, *bosenský*, *česko* etc. Although these words may be of interest, they do not represent the core of language development. Unlike the domain-specific words, their distribution is more even and does not depend on selection of particular texts. The third group contains frequent general language words: *on*, *se*, *můj*, *strana* etc. The frequency differences are rather small but statistically significant and are caused mostly by non-randomness of language, although corpus composition issues are of minor importance, too. Finally, the fourth group consists of various errors or generally trash that got pinpointed by the comparison. These include abbreviations (*b*, *m*, *cm*), results of insufficient corpus cleanup (*xxxx*, *hn*, *foto*) or incorrect lemmatisation (*prag* and *česko* should have been lemmatised as proper names). Perhaps the most notable error encountered here concerns lemma *podzol*: by far the most prevailing word form lemmatized as *podzol* is *PZ*, an abbreviation of "podzimní zkoušky" – autumn tests of hunting dogs. The above mentioned classification should thus be slightly adjusted. However, it is more important to mention that the groups can overlap (e.g. *logistický* is both domain-specific and period-specific term) and that it is not always clear what are the real reasons of observed differences. For instance, to what extent the frequency difference of lemma *fax* was caused by corpus composition issues or different level of corpus cleanup, as it often occurs as a part of the address in footnotes (its nature is similar to *foto* in a sense).

Taking these findings into consideration, it should not be surprising that even careful examination of lower-ranking items in the lists does not easily give true examples of lexical variation. For instance, *esemeska* (*SMS message*) is a perfect example of neologism brought into common use by technical development. It does not occur at all in SYN2000, but its normalised frequency in SYN2005 is 217. However, this frequency difference itself cannot be viewed as an evidence of its novelty. There are number of lemmas with frequency characteristics similar to the 0 : 220 ratio, e.g. *humerus* (*humerus* – shoulder bone, professional medical term), *spagyrikum* (*spagyrics* – herbal alchemy agent), *kanovnice* (*canon* – clergywoman), *lůmek* (*small quarry*) etc. Obviously, this is caused by the lack of some kind of distribution information, because all

these words are domain-specific and more or less unevenly distributed. However, the evenness of distribution of *esemeska* is not notably even as well: all its occurrences in SYN2005 are from the newspaper part of the corpus (except for one occurrence from fiction). Moreover, one more lemma can be added to these above: *beďar* (*furuncle* – colloquial). Although it is a widely used informal expression with completely different nature from linguistic point of view, its frequency distribution is uneven as well and can be statistically hardly distinguished from the domain-specific words above. Its absence in SYN2000 can thus be seen as a shortcoming of the corpus data without any linguistic explanation.

Let us take more frequent function word as another example. Lemma *-li* (morpheme meaning *if* attached mostly to verbs) is becoming archaic and it would thus be desirable to find evidence for this. The respective frequencies correspond to this expectation being 71901 for SYN2000 and 54572 for SYN2005. Although they are rather high so the difference may be significant, it is difficult to determine its exact cause. Both the frequency differences and distribution of occurrences are very similar to that of *strana* or other frequent general language words mentioned in the previous paragraphs and not showing any tendency to disappear from the language. It can thus be concluded that even the distribution information may not help to distinguish cases of lexicon development from mere corpus composition issues.

5 Conclusions and further work

It was shown that the measures do not give satisfactory results and their use is thus limited. There are several reasons for this failure. First, it is not possible to produce "ideal" ranking the measures should approach as close as possible. Second, normalised frequencies in the CFLs do not provide sufficient information. Frequency distribution, document frequency or any other kind of additional information would be of great help, especially in case of the domain-specific words, but they cannot be expected to solve all the encountered problems. Third, nature of the language data makes the task more difficult. Even statistically highly significant differences in frequency of very frequent function words are often caused by simple fact that corpora contain different texts, not by any language development tendencies. This points out to the importance of manual inspection and verification of the statistically-ranked items that can hardly be avoided.

We should be also aware of another limitation of this approach: the only language changes considered so far were dealing with introducing new words into usage or their disappearing from the lexicon. However, we should also take into account often neglected fact that the most common manifestations of language variation are related to semantic shift, polysemy or collocability preferences. Of course, this kind of lexicon development can be hardly traced statistically: collocability issues would require more data, while the semantic

features are virtually undetectable. Furthermore, it showed that frequency of neologisms almost never grows significantly enough within a span of few years to be detectable by statistical methods. It is therefore possible to discover only the most salient examples. The same holds also for function words that are usually much more frequent: their usage seems to change too slowly over time to make the frequency differences significant. Employing additional distribution information would no doubt reduce the number of false findings, but perhaps would not suffice, being overlapped by the non-randomness of language together with corpus composition issues that cannot be completely obliterated. However, utilising various kinds of distribution information in addition to the normalised frequencies is certainly desirable and further research should be therefore aimed in this direction.

Acknowledgement

This research has been supported by MSM0021620823 grant.

References

1. Czech National Corpus: SYN2000. Prague, 2000. Available on-line from http://ucnk.ff.cuni.cz
2. Czech National Corpus: SYN2005. Prague, 2005. Available on-line from http://ucnk.ff.cuni.cz
3. Czech National Corpus: Comparative frequency lists based on SYN2000 and SYN2005. Prague, 2006. Available on-line from http://ucnk.ff.cuni.cz/srovnani.html
4. Čermák F. (1997): Czech National Corpus: A Case in Many Contexts. In: *International Journal of Corpus Linguistics*. 2 (2), 181–197.
5. Čermák F. (1998): Czech National Corpus: Its Character, Goal and Background. In: *Text, Speech, Dialogue: Proceedings of the First Workshop on Text, Speech, Dialogue: TSD–98*. Brno: Masaryk University. 9–14.
6. Dunning T. (1993): Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational Linguistics*, 19 (1), 61–74.
7. Hajič J. (2004): *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha: Karolinum.
8. Hofland K., Johansson S. (1982): *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.
9. Kilgarriff A. (1996): Which words are particularly characteristic of a text? A survey of statistical approaches. In: *Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition*. Brighton: Sussex University. 33–40.

10. Kilgarriff A., Salkie R. (1996): Corpus similarity and homogeneity via word frequency. In: EURALEX '96 Proceedings. Göteborg: Göteborg University. 121–130.
11. Kilgarriff A., Rose T. (1998): Measures for Corpus Similarity and Homogeneity. In: *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP–3)*. 46–52.
12. Křen M. (2006): SYN2000 vs. SYN2005: Comparing the Large Synchronic Corpora of Czech. In: *Proceedings of the International Conference "Corpus Linguistics – 2006"*. St.-Petersburg: St.-Petersburg University Press. 182–189.
13. Oakes M. (1998): *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
14. Rayson P., Garside R. (2000): Comparing Corpora using Frequency Profiling. In: *Proceedings of the Workshop on Comparing Corpora, Annual Meeting of the ACL Archive*. 2000. no. 9, 1–6.
15. Spoustová D. (2007): *Kombinované statisticko-pravidlové metody značkování češtiny*. Praha: MFF UK.

Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora

Karel Kučera

Charles University, Czech National Corpus Institute
Náměstí Jana Palacha 2,
11638 Praha 1, Czech Republic
karel.kucera@ff.cuni.cz

Abstract. In lemmatizing a diachronic corpus, which includes texts from the entire history of a language, it is advantageous to adopt a broader notion of lemmatization to handle the diversity of morphological, phonological and orthographic forms. Working along these lines, the concept of hyperlemma has been implemented in the development of a lemmatizer for the diachronic part of the Czech National Corpus, now under way. Like the lemma, the hyperlemma represents all inflectional forms of a word, but unlike the lemma it also represents all historical and dialectal phonological varieties and modern spelling varieties. Moreover, the hyperlemma can be a set of lemmata in compounds like *naň, tys* or in forms shared by two coexisting paradigms.

At a rather general level one can say that the very basic reason for the lemmatization of a corpus is to facilitate the search for all the forms of a particular word. However, what exactly is considered a form of a word, often remains virtually undefined. I believe there is a general consensus that it is any inflectional form, i.e. any form resulting from conjugation of a verb (i.e. forms such as *go, goes, going, went, gone*) or declination of a noun, pronoun, adjective or numeral (such as *stone, stones, stone's* or *they, them*). On the other hand, how to handle different spelling forms (e.g. *theatre:theater*, *jail:gaol, night:nite* or *4 U* 'for you'), different sound forms, dialectal or individual pronunciations reflected in different spellings (*get:git, them:'em, that:dat* etc.) or even different but formally similar derivatives (e.g. *aluminium:aluminum*) often rests on arbitrary decisions. All these forms may be included under the same lemma, especially if the forms are not far removed from one another, or they may be represented by two different lemmata.

In the lemmatization of diachronic corpora, which include texts from the entire history of a language, i.e. texts from a number of its historical stages, the problem of lemmatization becomes much more complicated. Generally, one has to face, in addition to the above problems, the necessity to handle different historical forms, scribes' or printers' contractions and often also several different writing systems or competing orthographies. What may appear as a marginal rarity in a synchronic corpus, as well as in the contemporary language in general (e.g. Czech forms like *bylť, dejž*), is often a central, widespread

phenomenon both in older texts and historical corpora, and should be lemmatized in a systematic way.

To handle the diversity and competition of morphological, phonological, orthographic and other forms in lemmatizing diachronic corpora, it appears to be advantageous to adopt a rather broader notion of lemmatization. Working along these lines, the concept of *hyperlemma* has been implemented in the development of a lemmatizer for the diachronic part of the Czech National Corpus (DCNC), now under way. As far as I know, the label "hyperlemma" itself is new; however, a similar notion, limited largely to historical and dialectal variety of phonological and orthographic features, has been already used in the project of Tesoro della Lingua Italiana delle Origini (TLIO, [1], [2]). In the DCNC, which includes the language material from seven centuries of Czech written texts, the concept has been applied in a broader way. The basic characteristics of the hyperlemma can be described as follows:

1. Like the common synchronic lemma, the hyperlemma represents all inflectional forms of a word (in the case of old Czech texts, this means that it also represents forms nonexistent in contemporary Czech, like the aorist, the imperfect or the dual).

2. Unlike the synchronic lemma, but similar to the lemmatization implemented in the above-mentioned TLIO project, the hyperlemma also represents all historical and dialectal phonological varieties (e.g. *mouka, múka, muka, móka*) and all spelling varieties (in the case of Czech, from the last orthographic reform, realized in 1849, on). However, unlike the TLIO, one of the goals of DCNC is to make text search as easy as possible and, consequently, one of the fundamental principles is to transcribe old texts which use orthographic systems exceedingly different from the one used in modern Czech; older spelling varieties, characterized by different links between graphemes and phonemes as well as by widespread unsystematicity in the use of many letters and their combinations, are therefore transcribed for the corpus. This means that, for example, new systematic varieties used after 1849, like *engagement* (*angažmá* in modern Czech), *prosa* (now *próza*) are preserved in the corpus and subsumed under one hyperlemma; on the other hand, older spellings like *gegj* or *wedau* are transcribed as *její* and *vedou*, according to the accepted standards of transcription applied in editions of Czech historical texts.

3. In the words that still exist in modern Czech, the form of the hyperlemma is identical with the form used in Czech synchronic corpora, thus paving the way for a smooth transition of aging texts from the synchronic to the diachronic corpus. In the words that do not exist today, the hyperlemma is identical to the newest form attested in the texts or dictionaries. Among other things, this means that the hyperlemmata of the words that – as far as we

know – had fallen out of use before certain phonological and morphological changes were realized in Czech, have the forms existing before the changes; thus, for example, the hyperlemma of the Old Czech verb *lunúti* 'fling oneself after someone', which does not exist in Middle or New Czech, is <lunúti> (not <lunout>, which would be its standard extrapolation in New Czech), because the verb seems to have ceased to exist before the phonological change *ú>ou* was realized in Czech and before infinitive forms ended in -*t* came into use.

4. Unlike the common standards of lemmatization applied in synchronic corpora, the hyperlemma can be a set of lemmata. This rather novel principle has been adopted to make the lemmatization of some forms more logical and consistent, and has been applied in the following two cases:

a. orthographic compounds, such as the English *can't, it's* or Czech *naň, tys, onať, dejtež, dejžť*. In spite of the traditional one-word spelling, these units are combinations of two or more words (or their contractions) and in our opinion each of them should be represented by its own lemma. There is no logical reason, why e.g. *it's* should only be lemmatized as <it> or <be>, or *naň* as <na> or <on>; they should be lemmatized as both <it> and <be>, both <na> and <on>, i.e. they should be represented by sets of lemmata united in one hyperlemma. Thus the hyperlemma of the English *can't* consists of <can> + <not> and the hyperlemmata of the rest of the above Czech examples include respectively <ty> + <být>, <ona> + <ť>, <dát> + <ž>, and <dát> + <ž> + <ť>. In this way, a user of the corpus searching, for example, for all the instances of the verb *být* 'to be' and using the hyperlemma <být> will find, among other forms, the combination *tys*; and of course, he or she will find the same form if they are looking for the instances of the pronoun *ty*.

b. forms shared by coexisting paradigms, such as the Czech plural form *brambory*, which can belong both to the lemma <brambor> and <brambora>. Typically, such cases are results of the transition of words from one paradigm to another, and can be found during the periods of time when forms following both of the paradigms are in use. The phenomenon is relatively rare in modern Czech, but was widespread for centuries in the past – cf. examples like *vrhl* (possible lemmata <vrci> and <vrhnout>), *padl* (<pásti> and <padnout>), *musí* (<musit> and <muset>) etc. In such cases, the hyperlemma includes both of the possible lemmata, reflecting the fact that it is impossible to connect the form to one lemma only. The common practice of using in such cases just one of the two lemmata arbitrarily is irrational and may completely distort our notion about the dynamics of the transition of words from one paradigm to another.

The following graph represents an example of what kind of information can be easily retrieved from a diachronic corpus when hyperlemmata are used in its lemmatization. Since the advantages of the use of hyperlemmata are most obvious in very general searches, the example is of a very general nature indeed.

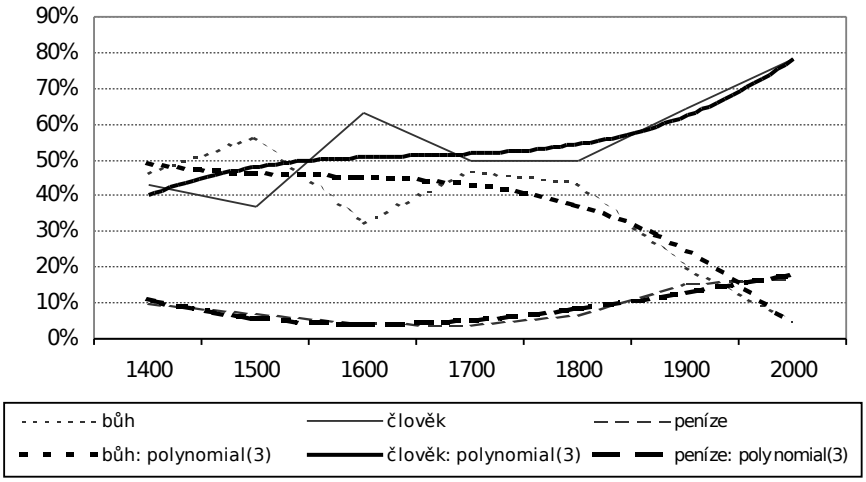


Fig. 1. Relative frequencies of the hyperlemmata bůh ('god'), člověk ('man'), and peníze ('money').

In the graph, DCNC has been used to show how the relative frequencies of the words *bůh* 'god', *člověk* 'man' and *peníze* 'money' were changing in Czech texts during the seven-hundred-year history of the Czech literary language. If normal lemmata had been used to get the results, one would have to search separately for all the phonological forms of the words *bůh, člověk* and *peníze* that have existed in the past (i.e. *bóh, buoh, bůh; ľudé, ľudie, lidé; penieže, penieze, peníze*); with all the forms subsumed under appropriate hyperlemmata, one must only perform three searches, one for each of the hyperlemmata *bůh, člověk* and *peníze.* The results have been gained from virtual subcorpora of DCNC, in which texts had been grouped in one-hundred-year clusters to minimize the influence of uneven representation of various text types and styles in shorter periods of time such as decades.

Looking at the graph, one may be tempted to start to analyze and interpret it in a rather philosophical way. Nevertheless, the example is only intended to demonstrate the use of hyperlemmata and should be taken as such; it is not intended to provoke a debate about changing social values and about the

general validity of the results presented in the graph. In fact, the validity of the results may be rather limited, since the DCNC – with its current modest size of over 2 million running words, about one-third of it accessible on the Internet – can hardly be called representative or sufficient for detailed analyses critically dependent on changes in the proportions of various text types and domains at different periods of time. However, even if the present results can hardly provide a firm basis for far-reaching conclusions about Czech society, and may even undergo some corrections in the future, the example can well represent the advantages of the use of hyperlemmata in diachronic corpora.

References

1. Bollettino dell Opera del Vocabolario Italiano, II (1997)
2. Bollettino dell Opera del Vocabolario Italiano, III (1998)

Semi-automatic Semantic Annotation of Slovak Texts*

Michal Laclavík[1], Marek Ciglan[1], Martin Šeleng[1], Stanislav Krajčí[2], Peter Vojtek[3], Ladislav Hluchý[1]

[1] Institute of Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, Bratislava, 845 07, Slovakia
laclavik.ui@savba.sk
http://ikt.ui.sav.sk/

[2] Ústav informatiky, Prírodovedecká fakulta, UPJŠ, Park Angelinum 9, 040 01 Košice, Slovakia
stanislav.krajci@upjs.sk

[3] Ústav informatiky a softvérového inžinierstva, Fakulta informatiky a informačných technológií, Slovenská technická univerzita, Ilkovičova 3, 842 16 Bratislava
peter.vojtek@fiit.stuba.sk

Abstract. Automated annotation of the Web documents is a key challenge of the Semantic Web effort. Web documents are structured but their structure is understandable mainly for humans, which is the major problem of the Semantic Web. Many solutions for semi-automatic annotation exist based on neural networks, structure analysis or supervised learning techniques. Another possibility is to use pattern based methods for semantic annotation such as SemTag or C-PANKOW. Mentioned methods and solutions are applicable mainly in English and could not work well on highly inflective languages such as Slovak. We have developed the Ontea tool for semi automatic semantic annotation based on regular expression patterns, which together with tools for natural language identification, lemmatization or stemming of Slovak and specialized indexing mechanism provide promising results for semantic annotation of Slovak texts. Language identification is based on Markov processes, which enables to accommodate granularity of text modeling according to attributes of input text. Lemmatization can be done via existing lemmatizer or the simple lemmatizer based on finding same word suffix. The annotation method has been evaluated and success rate measured using recall, precision and F1-measures is over 70%. We can identify objects such as geographical locations: cities, villages, rivers; company names; or other application specific objects. Results can be used for further computer processing and for partial understanding of text by a machine.

* This work is supported by projects NAZOU SPVV 1025/2004, RAPORT APVT-51-024604, SEMCO-WS APVV-0391-06, VEGA 2/7098/27.

1 Introduction

Adding machine understandable information about documents content is one of main challenges of emerging semantic oriented systems. An ultimate goal is to allow machine based reasoning about content of vast quantity of documents produced by human experts and to allow automatic inference of new knowledge. One step towards this goal is to enable automatic and semi-automatic semantic annotation (ASA) of unstructured texts such as web pages, office documents that provides the means to transfer useful information from the documents to the ontology structures.

In this paper, we present combination of a traditional method from information retrieval domain and an annotation method from semantic web, which increase the relevance of automatic annotation. Namely, we have integrated lemmatization, full-text indexing and search mechanism with ASA based on regular expression matching. Indexes of full-text search engine are exploited to gather statistical information about words occurrences in the document collection to estimate the relevance of the ASA outputs.

We present ASA and full-text indexing, including the summarization of the state-of-the-art and description of our approaches (Sections 2, 3, 4); we continue by depicting the integration of the methods and present the experimental results. We conclude the paper with the summary and future work description.

2 Semantic annotation

Automated annotation of the Web documents is a key challenge of the Semantic Web effort. Web documents are structured but their structure is understandable only for humans, which is the major problem of the Semantic Web. Annotation solutions can be divided into manual and semi-automatic methods. This different strategy depends on a use of the annotation. There is number of annotation tools and approaches such as CREAM [15] or Magpie [16] which follow the idea to provide users with useful visual tools for manual annotation, web page navigation, reading semantic tags and browsing [18] or provide infrastructure and protocols for manual stamping documents with semantic tags such as Annotea[1], Rubby[2] or RDF annotation[3].

Semi-automatic solutions focus on creating semantic metadata for further computer processing, using semantic data in knowledge management [19] or in information extraction application. Semi-automatic approaches are based on natural language processing [11] [12], a document structure analysis [13] or

1 http://www.w3.org/2001/Annotea/

2 http://www.w3.org/TR/ruby/

3 http://ilrt.org/discovery/2001/04/annotations/

learning requiring training sets or supervision [14]. Moreover, other pattern-based semi-automatic solutions such as PANKOW and C-PANKOW [10] exist, using also Google API for automatic annotation. Other methods use a variety of pattern matching mechanisms. Another relevant automatic semantic annotation solution and the only one which runs on distributed architecture is SemTag [20]. SemTag uses Seeker [20] information retrieval platform to support annotation tasks. SemTag annotates web pages using Stanford TAP ontology [21].

2.1 Ontea

One of pattern based solutions is also Ontea [19] developed in the NAZOU project. Ontea works on text, in particular domain described by domain ontology and uses regular expression patterns for semi-automatic semantic annotation. Ontea detects or creates ontology elements/individuals within the existing application/domain ontology model according to defined patterns. Ontea tool analyzes text using a regular expression patterns and detects equivalent semantic elements according to the defined domain ontology. Several cross application patterns are defined but in order to achieve good results, new patterns need to be defined for each application. In addition, Ontea creates a new ontology individual of a defined class and assigns detected ontology elements/individuals as properties of the defined ontology class.

3 Document indexing and search

Information retrieval is a process of identifying the text resources of interest from a large collection of documents that would satisfy the user needs. Full text search is a wildly used concept in today information systems for information retrieval. Full-text search engines usually operate over the index structure which keeps information about documents content. Indexes of the documents content are exploited because of the time efficiency of retrieving information from those structures.

3.1 Related work

Documents content indexing is a well established method for information retrieval which crossed the border of academic research and become a part of every day life. Full text search engines are used to find documents stored at users workstations (desktop search engines) as well as to locate resources in intranet and Internet. Main technological challenges addressed by document indexing and search solutions are: index Data Structures, performance (Maintenance, Lookup speed), transformation of words to their base form – usually done by stemming or lemmatization (this topic is discussed in detail in section 3.3 and 3.4 for Slovak), provide rich query mechanism (phrase queries, wildcard queries, proximity queries, range queries), stop words filtering, search

results ranking. A lot of document indexing solutions are available both under commercial and open source licenses with different level of features implementation. We mention several popular systems suitable for intranet and document repository indexing and searching:

Apache Lucene [1] is a search engine designed for high-performance search, supporting large number of query mechanisms. Another search engine is OpenFTS [2] (Open Source Full Text Search engine) using relational database PostgreSQL as backend for storing indexes, provides online indexing of data and relevance ranking. MnoGoSearch [3] is a search engine designed primary for indexing HTML content with HTML specific features such as META tags support, robots exclusion standard support.

3.2 Language identification using Markov processes

Before providing text operations such as lemmatization or stemming, we need to identify language of the text. This can be done using a variety of methods. In our annotation solution we have used the NALIT method which uses Markov processes [23, 24].

The categorization method used in the NALIT tool (Markov Processes based Categorization) [25] is based on method proposed by Dunning [24]. First a statistical model for each category is created in a learning phase. Each of these category models is constructed from pre-selected training text documents, every document represents a certain category in selected categorization. In our case categories represents documents in different languages. The NALIT tool allows to execute proper lematization algorithm as well as to use proper regular expression paterns which are also language dependent. In our experiments NALIT identified correct language of document in 100% cases. In other more general evaluation [25] NALIT was successful in more than 95%.

3.3 Words base forms

One of the driving factors for developing yet another indexing and search engine was the study of stemming and lemmatization methods for the Slovak language and subsequent integration of suitable methods with the search engine. Therefore, we describe in this subsection basic approaches to the words' base form acquisition. Different morphological variants of the natural languages words have in most cases the same or very similar semantic interpretations and can be considered as equivalent for the purpose of information retrieval systems. This means that different morphological forms can be represented by a single representative term. Queries can then produce more relevant responses and the dictionary size needed for representing a set of documents decrease. A smaller dictionary size results in a saving of storage space and processing time. Two main approaches to words' base form

acquisition are lemmatization and stemming. Lemmatization uses the dictionaries produced by human experts to retrieve the base form of a given word. Wordnet [4, 5] is one of sophisticated dictionaries for the English language that can be used for lemmatization. Stemming is a method, which algorithmically derives the stem of a given word; stems produced by stemming algorithms often do not belong to the given natural language, however they identify the class of words from natural language. Popular stemming algorithm for the English language is Porter algorithm [8, 9].

3.4 Lematization of Slovak – Tvaroslovník

One of the tools for lemmatization of Slovak texts is Tvaroslovnik [22]. Its algorithm works on finding the same longest word endings in dictionary words.

For evaluation of Tvaroslovnik we have used 8 documents containing Job offers in Slovak. We ran lemmatization on these documents and results are summarized in the table below.

words	lemmas	percentage
67	42	63%
82	50	61%
133	78	59%
114	71	62%
90	55	61%
82	51	62%
79	56	71%
148	97	66%
795	500	63%

Table 1. Tvaroslovnik lemmatization results

Words column represents word count in the documents. Lemmas column represents found words in other than lemma form, where lemmas were identified. Over half of content is in other than basic word forms even in partially formalized documents as job offers. In this article we evaluate the Ontea algorithm on company and location objects, which names are usually in basic form in job offer documents. Thus we do not compare algorithm success rate with and without lemmatization. Nevertheless preliminary results are promising as demonstrated also by the table above. For example objects related to education or job type are better identified in case of using lemmatization. Examples are:

- Text: „*práca s pokladňou*“ lemma: „*pokladňa*“
- Text: „*stredoškolské s maturitou*“ lema: „*maturita*“

Files used in the experiment as well as log output information can be found on the web[4].

4 `http://ikt.ui.sav.sk/archive/Tvaroslovnik/test_tvaroslovnik_ontea.zip`

3.5 RFTS

We have developed a tool for document indexing and document search, named RFTS (Rich full-text search). The motivation for implementing another search engine was to have an easily extendable and configurable document indexing tool to evaluate novel methods for information retrieval, documents statistical analysis and lemmatization and stemming methods for the Slovak language. Detailed information about documents content is stored in the index structure, including the positions of the word in the documents, phrase number within the document. The words in tool's dictionary are kept in the basic form; different stemming algorithms are used for documents in different languages. The tool exploits relational database to store all the information about documents and its contents. From engineering point of view, it is worth to mention that RFTS functionality in conjunction with Corporate Memory [7] (also developed within the project NAZOU [6]) can be accessed locally (using JAVA interfaces or command line tools) as well as remotely using RPC calls or Web Service interface. The remote access and Web Service interface allows easy integration of the RFTS indexing and search solution in other components and allows rapid prototyping of new tools that require full-text search or some form of statistical analysis of document collection.

4 Integrated annotation method

Ontea's method of automatic annotation based on regular expressions matching showed promising results for domain specific texts. However it suffers from frequent mismatching which creates imprecise instances of ontological concepts. We propose to overcome this obstacle by evaluating the relevance of candidate instances by the means of statistical analysis of the occurrence of the matched words in the document collection. Based on regular expression, Ontea identifies part of a text related to semantic context and match the subsequent sequence of characters to create an instance of the concept. Let us denote the sequence of words related to semantic context by C and word sequence identified as a candidate instance as I. We evaluate the relevance of the new instance by computing the ration of the close occurrence of C and I and occurrence of I:

$$\frac{close\_occurrence(C, I)}{occurrence(I)}$$

RFST indexing tool provides us with enough functionality to retrieve required statistical values computed from the whole collection of documents stored in RFTS index structures.

Let *COLL* be a collection of the documents $d_1, d_2, \ldots, d_n$: $COLL = d_1, d_2, \ldots, d_n$
Let d in *COLL*, *distance* N, and $w_1, w_2, \ldots, w_k$ be the words from natural language.
Function $dist(d, distance, w_1, w_2, \ldots, w_k)$, where $k \leq distance$, denotes the number of distinct word sequences of the length distance containing the words $w_1, w_2, \ldots, w_k$.

We compute the relevance of candidate instance as:

$$relevance(C, I, wordsdist) = \frac{\Sigma\, dist(d, wordsdist, C \cup I)}{\Sigma\, dist(d, (I), I)}$$

If the resulting relevance value exceeds defined threshold, the candidate word sequence I is considered to be a valid instance of the semantic concept related to sequence C. For the experimental evaluation of the approach, the threshold was set manually after inspecting the preliminary relevance values of the generated candidate instances.

The utilization of RFTS brings also an important benefit of treating different morphological word forms as a single class of equivalence represented by the word stem or lemma.

All the documents that are subject to semantic annotation by Ontea must be part of the document collection indexed by RFTS tool.

4.1 Ontea with Lucene

The Ontea annotation method can be also used with Lucene [1] information retrieval library. When connected with RFTS indexing, Ontea asks for relevance based on words distance. When connecting with Lucene, Ontea asks for percentage of occurrence of matched regular expression pattern to detected element represented by word. Example can be Google, Inc. matched by pattern for company search: [\\s]+([-A-Za-z0-9][ ]*[A-Za-z0-9]*),[ ]*Inc[\\.\\s]+, where relevance is computed as Google, Inc. occurrence divided by Google occurrence. RFTS indexing tool also supports this type of queries, however Lucene can achieve better performance. So far we did not compare those two methods of finding relevance of new created individual, but both are implemented.

5 Evaluation

In this chapter we discuss the algorithm evaluation and success rate.

To evaluate the performance of annotation, we used the standard recall, precision and F1 measures. Recall is defined as the ratio of correct positive predictions made by the system and the total number of positive examples. Precision is defined as the ratio of correct positive predictions made by the system and the total number of positive predictions made by the system:

$$Recall = \frac{Match}{Count} = \frac{Relevant\ retrieved}{All\ relevant}$$
$$Precision = \frac{Match}{Ontea} = \frac{Relevant\ retrieved}{All\ retrieved}$$

Recall and precision measures reflect the different aspects of annotation performance. Usually, if one of the two measures is increasing, the other will decrease. These measures were first used to measure IR (Information retrieval) system by Cleverdon [11]. To obtain a better measure to describe performance, we use the F1 measure (first introduced by van Rijsbergen [12]) which combines precision and recall measures, with equal importance, into a single parameter for optimization. F1 measure is weighted average of the precision and recall measures and is defined as follows:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

5.1 Test set of documents

As reference test data, we used 500 job offers downloaded from web using wrapper which prepared us some structured data. This was converted to a defined ontology and manually checked and edited according to 500 html documents representing reference job offers. Ontea processed reference html documents using the reference ontology resulting in new ontology metadata consisting of 500 job offers, which were automatically compared with reference manually checked job offers ontology metadata.

5.2 Target ontological concepts for identification

In this test, Ontea used simple regular expressions matching from 1 to 4 words starting with a capital letter. This experiment is referred to as "Ontea" in next chapter. In the second case we used domain specific regular expressions which identified locations and company names in text of job offers and Ontea was also creating individuals in knowledge base, while in the first case Ontea did not create extra new property individuals and only searched for relevant individuals in knowledge base. This second case is referred to as "Ontea creation". The third case used previously described RTFS indexing tool to find out if it is feasible to create new individual using word occurrence functionality of RTFS this case is referred as "Ontea creation, RFTS"

So we did our experiments in 3 cases:

- Ontea: searching relevant concepts in knowledge base (KB) according to generic patterns
- Ontea creation: creating new individuals of concrete application specific objects found in text
- Ontea creation, RTFS: Similar as previous with the feedback of RTFS to get relevance computed above word occurrence. Individuals were created only when relevance was above defined threshold which was set up to 10%

We have used following regular expressions:

- Generic expression matching one or more words in text. This was used only to search concepts in KB.
  ([A-Z][-A-Za-z0-9]+[\\s]+ [-a-zA-Z]+)
- Identifying geographical location in text and if not found in KB individual was created
  Location:[\\s]*([A-Z][-a-zA-Z]+[ ]*[A-Za-z0-9]*) used for English
  [0-9]{3}[ ]*[0-9]{2}[ ]+([A-Z][^\\s,\\.]+[ ]*[^0-9\\s,\\.]*)[ ]*[0-9\\n,]+
  used for Slovak text where settlement name is usually next to ZIP code
- Identifying company in the text, this was used also with other abbreviations such as Ltd or a.s., s.r.o. for the Slovak language
  [\\s]+([-A-Za-z0-9][ ]*[A-Za-z0-9]*),[ ]*Inc[\\.\\s]+ for English
  [\\s]+([A-Z][^\\s,\\.]+[ ]*[^\\s,\\.]*[ ]*[^\\s,\\.]*)[,]*s\\.r\\.o\\.[\\s]+
  used for Slovak texts

5.3 Experimental results

Experimental results using precession, recall and F1-measures are in Table 1. In the table we compare our results with other semantic annotation approaches and we also list some advantages and disadvantages. The column "relevance" corresponds to F1-measures in case of Ontea, in case of other methods, it can corespondto other evaluation techniques. For example for C-PANKOW, relevance is referred to as recall.

The experimental results are summarized in Table 1. Rows relevant to our annotation approach are in grey color, showing success rate of three evaluation cases mentioned in the previous chapter. The row "Ontea creation, RTFS" case is the most important concerning evaluation where we combined indexing and annotation techniques. By using this combination we were able to eliminate some not correctly annotated results. For example by using [Cc]ompany[:\\s]*([A-Z][-A-Za-z0-9][ ]*[A-Za-z0-9]* regular expression in the second case we have created and identified companies such as This position or International company which were identified as not relevant in the third case with RTFS.

Similarly Ontea creation identified also companies like Microsoft or Oracle which is correct and in combination with RTFS this was eliminated. Because of this issue, recall is decreasing while precession is increasing. Here it seems that RTFS case is not successful but the opposite is true because in many texts Microsoft is identified as products e.g. Microsoft Office so if we take more text to annotate it is better to not annotate Microsoft as company and decrease recall. If we annotate Microsoft as company in other texts, used in context of Microsoft Office we would decrease precision of annotation.

So it is very powerful to use presented annotation technique in combination with indexing in applications where precision needs to be high.

	Method	Rel. %	Prec. %	Recall %	Disadvantages	Advantages
Ontea	**regular expresions, search in knowledge base (KB)**	**71**	**64**	**83**	**high recall, lower precision**	**high success rate, generic solution, solves the duplicity problem, fast algorithm**
SemTag	disambiguity check, search in KB	high	high		works only for TAP KB and English	fast and generic solution
Ontea creation	**regular expressions (RE), creation of individuals in KB**	**83**	**90**	**76**	**application specific patterns are needed**	**supports Slovak**
Ontea creation RFTS, TS	**RE, creation of individuals in KB + RFTS**	**73**	**94**	**69**	**low recall**	**disambiguities are found and not annotated**
Wrapper	document structure	high	high		zero success with unknown structure	high success with known structure
PANKOW	pattern matching	59			low success rate	generic solution
C-PANKOW	POS tagging, and pattern matching Qtag library	74		74	suitable only for English, slow algorithm	generic solution
Hahn et al.	semantic and syntactic analysis	76			works only for English, not Slovak	
Evans	clustering	41			low success rate	
Human	manual annotation	high	high	high	problem with creation of individuals duplicities, inaccuracy	high recall and precision

Table 2. Annotation experimental results

6 Conclusion

By integrating information retrieval system based on lemmatization (Tvaroslovnik), full-text indexing and search (RFTS) and the semantic annotation tool (Ontea), we were able to improve the results of the automatic semantic annotation process for domain specific documents – increasing the precision of newly created instances at least by 4%. However, the recall of identified instances decreased. This is an advantageous trade-off as the ontological data precision is the primary goal of our work on automatic annotation.

We have also identified, but not proved yet, that using a large collection of experimental texts or documents "Ontea creation" (without RTFS indexing) precission will decrease and in combination with RTFS precision will still stay over 90%, which is very high for semi-automatic annotation solution.

Ontea algorithm disadvantage is a requirement to set up domain specific patterns. While annotation methods as C-PANKOW are more generic, Ontea is a simpler, faster solution with a better success rate, suitable for knowledge management, information extraction or knowledge acquisition applications, where large number of documents needs to be annotated. SemTag on the other hand is faster than Ontea but is not able to create new ontology metadata, only identify their existence in the knowledge base.

Main advantages of described method are: supporting the Slovak language, fast algorithm comparing to other methods, instance duplicity identification and very high precision.

7 Future work

The presented work is an intermediate result on our research on automatic and semi-automatic semantic annotation. Subsequent effort will be focused on studying and tuning instance relevance computation from the document collection. We plan to study the effect of increasing the distance parameter (distance larger than cardinality of C and I sequences), relevance computation based on C and I membership in a phrase in documents (instead of word distance concept), document preprocessing methods that would pre-format the selected terms in order to increase regular expressions matching precision. We will study the effects of extending the document collection (that form the basis for our statistical analysis) by text, which do not belong to the specific domain and we will examine the results obtained from document from different domains. We will also analyze the effects of document collection size on the instance relevance identification.

We would also like to evaluate Ontea with use of Lucene and evaluate better Ontea method on Slovak texts.

References

1. Hatcher E., Gospodnetić O., Lucene in Action, Manning (12 Jan 2005), ISBN: 1932394281
2. OpenFTS – http://openfts.sourceforge.net
3. mnoGoSearch – http://www.mnogosearch.org
4. Miller, G.:WordNet: An On-line Lexical Database, Special Issue, International Journal of Lexicography, Vol. 3, Num. 4, 1990
5. WordNet: http://wordnet.princeton.edu/
6. Návrat, P., Bieliková, M., Rozinajová, V.: Methods and Tools for Acquiring and Presenting Information and Knowledge in the Web. In: CompSysTech 2005, B. Rachev, A. Smrikarov (Eds.), Varna, Bulgaria, June 2005. pp. IIIB.7.1-IIIB.7.6.
7. M. Ciglan, M. Babik, M. Laclavik, I. Budinska, and L. Hluchy. Corporate memory: A framework for supporting tools for acquisition, organization and maintenance of information and knowledge. In J. Zendulka, editor, Proc. of 9th Int. Conf. on Information Systems Implementation and Modelling (ISIM 2006), pages 185--192. MARQ, Ostrava, 2006.
8. Jones, K. S., Willet, P: Readings in Information Retrieval, San Francisco: Morgan Kaufmann,1997, ISBN 1-55860-454-4.
9. Van Rijsbergen, C. J., Robertson, S. E., Porter, M. F. New models in probabilistic information retrieval. British Library Research and Development Report, no. 5587, British Library, London, 1980.
10. Cimiano P., Ladwig G., Staab S.: Gimme' the context: context-driven automatic semantic annotation with c-pankow. In WWW '05, pages 332-341, NY, USA, 2005. ACM Press. ISBN 1-59593-046-9.
11. Madche A., Staab S.: Ontology learning for the semantic web. IEEE Intelligent Syst., 16(2):72-79, 2001
12. Charniak E., Berland M.: Finding parts in very large corpora. In Proceedings of the 37th Annual Meeting of the ACL, pages 57-64, 1999.
13. Glover E., Tsioutsiouliklis K., Lawrence S., Pennock D., Flake G.: Using web structure for classifying and describing web pages. In Proc. of the 11th WWW Conference, pages 562-569. ACM Press, 2002.
14. Reeve L., Hyoil Han: Survey of semantic annotation platforms. In SAC '05, pages 1634-1638, NY, USA, 2005. ACM Press. ISBN 1-58113-964-0. doi: doi.acm.org/10.1145/1066677.1067049
15. Handschuh S., Staab S.: Authoring and annotation of web pages in cream. In WWW '02, pages 462-473, NY, USA, 2002. ACM Press. ISBN 1-58113-449-5. doi: http://doi.acm.org/10.1145/511446.511506.
16. Domingue J., Dzbor M.: Magpie: supporting browsing and navigation on the semantic web. In IUI '04, pages 191-197, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-815-6.

17. Uren V. et al.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics: Science, Services and Agents on the WWW, 4(1):14-28, 2005.
18. Uren V. et al.: Browsing for information by highlighting automatically generated annotations: a user study and evaluation. In K-CAP '05, pages 75-82, NY, USA, 2005b. ACM Press. ISBN 1-59593-163-5
19. Michal Laclavik, Martin Seleng, Emil Gatial, Zoltan Balogh, Ladislav Hluchy: Ontology based Text Annotation – OnTeA; Information Modelling and Knowledge Bases XVIII. IOS Press, Amsterdam, Marie Duzi, Hannu Jaakkola, Yasushi Kiyoki, Hannu Kangassalo (Eds.), Frontiers in Artificial Intelligence and Applications, Vol. 154, February 2007, pp.311-315. ISBN 978-1-58603-710-9, ISSN 0922-6389.
20. S Dill, N Eiron, et al.: A Case for Automated Large-Scale Semantic Annotation; Journal of Web Semantics, 2003
21. R.Guha and R. McCool. Tap: Towards a web of data. `http://tap.stanford.edu/`.
22. Stanislav Krajči, Róbert Novotný: Hľadanie základného tvaru slovenského slova na základe spoločného konca slov; In: 1st Workshop on Intelligent and Knowledge oriented Technologies – WIKT 2006 Proceedings; Michal Laclavik, Ivana Budinska, Ladislav Hluchy (Eds.); November 28 – 29, 2006; Bratislava, Slovakia; March 2007; ISBN 978-80-969202-5-9; March 2007
23. William J. Teahan (2000), Text classification and segmentation using minimum cross entropy, In: RIAO-00, 6th International Conference "Recherche d'Information Assistee par Ordinateur", Paris, FR.
24. Ted Dunning (1994), Statistical identification of language. Techical Report MCCS-94-273, Computing Research Lab (CRL), New Mexico State University.
25. Peter Vojtek, Mária Bieliková (2007), Comparing Natural Language Identification Methods based on Markov Processes. In: Slovko – International Seminar on Computer Treatment of Slavic and East European Languages, Bratislava. Accepted.

Terminology and Terminological Activities in the Present-Day Slovakia

Jana Levická

Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia
janal@korpus.juls.savba.sk
https://data.juls.savba.sk/std

Abstract. The author of this paper aims to present current situation and problems of terminology monitoring and administration in Slovakia that lacks appropriate terminology discussion and terminography activities. In order to raise the awareness of the terminological issues, a centralised national term base was created intending to draw on the subcorpora of respective fields. The subcorpora not having been completed so far, terminology entries are to be filled in with available internet data. The author draws special attention to the key information of the terminology entry – definition, which, however, is not always to be found on the internet. Therefore, the team opted for the so-called defining context as a temporary substitute. The structure and content of both definition and defining context is to be analysed, which may serve for further semi-automatic retrieval in the specialised corpora.

Slovak language has seen spontaneous and uncontrolled accumulation of terms for more than 17 years due to the political, economic and social transformation of Slovakia after 1989, which resulted in coinage of excessive and often unnecessary terminological variants of both domestic and foreign provenience.

In spite of a rich history of terminological activities, Slovak society has been facing a double vaccum in its post-war history – on one hand in terms of analysis and development of foreign and domestic terminological theories, technologies and methodologies and on the other studies comparing foreign and Slovak terminological systems. Moreover, the new political and economic situation has caused a massive braindrain in the academic and scientific sphere, thus Slovakia lacks available full-time and skilled terminologists not to mention sufficient linguistic curricula and existent terminological education.

Although terminological activity in Slovakia has suffered the greatest fallout in the last 50 years, it was not suppressed completely. Quality bilingual or unilingual specialised dictionaries have been sporadically published, quality theses elaborated and articles on the theory of terminology written and published especially in the revue *Kultúra slova* but have only seen modest feedback.

Some institutions, aware of the urgent need, started with scarce but key terminological activities, i.e. setting up of terminology databases – e.g. the National Bank of Slovakia and the Slovak Institute of Technical Normalisation. However, those are only domain limited databases with specific, very narrowly defined aims – the former is an in-house tool for employees of the bank while the latter is used by the creators and translators of technical norms. As a mater of fact, these banks do not provide any access for lay public.

At the same time, there has been an external demand for consistent and unified Slovak terminologies for the purpose of drafting and translating European legislation into Slovak as they have been revealed to be unsatisfactory.

1 Place of a corpus in the context of terminological activities

It is common knowledge that effective specialised communication requires unambiguous terminology. The contemporary Slovak state of affairs does not contribute to coherent and intelligible science neither for specialists, producers and lawmakers nor for teachers, translators and interpreters. Therefore terminology monitoring, coordination, analysis of the special vocabularies and unification of pertinent results in the form of glossaries, dictionaries, terminological standards or terminology databases, which present-day Slovakia lacks, is considered by the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences to be a priority.

Centralisation of various terminologies under one administration, continuous modification and updating of term records and narrow collaboration of terminological boards, translators and specialists is nowadays regarded to be the only way of terminological harmonisation, and consequently standardisation. Contemporary terminological tendencies stress the model of the text and corpus approach as a sine qua non prerequisite of every terminological project. The process of systematic gathering of terms is based exclusively on representative corpora, supervised and validated by specialists and terminologists. As Sager (1990:131) puts it, information extracted from a text represents a reliable indicator of changes and ensures the only plausible data for building and revising terminological records.

It is therefore only natural that a terminology database project be started by the Corpus Department of the Institute for it had all the resources and tools at its disposal – textual base of the Slovak National Corpus itself and software ones for automatic annotation of Slovak texts.

2 Project of the Slovak National Terminology Database

The SNC project aims to set up a terminological database provided with both conceptual and linguistic information, inspired by foreign examples, mostly Canadians but, of course, adapted to Slovak needs and present-day possibilities of the L. Štúr Institute of Linguistics. The team expects to cooperate and exchange the data with leading European database IATE which is why the EUROVOC 4.2 Thesaurus was chosen as the classification system.

The starting point of the Slovak Terminology Database Project dates back to the autumn of 2005 when the SNC team 2005 launched the analysis of existing terminology databases, subsequently proceeding to the design of the term record layout based partly on translation and interpreting needs survey. The team chose and adapted the appropriate software for the database along the way. The SNC policy of text acquisition also had to be modified and the focus was shifted towards economic and legal texts for the purpose of creating specialised subcorpora and further automatic extraction of terms and possibly definitions as well as other terminological data from specialised corpora.

Project methodology, as we already mentioned, has drawn inspiration from the textual terminology approach to the terminology extraction of lexical units – potential terminological units from running specialised texts and identification of the concept they refer to. No less importance is to be attached to the generation of software tools for automatic extraction of terms and possibly definitions from specialised subcorpora.

As far as the data categories of the term record are concerned, the team opted for an 11-item term record containing 7 obligatory fields. In order to satisfy the needs of professionals, lay public and last but not least the translation and interpreting public, the team decided to not only make the field of definition and domain obligatory, but also context, related terms and sources of definition and context. The remaining 4 optional fields feature synonym, foreign language equivalent, comment and links to reliable web pages.

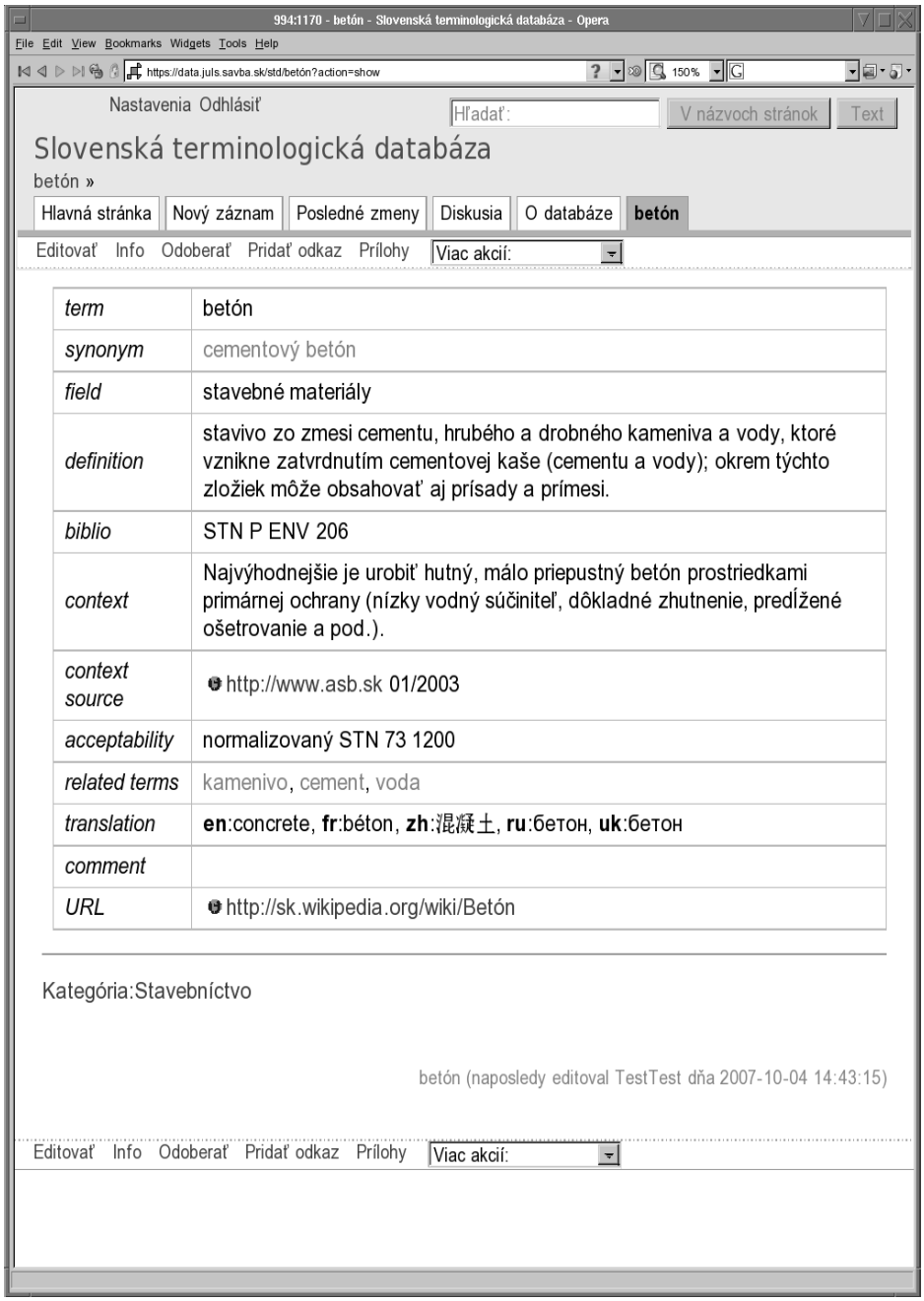


Fig. 1. Example of a term entry.

3 Changes in the project strategy

However, the project did not receive funding from the state run Slovak Research and Development Agency. The team had to take a different path and began to work within a reduced scale, i.e. instead of creating new records by filling in obligatory fields (1, 3, 4, 5, and 6), the emphasis was shifted on re-using and adapting existing quality terminology resources published in Kultúra slova in particular, as the team received a copyright license for their non-commercial use as well as some of those that had been elaborated by our close collaborators.

For the time being, the database offers almost 3000 terminological records covered by 8 domains (Astronomy, Security and Law, Migration Policy, Construction, Corpus Linguistics, Phraseology, Phonetics and Phonology, Bilingualism, Civil Security, Historical Linguistics, Fire Protection) with more or less completed terminology fields – term and usually definition, source, less often synonym, sometimes related terms, and comment.

In order to meet one common form of term record we had to proceed to the harmonisation of different editorial practices. This alternative thus brought about different issues to discuss and deal with.

On the formal level, the most frequent flaws include term records written in incorrect Slovak, excessive punctuation within the definition and formal treatment of polysemous terms.

As far as the content of the term record is concerned, problem issues cover harmonisation of EUROVOC descriptors with the classification of the original terminological ones, which are usually fine-grained; evaluation of relevancy of terms belonging to the terminology of a specific author or school; identification and distinction of different types of related terms; ascribing the status of terms; limitation of inconsistency of terms records as such; treatment of nomenclatures and accompanying data etc.

In spite of all rules and efforts, definition-related discrepancies include incompleteness, inconsistency, amateurishness and subjectivity as well as distinction and splitting of original entries into definition, context and comment fields.

4 Definition and context analysis

The second part of the paper will be focused on the key obligatory field of the proposed term record – definition. Upon a brief introduction dealing with its typology and function, the attention will be paid to its delimitation and

structure, which will subsequently help to draw analogy with the so-called defining context as the nearest substitute of the definition.

4.1 Definition rôle and typology

Definition represents a sort of microsystem consisting of hierarchically ordered characteristics of a concept and their relations, which enable to describe, circumscribe and distinguish the concept. However, features included in a definition reflect the concept structure but can never cover the totality of a concept (Seppälä 2004: 37), hence the origin of variant definitions of the same concept.

Terminological practice has recorded numerous typologies of definitions based on different perspectives – e.g. situation of use, defining mode, formal composition, content of the definition, rôle, and editing practice, the choice being dictated by the target audience, aim of terminographic project and respective domain.

Terminological theory fosters traditional and most frequent Aristotelian definition, which begins with the nearest superordinate concept and specific features (*genus term* and *differentia* or *characteristics*), i.e. "systematically identifies a concept with respect to all others in the particular subject field" Sager (1990:42). This so-called ideal definition with specific editing criteria to follow is referred to as **classic, intensional or comprehensive definition**.

Example: Samozhutniteľný betón: *betón, ktorý je schopný tiecť a spevňovať sa účinkom vlastnej hmotnosti, úplne vyplniť debnenie, aj vo vysoko vystuženom priereze, za súčasného udržiavania homogenity a bez potreby hocijakého dodatočného zhutnenia.*

"Samozhutniteľný betón" is defined with the aid of the closest genus, which is "concrete" as well as the characteristics, which distinguish this specific concrete from all other types of "concretes":

- *schopný tiecť a spevňovať sa účinkom vlastnej hmotnosti*,
- *úplne vyplniť debnenie, aj vo vysoko vystuženom priereze, za súčasného udržiavania homogenity a bez potreby hocijakého dodatočného zhutnenia.*

Comprehensive definition has its counterpoint in the **extensional definition** that ISO 740 defines as "an enumeration of all species which are all on the same level of abstraction".

Example from the STD: materiálno-technické vybavenie jednotiek požiarnej ochrany: *vybavenie jednotiek požiarnej ochrany zahŕňajúce požiarnu techniku, vecné prostriedky požiarnej ochrany na výkon odborných služieb a špecializovaných činností, hasiace látky, ako aj ochranné prostriedky na účinné vykonávanie zásahu a činnosti na požiarnej stanici.*

It is a common phenomenon to find so-called mixed definitions that consist of a comprehensive as well as extensional part.

4.2 Context rôle and typology

What is meant by the "context"? **Context** as such is defined in the ISO 12620 as a "text which illustrates a concept or the use of a designation", or "a text or part of a text in which a term occurs". The two definitions indicate several functions of the context exploited in terminological works according to which it is possible to distinguish three types of contexts:

1. **language context** that can be identified with *lexicographic example*, which is a simple occurrence of a term indicating neither conceptual nor linguistic information.
2. **linguistic context** is a context that illustrates the linguistic function of a term in discourse but provides no conceptual information e.g. typical syntactic structures or collocations.
3. **defining context** that Canadian term base Termium considers as the one that sheds light on the distinctive characteristics of the concept. ISO 12320:1995 informs that it "contains substantial information about a concept but does not possess the formal rigor of a definition", i.e. the substantial information can cover essential characteristics of the concept, its purpose, consequence of the action/event etc. Sue Ellen Wright offers a slightly extended ISO definition and at the same time a less restricted one: „defining context contains definitive information that may look very much like a definition, but is *incomplete* or doesn't have the right form for a definition".

In conclusion, the comprehensive definition is made up of a genus term and essential characteristics while the defining context does not have to explicitly express the superordinate genus term and all the essential characteristics but must include enough information to identify the concept.

Due to the lack of definitions, not speaking of the quality of existing ones, it is planned to use the defining context in Sue Ellen Wright's sense, as a provisional terminological data until a proper definition is found or formulated as it is the closest one to the definition and the most pertinent one for classification and identification purposes of a term.

5 Definition and context mining: Slovak National Corpus vs Internet

Maintaining the idea of creating new term records in the near future with only a limited team of unskilled persons (for the specialists can be contacted only for the revision/validation process), the team has decided to use ready-made definitions and defining contexts that may be available in our Slovak National Corpus (SNC) and Internet.

Therefore, research focused on types, frequency and quality of definitions and defining contexts in both sources has been carried out within domain specific areas. For the purpose pilot research on the prescriptive domain of construction was selected because, i.e. terms referring to construction materials. The idea was to identify the linguistic and conceptual structure of the definition and thus find a repeating structure or pattern of defining elements, i.e. genus term and differentia and key words expressing them might enable semi-automated extraction of both terminological types of data.

The analysis is based on the French-written thesis by Selja Seppälä (2004) titled *Conceptual Composition and Formalisation of the Terminograhical Definition* and her two typologies: 13-item typology of genus conceptual classes and 22-item typology of differentia, which she used for manual annotation of a corpus made of 500 definitions after their identification and isolation.

5.1 Corpus search

Since a construction subcorpus is not available yet, we started our analysis by searching the 3 juls-all version of the SNC within the text annotated as TEC domain, which yielded 801 occurrences of the lemma *betón* by means of regular expressions. On the basis of the collocation statistics we could identify the most frequent complex terms out of more than 40.

As for the typology of sources, the virtual subcorpus consists of specialised magazines *ASB* and *Materiálové inžinierstvo* (25 issues and 2 issues respectively), a commercial leaflet and a semi-specialised book representing the two remaining sources. Occurrences of *betón* from a popular internet magazine inZIne and the specialised IT magazine PC REVUE had to be classified as unacceptable and therefore left out from further analysis.

According to the Canadian *Handbook of Terminology* classification of relevant and reliable terminological sources, specialised and popularised periodicals are ranked 4th while the brochures and publicity flyers are 5th.

However, manual search of the abovementioned occurrences revealed to be highly disappointing for it yielded only 3 relevant results: 1 definition and 2 defining contexts deriving from the same source – *ASB magazine*. We present their annotated structure and content as follows:

Definition

> **SSC** *je* [*betón*] INANIMATE→ARTIFICIAL *[s veľkou pohyblivosťou a schopnosťou tiecť bez pôsobenia vonkajších dynamických síl]* PHYSICAL PROPERTY, [*s mimoriadnou odolnosťou proti rozmiešavaniu a segregácií hrubých zložiek čerstvého betónu]* UTILITY. Source: ASB – 2004/07

Defining contexts

> *Výroba sa sústreďuje do centrálnych výrobní betónu, kde sa vyrába tzv.* **transportbetón**, *t.j. čerstvý betón* INANIMATE→ARTIFICIAL *[mäkkej alebo tekutej konzistencie]* PHYSICAL PROPERTY, *ktorý [sa prevezie auto-*

domiešavačom] INSTRUMENT *a ktorý [sa na stavbe ukladá do debnenia]* UTILITY. Source: ASB 2003/07

Vysokopevnostný betón HSC *(High Strength Concrete) [patrí do pomerne nedávno (v roku 1993) vytvorenej skupiny vysokohodnotných betónov – HPC]* WHOLE *(High Performance Concrete), pre ktoré je charakteristická [pevnosť v tlaku vyššia než 65 MPa]* MEASURABLE PROPERTY Source: ASB 2005/07-08

Comparing the two defining contexts to the definition, the first one introduces defining elements by means of the explanatory structure *t.j.* (i.e.), which is synonymous to the verb *be.* The other uses a sort of paraphrase of the conceptual class indicating membership of this concrete to a specific group and differentia are launched with the syntactic structure: *je charakteristický.*

5.2 Internet search

The research proceeded by searching definitions and defining contexts of 12 complex terms in their nominative form selected with respect to their highest absolute frequency in the 3-juls-all SNC within texts annotated as TEC domain.

Internet sources include again the specialised magazine ASB, two association portals (`http://www.betonracio.sk` and `http://www.beton.sk`), two educational pages, European Directive on Self-Compacting Concrete and the individual professional portal `http://www.dalnice.com`.

	Complex term	Frequency in the SNC	Google search occurrences
1.	čerstvý betón	72	153
2.	pohľadový betón	31	92
3.	podkladový betón	25	116
4.	asfaltový betón	19	82
5.	cementový betón	11	32
6.	predpätý betón	9	106
7.	vysokopevnostný betón	8	20
8.	vystužený betón	7	40
9.	samozhutňujúci betón	6	14
10.	vodotesný betón	5	51
11.	vysokohodnotný betón	5	23
12.	samozhutniteľný betón	4	21

Table 1.

In spite of a relatively high number of occurrences, this research resulted in identifying only 5 defining contexts and 6 definitions of googled complex terms. In the case of five complex terms we did not get any relevant result. We present the annotated conceptual structure of genus terms and differentia as follows:

Definitions

1. **čerstvý betón** – betón INANIMATE→ARTIFICIAL, ktorý je [úplne zamiešaný a je ešte v takom stave, ktorý umožňuje jeho zhutnenie zvoleným spôsobom] CONDITION. `http://www.beton.sk`
2. **vyysokopevnostný betón** – *betón* INANIMATE→ARTIFICIAL*, ktorý má [pevnostnú triedu v tlaku väčšiu ako C 50/60 (B60)]* MEASURABLE PROPERTY *[pre obyčajný a ťažký betón]* TYPE *a [LC 50/55 (B55)]* MEASURABLE PROPERTY *[pre ľahký betón]* TYPE. `http://ww.beton.sk`.
3. **samozhutniteľný betón** *je inovovaný betón* INANIMATE→ARTIFICIAL*, ktorý [nevyžaduje vibrovanie pri ukladaní a zhutňovaní.]* CONDITION `http://w.savt.sk/dokumenty/public/eur-pske-smernice-pre-samozhutnite-318-ny-bet-n/view.html`
4. **samozhutniteľný betón** – *betón* INANIMATE→ARTIFICIAL*, ktorý je [schopný tiecť a spevňovať sa účinkom vlastnej hmotnosti]* PHYSICAL PROPERTY*, [úplne vyplniť debnenie, aj vo vysoko vystuženom priereze, za súčasného udržiavania homogenity a bez potreby hocijakého dodatočného zhutnenia]* UTILITY
   `http://ww.savt.sk/dokumenty/public/eur-pske-smernice-pre-samozhutnite-318-ny-bet-n/view.html`
5. **cementový betón** *je zmes* INANIMATE→ARTIFICIAL *vysoko kvalitných drtených kamenív, cementu a vody* CONTENT.
   `http://www.dalnice.com/pojmy/slovnicek.htm`
6. **asfaltový betón** – *je to [zmes]* INANIMATE→ARTIFICIAL *[hutného kameniva [s uzavretou zrnitosťou zahorúca obalená asfaltom]* PHYSICAL PROPERTY*]* CONTENT. `http://ww.asb.sk`

All definitions, being comprehensive ones, rank defined terms among the conceptual classes INANIMATE→ARTIFICIAL expressed by the designation */betón/* linked with the rest of defining elements by verb junctors *be* or *have,* which confirms Seppälä's findings (2004:138).

As for the categories of characteristics, the most frequently identified are MEASURABLE PROPERTY (2x) and CONDITION (2x), UTILITY, TYPE, CONTENT and PHYSICAL PROPERTY.

Defining contexts

1. **čerstvý betón** *je [v nezatvrdnutej forme dopravovaný na miesto spotreby]* CONDITION – *[stavenisko]* PLACE INDICATOR, *kde je [pripravený pre uloženie do debnenia a po zatvrdnutí je spravidla hlavnou nosnou časťou stavebných konštrukcií]* UTILITY
   `http://www.savt.sk/r-zne/charakteristika-vyroby-transportbet-nu.html`
2. **samozhutniteľný betón** *možno charakterizovať ako [extrémne tekutý]* PHYSICAL PROPERTY *[betón]* INANIMATE→ARTIFICIAL, *ktorý [dokáže úplne vyplniť priestor debnenia alebo formy a zhutniť sa bez použitia vibrácie alebo iného spôsobu zhutňovania].* UTILITY
   `http://www.betonracio.sk`
3. *Hlavnou charakteristickou vlastnosťou* ***SCC*** *je [schopnosť tečenia čerstvého betónu bez pôsobenia vonkajších dynamických síl]* PHYSICAL PROPERTY, *[odolnosť proti rozmiešaniu a segregácii hrubých zŕn kameniva a schopnosť zhutnenia vlastnou hmotnosťou]* UTILITY.
   `http://www.asb.sk`
4. **Predpätý betón** *vzniká [kombináciou betónu a predpätej výstuže]* CONTENT `http://www.soustavebne.sk/betony/predpaty_beton.htm`
5. **Vodotesný betón** *je [betón]* INANIMATE→ARTIFICIAL, *ktorý [odoláva tlakovej vode tak, že na jeho vzdušnej strane nevzniknú viditeľné priesaky, prípadne vlhké škvrny]* PHYSICAL PROPERTY.
   `http://fzki.uniag.sk/`

Only two out of five defining contexts rank defined terms among conceptual classes INANIMATE→ARTIFICIAL expressed by the designation */betón/* linked with the rest of defining elements by verb junctors *be*, *vznikať* or by means of classifying paraphrases *možno charakterizovať, charakteristickou vlastnosťou.* Lack of formal rigour is evident in the wording of all texts.

As for the categories of characteristics, we identified most frequently PHYSICAL PROPERTY (3x) and UTILITY (3x), followed by CONDITION, PLACE INDICATOR, CONTENT.

Conclusion

Due to the insignificant number of found definitions and defining contexts, their comparison is not of much relevancy, moreover, only in the case of two complex terms the search revealed both defining contexts and definitions. However, analysed definitions and contexts show the same conceptual class of the genus term and three same characteristics (UTILITY, PHYSICAL PROPERTY,

CONDITION). Excerpted definitions seem to be more precise (see number of MEASURABLE PROPERTY) while defining contexts appear to be more function-oriented (see UTILITY). The author is aware that these tendencies must be verified in a representative and balanced specialised subcorpus in order to ensure automated research of collocated defined words with their genus and differentia terms.

References

BÉJOINT, Henri: La définition en terminographie. In: Aspects du vocabulaire. Ed. P. J. L. Arnaud et Ph. Thoiron. Travaux du CRTT: Lyon, Presses Universitaires de Lyon, pp. 19 – 25.

BOURIGAULT, Didier, AUSSENAC-GILLES, Nathalie, CHARLET, Jean: Construction de ressources terminologiques ou ontologiques à partir de textes: un cadre unificateur pour trois études de cas. In *Revue d'Intelligence Artificielle (RIA)*, Techniques Informatiques et structuration de terminologiques, Pierrel J.-M. Et Slodzian M. (Ed.), Paris : Hermčs. 2004, vol. 18, n. 1, pp. 87 – 110.

CABRÉ, Maria Teresa: La terminologie – théorie, méthode et applications. Ottawa: Armand Colin/PUO 1998.

ČERMÁK, F. et alii.: Manuál lexikografie. Praha: H&H 1995.

GAUDIN, François: Socioterminologie. Une approche sociolinguistique de la terminologie. Bruxelles: De Boeck et Larcier 2003.

GESCHÉ, Véronique: Évaluation des définitions d'ouvrages. In Meta, 1997, roč. 42, n. 2, pp. 374 – 390.

HORECKÝ, Ján: Základy slovenskej terminológie. Bratislava: Vydavateľstvo SAV 1956.

ISO 704 (2000): Travail terminologique – principes et méthodes. International Organization for Standardization.

ISO 860 (1996): Travaux terminologiques – Harmonisation des termes. International Organization for Standardization.

ISO 1087-1 (2000): Travaux terminologiques – Vocabulaire. International Organization for Standardization.

ISO 12620 (1999): Aides informatique en terminologie. International Organization for Standardization.

KOCOUREK, Rostislav: La langue française de la technique et de la science. Wiesbaden: Brandstetter 1991.

LERAT, Pierre: Les langues spécialisées. Paris: PUF 1995.

LINO, Maria Teresa Rijo da Fonseca: Base de données textuelles et terminographiques. In Meta, 1994, vol. 39, n. 4, pp. 786 – 789.

MASÁR, Ivan: Príručka slovenskej terminológie. Bratislava: VEDA 1991.

NAZARENKO, Adeline, HABERT, Benoît, SALEM, André: Les linguistiques de corpus. Paris: Armand Colin 1997.

NUOPPONEN, Anita: Terminological information and activities in the world wide web. In Christian Galinski & Klaus- Dirk Schmitz (eds.), *TKE'96, Terminology and Knowledge Engineering.* 92–99. Frankfurt: INDEKS-Verlag.

OTMAN, Gabriel: Les bases de connaissances terminologiques: les banques de terminologie de seconde génération. In Meta, 1997, vol. 42, n. 2, pp. 244 – 256.

PAVEL, Silvia, NOLET, Diane: Handbook of Terminology. Translation Bureau, 2001 http://www.translationbureau.gc.ca

SAGER, Juan Carlos: A Practical Course in Terminology Processing. Amsterdam/ Philadelphia: John Benjamins 1990.

SEPPÄLÄ. Selja: Composition et formalisation conceptuelles de la définition terminographique. *Mémoire pour l'obtention du DEA*. Genève: Université de Genève 2004.

SCHWARZ, Jozef: Vybrané teoretické a metodologické problémy terminografie: poznatky z tvorby České terminologické databáze knihovnictví a informační vědy. In Národní knihovna, 2003, vol. 13, n. 1, pp. 21 – 41.

Internet sources

http://www.termium.com

http://www.nkp.cz/o_knihovnach/Slovnik/index.htm

http://www.yourdictionary.com

http://www.cilf.org

http://www.termisti.refer.org

http://www.erudit.org/revue/meta/

http://www.rint.org

http://www.infoterm.info/index.php

http://linux.termnet.org/

http://europa.eu/eurovoc/

Beyond Syntactic Valence: FrameNet Markup of Example Sentences in a Slovenian-German Online Dictionary

Birte Lönneker-Rodman

International Computer Science Institute, 1947 Center Street, Suite 600
Berkeley, CA 94704, USA
loenneke@ICSI.berkeley.edu

Abstract. This contribution focuses on the representation of syntactic and semantic valence in a bilingual electronic dictionary, *Online SLO-DE-SLO*. To overcome valence representation problems caused by cross-lingual divergences, FrameNet-style annotation is applied to usage examples. Results and possible future directions are discussed.

1 Introduction

Many factors must be considered when compiling electronic dictionaries (de Schryver 2003). This contribution focuses on the representation of syntactic and semantic valence in a bilingual dictionary, using examples drawn from an existing online dictionary. The *Online SLO-DE-SLO* dictionary is presented briefly in Section 2. Section 3 summarizes the FrameNet-approach to corpus-based semantic annotation. FrameNet-style annotation is applied to *Online SLO-DE-SLO* usage examples (Section 4), resolving valence representation problems caused by cross-lingual divergences (Dorr 1994). Section 5 discusses the results and indicates possible future directions.

2 A Slovenian-German online dictionary

Online SLO-DE-SLO is a Slovenian-German/German-Slovenian online dictionary; an interface in German and Slovenian is accessible via the Web.[1] Earlier stages of dictionary development and evaluation have been presented in Lönneker and Jakopin (2003), Jakopin and Lönneker (2004), and Lönneker and Rozman (2004). The motivation behind *Online SLO-DE-SLO* is to create a lexical resource useful to both human users and Natural Language Processing. Recent additions include a full-form morphological generator for Slovenian, and an improved dictionary structure: Translated example phrases and example sentences have been separated from the main component of the dictionary, which is now restricted to correspondences between single words and multi-words.

[1] http://webapp.rrz.uni-hamburg.de/~slowenisch/ [July 30, 2007].

As of July 2007, the dictionary contains over 8,800 word and multi-word correspondences, as well as 2,150 bilingual usage examples. The dictionary records on average more than 80,000 requests per month. The subsequent subsections are devoted to two aspects of *Online SLO-DE-SLO* central to the focus of the paper: valence information (2.1) and example sentences (2.2).

2.1 Valence information in online SLO-DE-SLO

Online SLO-DE-SLO provides grammatical information for both languages, such as part of speech for each single word lemma or aspect information for Slovenian verbs. Syntactic valence patterns are also given; these apply to words which can function as syntactic governors (head words of phrases). Generally, syntactic valence information is available for verbs, but it can also appear with other parts of speech.

Example (1) contains *Online SLO-DE-SLO* grammatical information for the German and Slovenian verbs corresponding to English *(to) introduce [x to y]*.

(1) [Slov.] predstaviti (perf) V [+ DAT.] [+ AKK.]
[Ge.] vorstellen V [+ DAT.] [+ AKK.]

In the example, the first item in each line is the base form of the verb. Information on verb aspect is relevant for Slovenian only; *(perf)* indicates perfective aspect. This is followed by an abbreviation indicating part of speech; *V* stands for verb. Finally, arguments are represented by their grammatical case and delimited by square brackets. The order of the arguments corresponds to their canonical order in an unmarked declarative sentence. Only arguments typically following the verb are represented; the (unrepresented) subject is implicitly assumed to take nominative case. The verbs in (1) thus take two non-subject arguments, the first in the dative case (*[+ DAT.]*, corresponding to the indirect object in English) and the second in the accusative case (*[+ AKK.]*, corresponding to the English direct object).

Currently, each syntactic valence is a separate dictionary entry, whether or not it results in different cross-linguistic equivalents. To illustrate, the Slovenien verb *potresti* – 'strew' has several valence patterns, among them (2) and (3), which translate differently into German. – The display convention of dependent prepositional phrases is "preposition plus case".

(2) [Slov.] potresti (perf) V [+ AKK.] [s/z + INSTR.]
[Ge.] bestreuen V [+ AKK.] [mit + DAT.]
'strew [sth. with sth.]'

(3) [Slov.] potresti (perf) V [+ AKK.] [po + LOK.]
[Ge.] streuen V [+ AKK.] [über + AKK.]
'strew [sth. over sth.]'

Dictionary data is stored in a relational database. There is a separate table for valence information, with a pointer to the relevant lemma and with a number indicating the order of the dependent phrases. For example, the two valence patterns of *potresti* shown in (2) and (3) are represented as in Table 1.

PhraseType	Lemma_sl_Ref	PhraseOrder
+ AKK.	2	1
s/z + INSTR.	2	2
+ AKK.	3	1
po + LOK.	3	2

Table 1. Slovenian valence information in dictionary database

2.2 Example sentences in online SLO-DE-SLO

Syntactic valence information as introduced in 2.1 is valuable to advanced learners with appropriate linguistic knowledge. It is also intended to be useful to automatic systems which might process the bilingual data. However, the occasional user lacking sufficient background in linguistics would benefit from (additional) examples illustrating the usage of the words. For instance, the entry for the Slovenian verb *telefonirati* – '(to) phone [so.], to call [so.]' says that its first non-subject argument takes the dative case. The closest German translation equivalent is *anrufen*, which takes an accusative object, as in (4).

(4) [Slov.] telefonirati (perf, impf) V [+ DAT.]
[Ge.] anrufen V [+ AKK.]

Example (5) illustrates this difference in context.

(5) a. [Slov.] Če bom le mogla, **ti** bom telefonirala.
If will-I only can-PART, **you-DAT** will-I phone.

b. [Ge.] Wenn ich nur kann, werde ich **dich** anrufen.
If I only can, will I **you-ACC** call.

'If at all possible, I will call you.'

For *Online SLO-DE-SLO*, examples are considered particularly useful when valence patterns differ across languages. Example (5) illustrates a difference in grammatical case assigned to a particular argument, but cross-lingual divergences can be more complicated than that. *Thematic divergence* (Dorr 1994, pp. 607–609) is the repositioning of arguments with respect to a given verb or other lexical head. It can also be described as the realization of what is the subject in one language by a different grammatical function (e.g., direct object) in the other language. For instance, in its simplest valence pattern the Slovenian verb *zebsti* takes a zero subject and an accusative object, the object phrase indicating a living being that feels cold. With the corresponding German verb *frieren*, this living being is realized as subject, in the most common valence pattern. The dictionary represents this fact as in (6).

(6) [Slov.] zebsti (perf) V [+ AKK.]
[Ge.] frieren V

Such cases present a problem in the current *Online SLO-DE-SLO* representation of syntactic valence, because it relies on formal and semantic cross-linguistic equivalence of the subject. Therefore, for further illustration the dictionary includes translation equivalent example sentences, as in (7).

(7) a. [Slov.] Zebe me.
Freezes me-ACC.

b. [Ge.] Ich friere.
I freeze.

'I am cold.'

The correspondence between the Slovenian object and the German subject of sentences (7a) and (7b) is best captured in semantic terms. For English, information on syntactic and semantic valence has been collected in the lexical database of the FrameNet project (Fillmore *et al.* 2003, Ruppenhofer *et al.* 2006). After a brief overview of the main principles of FrameNet in Section 3, their possible applications in a bilingual dictionary such as *Online SLO-DE-SLO* will be shown in Section 4.

3 FrameNet

FrameNet is an on-line lexical resource for English based on frame semantics (Fillmore 1978; Fillmore *et al.* 2003). It represents each word sense or "lexical unit (LU)" as a unique combination of semantico-conceptual and morpho-syntactic information. Morpho-syntactic information in FrameNet comprises part of speech, word forms and the order of units within multi-word terms. Semantico-conceptual information of a lexical unit is provided through membership in a particular *semantic frame*, the background for understanding the lexical unit in context (3.1). FrameNet annotates occurrences of LUs with respect to their frame-semantic and syntactic behavior in example sentences retrieved from electronic corpora (3.2). A set of analyses of annotated examples taken together then shows the valence of a LU (3.3).[2]

3.1 The semantic frame

A frame is a conceptual structure that describes a particular type of situation, object, or event along with its participants and props, which are referred to as frame elements (FEs). Table 2 gives a very condensed overview of a frame and its elements.[3] Notably, each frame has a name (caption), a description in free text (top line of Table 2), and is associated with frame elements (subsequent lines of the table), which are given a name and a textual definition, usually referring back to the description of the frame as a whole.

[2] FrameNet is a very rich lexical resource and many of its features cannot be explained in this paper. More information on the FrameNet data model can be found in (Baker *et al.* 2003), (Ruppenhofer *et al.* 2006) and (Lönneker-Rodman 2007).

[3] For full definitions of the frames mentioned in this paper, please refer to the FrameNet website: `http://framenet.icsi.berkeley.edu/` [July 29, 2007].

This frame contains words describing physical experiences that can affect virtually any part of the body. The body part affected is almost always mentioned with these words.	
Frame Element	**Definition**
Body_part	This FE is the location on the body where the physical experience takes place [...].
Experiencer	The Experiencer is the being who has a physical experience on some part of his or her body, or internally.

Table 2. FrameNet Frame `Perception_body`

In spite of its conceptual nature, a frame cannot be defined without knowledge about the lexical units that *evoke* the frame. For example, in English, the verbs *ache* and *itch* (among others) evoke the `Perception_body` frame; the verbs are *members* of the frame.

3.2 Example sentences in FrameNet

Information provided by FrameNet would be very abstract if it were not supported by authentic example sentences where frame-evoking lexical units and frame elements are annotated. Color highlighting of each annotated constituent facilitates the recognition of its semantic role in the sentence. Frame-evoking words are displayed in white letters on black background. Highlighting of other constituents of the sentence makes reference to the colors defined for the respective frame elements of the frame. Examples (8) – (11) are taken from the FrameNet website and illustrate usages of the verb *itch*, evoking the `Perception_body` frame. Only the Experiencer frame element is realized in (8) with the noun phrase *he*; and only the Body_part FE is realized in (9) with the noun phrase *they* (i.e. the ears).

(8) Within hours, while he itched and writhed [...]

(9) The ears are thickened, [...] they itch and they hurt [...]

Most examples realize both of these frame elements, as in (10) and (11) below. Usually, syntactic constituents realizing different frame elements do not overlap. In the particular case of Examples (10) and (11), however, the frame element Experiencer is incorporated within the Body_part constituent. To represent this, FrameNet uses second-layer annotation (Fillmore *et al.* 2003, p. 318; Ruppenhofer *et al.* 2006, p. 37), which allows an annotator to single out the relevant possessor phrase and annotate it twice, once for Body_part (within the larger phrase) and then for Experiencer.

(10) Leith's right hand started to itch again.
Leith's

(11) Is your head itching now?
your

Besides the frame-evoking and frame element information provided by the semantic markup of sentence constituents, syntactic information about these constituents is given, including phrase type and grammatical function. More information on syntax in FrameNet can be found in (Fillmore *et al.* 2003) or (Ruppenhofer *et al.* 2006).

3.3 Valence information in FrameNet

In FrameNet, valence information is not stored statically as a feature of a lexical unit. Instead, it is derived dynamically from the annotation of sentences with respect to that lexical unit. The information characterizes both the semantic and syntactic combinatorial profile of the LU. The semantic profile is given in terms of number and specific combination of frame elements occurring with the LU, and the syntactic profile provides the phrase types and grammatical functions of the annotated constituents. Ideally, about 20 sentences are annotated for each LU, and the annotation summaries quantify each valence pattern in terms of its frequency within this set of annotations.

For example, a valence pattern table derived from only the four annotations of Examples (8) to (11) above is given in Table 3. The abbreviation *NP* stands for noun phrase and is a phrase type label; *Ext* stands for the grammatical function "external argument", or subject.

Number Annotated	**Patterns**	
1 TOTAL	Body_part	
(1)	NP	
	Ext	
2 TOTAL	Body_part	Experiencer
(2)	NP	2nd layer
	Ext	–
1 TOTAL	Experiencer	
(1)	NP	
	Ext	

Table 3. FrameNet-style valence information for *itch.v*

As syntactic information is already – at least partly – covered in *Online SLO-DE-SLO*, the following discussion will concentrate on the possible contribution of FrameNet-annotation to the representation of semantic valence and cross-lingual correspondences thereof.

4 Putting the pieces together

This section discusses how FrameNet annotation could be integrated into the bilingual dictionary *Online SLO-DE-SLO*. The focus of the case studies (4.1 to 4.3) is the usefulness of such markup to human users, especially in difficult (i.e. divergent) cases. A discussion will follow in Section 5.

4.1 General concept

Given that *Online SLO-DE-SLO* already includes syntactic valence information, the purpose of the semantic annotation of example sentences is to supplement that with information on correspondences between frame evoking words and their frame elements. The idea that frame element assignments can be visualized by colored highlighting is taken from FrameNet. When applied to translation-equivalent examples in two languages, the colors identify cross-linguistically corresponding portions of the sentence. This can be seen in Table 4illustrating the usage of Slovenian *telefonirati*; for glosses and translations, see Example (5) above. – Annotating bound morphemes such as the inflectional ending *-m* in the Slovenian auxiliary *bom* – 'I will' is not standard in FrameNet, but facilitates the display of cross-lingual equivalences.

Slovenian	German
Če bom le mogla,	Wenn ich nur kann,
ti bo m telefonirala .	werde ich dich anrufen .

Table 4. Example sentence for Slovenian *telefonirati*, with FrameNet markup

The actual semantics of the color-highlighted annotations is given in an abbreviated description of the relevant FrameNet frame, to appear directly above or below the example(s). Table 5 displays the necessary information from the `Contacting` frame, of which Slovenian *telefonirati* and German *anrufen* would be members. Only the first part of the definition and the relevant subset of frame elements are displayed. The complete frame description could be connected via a hyperlink to the FrameNet website.

4.2 Illustrating thematic divergence

In the language pair Slovenian-German, thematic divergence (cf. 2.2 above) often involves a zero subject in Slovenian, or the expletive (semantically empty) sub-

A Communicator [...] directs a Communication to an Addressee [...]	
Frame Element	**Definition**
Addressee	The person that receives the message from the Communicator.
Communicator	The person who uses language in the written or spoken modality to convey a message to another person.

Table 5. FrameNet Frame `Contacting`

ject *es* 'it' in German. Some of the verbs with which this phenomenon occurs belong to the `Perception_body` frame, a short description of which has been discussed in 3.1 above. By making reference to the frame, Example (7) discussed in 2.2, can be annotated in a straight-forward fashion, as in Table 6.

Slovenian	German
Zebe me.	Ich friere.

Table 6. Example sentence for Slovenian *zebsti*, with FrameNet markup

The Slovenian verb *srbeti* and its German counterpart *jucken* – '(to) itch' also evoke the `Perception_body` frame. Because of thematic divergences, three different valence patterns of German *jucken* have been defined as equivalents of one single syntactic valence in Slovenian, shown in (12).

(12) a. [Slov.] srbeti V [+ AKK.]
b. [Ge.] jucken V
[Ge.] jucken V [+ AKK.]
[Ge.] jucken V [+ AKK.] [an + DAT.]

Given the `Perception_body` frame, semantic valence patterns explain the cross-linguistic differences. Slovenian realizes the Body_part FE as subject of *srbeti*, and Experiencer as direct (accusative) object, as in (13a). In German, information about the Experiencer can be incorporated as a possessive determiner in the constituent expressing Body_Part (13b), precluding any other syntactic argument besides the subject. Experiencer information can also be expressed by a noun phrase in the accusative, similarly to Slovenian. In this case, the Body_part FE might be realized by a prepositional phrase rather than as subject, as in (14b).[4]

[4] Both example sentences come from the German DWDS corpus (`http://www.dwds.de` [24 July, 2007]) and have been translated into Slovenian by the author.

(13) a. [Slov.] Koža na glavi **ga** je srbela.
The-skin on head **him-ACC** has itched.

b. [Ge.] **Seine** Kopfhaut juckte.
His scalp itched.
'His scalp was itching.'

(14) a. [Slov.] Ko ga je srbela **noga**, ga je praskal eden izmed gospodov.
When him-ACC has itched **the-leg**, him has scratched one of the-sirs.

b. [Ge.] Als es ihn **am Bein** juckte, kratzte ihn einer der Herren.
When it him-ACC **at-the leg** itched, scratched him one of-the sirs.
'When his leg was itching, one of the sirs scratched him.'

Table 7 shows how the semantic correspondences in these thematically divergent sentences can be made explicit by frame semantic annotation, by using second-layer annotation (cf. 3.2 above) in German.

Slovenian	German
Koža na glavi ga je srbela.	Seine Kopfhaut juckte. Seine
Ko ga je srbela noga, ga je praskal eden izmed gospodov.	Als *es* ihn am Bein juckte, kratzte ihn einer der Herren.

Table 7. Example sentences for German *jucken*, with FrameNet markup

4.3 Illustrating categorial divergence

In a bilingual dictionary, the canonical word equivalence is usually between lexical items of the same part of speech, as in (15) and (16). However, due to what Dorr (1994, pp. 615–616) calls *categorial divergence*, semantic correspondence might sometimes be established between words of cross-linguistically different parts of speech.

(15) [Slov.] dolgčas N (m)
[Ge.] Langeweile N (f)
'boredom'

(16) [Slov.] dolgočasen ADJ
[Ge.] langweilig ADJ
'boring'

FrameNet frames can host lexical units of different part of speech (e.g., both verbs and nouns). FrameNet-style annotation is thus suitable for illustrating categorial divergence because the translation equivalent sentences in which this phenomenon appears still evoke the same frame (also Padó 2007, p. 42). In a bilingual dictionary, frame semantic markup can illustrate cross-linguistic correspondences between frame-evoking words of any part of speech and between their frame elements, as illustrated in Table 8. It shows that specific translations of the German adjective *langweilig* into Slovenian do not always reflect the canonical equivalence given in the word table; instead of a form of the corresponding adjective (first line), a noun can appear in Slovenian. – Annotation refers to the `Subject_stimulus` frame, a summary of which is given in Table 9.

Slovenian	German
Es war furchtbar langweilig.	Bilo je obupno dolgočasno.
Mir ist langweilig.	Dolgčas mi je.
Das wäre langweilig!	To bi bil dolgčas!

Table 8. Example sentences for German *langweilig*, with FrameNet markup

In this frame either a Stimulus brings about a particular emotion or experience in the Experiencer or saliently fails to bring about a particular experience. [...]	
Frame Element	**Definition**
Stimulus	The Stimulus is the object or event which brings about the emotion in the Experiencer.
Experiencer	The Experiencer experiences the emotion brought about by the Stimulus.

Table 9. FrameNet Frame `Subject_stimulus`

5 Discussion and outlook

The case studies presented in Section 4 show that FrameNet-style markup of example sentences provides a semantic complement to the syntactic valence patterns already incorporated in *Online SLO-DE-SLO*. Many cross-linguistic divergences which are difficult to capture at the syntactic level alone can be adequately described by frame semantic annotation. Still, the incorporation of semantic annotation into the dictionary database remains a long-term goal rather than an immediate project. Some of the issues that first must be resolved are related to dictionary structure (5.1), others to the availability of resources (5.2).

5.1 Dictionary structure

If *Online SLO-DE-SLO* is to be enhanced with semantic annotations of examples, the question of storing annotation information must be addressed. A possible approach would be to integrate the annotation of example sentences into the example table; for instance, by augmenting the text with in-line XML markup. For several reasons, however, annotations should be represented separately and linked to the examples. First of all, it should be possible to turn off the display of semantic annotation. Second, each annotation is done with respect to one particular frame evoking word within the sentence. Now, the same sentence might be displayed as an illustration not only of this particular word, but also of all the other words it contains, when the user actually queries the dictionary. Automatic display of semantic markup targeting a word that does not correspond to the user query might be confusing. While a sentence may contain several frame-evoking words in respect to which annotation would be provided, the user will be interested in at most one of them at a time. This makes it necessary to index each annotation by the target word, which can then be accessed and evaluated by the display function, to ensure that only relevant annotations are shown.

Finally, each annotation must be provided with a reference to the evoked frame, and minimal information about this frame should be held in the dictionary database. This includes short frame definitions, information about the colors for presenting the markup, and the URL of the original FrameNet definition.

5.2 Multilingual FrameNets

To ensure internal consistency and consistency with existing FrameNet resources, the example annotations should ideally make reference to previously established monolingual FrameNets in German and Slovenian. In the case studies presented, the appropriate FrameNet frame for each illustrated word had first to be found before a sentence could be annotated. However, proceeding example by example does not make sure that a frame is interpreted consistently by an annotator, and does not facilitate verifying whether the frame is actually suitable for the language in all respects.

In fact, a frame might need language specific modifications even if some individual sentences seem to fit it (see e.g. Burchardt *et al.* 2006, Lönneker-Rodman 2007). The overall picture of the semantics of a given frame and its lexical units can only be achieved by investigating a large monolingual corpus, following the empirical approach adopted by FrameNet. For German, a substantial step in this direction has been made by the SALSA project (Burchardt *et al.* 2006). Their data, once released, should be regarded as a reference when adding semantic annotations to *Online SLO-DE-SLO*. For Slovenian, no such resource has been developed yet and it is unlikely that one will be available in the near future. Annotation within *Online SLO-DE-SLO* is a very first step. Still, its main merit will consist in pointing out methodological problems and possible solutions for this language as well as bilingual issues, rather than in substantial coverage.

The case studies have been presented against the background assumption that a FrameNet frame providing semantics for the selected sentences exists. This is not always the case. FrameNet itself does not yet cover the entire vocabulary of English. At the time of writing, it was impossible to find a frame for the Introducing situation corresponding to Example (1). Incomplete coverage of English FrameNet and possible work-arounds for FrameNet-style annotation in other languages have been pointed out previously by Burchardt *et al.* (2006, p. 971).

Finally, in spite of the high level of universality of FrameNet frames, not all frames based on English data are suitable for other languages. To illustrate, it is not clear whether the `Perception_body` frame is actually suitable for Slov. *zebsti* and Ge. *frieren* – '(to) be cold' (see 4.2 above). The frame definition says that the body part affected is almost always mentioned with the words in the frame, which is not the case for the Slovenian and German verbs.

Acknowledgments

I would like to thank members of the AI group and FrameNet at the International Computer Science Institute for their support and hospitality. My research at ICSI is supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

References

Baker, C.F., Fillmore, C.J., Cronin, B.: The structure of the FrameNet database. International Journal of Lexicography **16**(3) (2003) 281–296

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., Pinkal, M.: The SALSA corpus: a German corpus resource for lexical semantics. In: Proceedings of LREC 2006, Genoa, Italy (2006) 969–974

Dorr, B.J.: Machine Translation divergences: A formal description and proposed solution. Computational Linguistics **20**(4) (1994) 597–633

Fillmore, C.J.: On the organization of semantic information in the lexicon. In: Papers from the Parasession on the Lexicon. Chicago Linguistics Society, Chicago, IL (1978) 148–173

Fillmore, C.J., Johnson, C.R., Petruck, M.R.L.: Background to FrameNet. International Journal of Lexicography **16**(3) (2003) 235–250

Fillmore, C.J., Petruck, M.R.L., Ruppenhofer, J., Wright, A.: FrameNet in action: The case of attaching. International Journal of Lexicography **16**(3) (2003) 297–332

Jakopin, P., Lönneker, B.: Query-driven dictionary enhancement. In: Proceedings of the Eleventh EURALEX International Congress, Lorient, France (2004) 273–284

Lönneker, B., Jakopin, P.: Contents and evaluation of the first German-Slovenian online dictionary. In: Proceedings of EACL 2003, Conference Companion, Budapest, Hungary, ACL (2003) 119–122

Lönneker, B., Rozman, K.: Online SLO-DE-SLO: Spletni slovensko-nemški in nemško-slovenski slovar. In: Proceedings of the Fourth Language Technologies Conference, Ljubljana, Slovenia (2004) 56–63

Lönneker-Rodman, B.: Multilinguality and FrameNet. Technical Report TR-07-001, International Computer Science Institute, Berkeley, CA (March 2007)
Padó, S.: Translational equivalence and cross-lingual parallelism: The case of FrameNet frames. In: Proceedings of the NODALIDA Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages, Tartu, Estonia (2007) 39–46
Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Scheffczyk, J.: FrameNet II: Extended theory and practice (August 2006)
de Schryver, G.M.: Lexicographers' dreams in the electronic-dictionary age. International Journal of Lexicography **2**(16) (2003) 143–199

Automatic Word Clustering in Russian Texts Based on Latent Semantic Analysis

Olga Mitrofanova, Polina Panicheva and Vyacheslav Savitsky

Department of Mathematical Linguistics
Faculty of Philology, St.-Petersburg State University
Universitetskaya emb. 11, 199034 St.-Petersburg, Russia
alkonost-om@yandex.ru, ppolin@yandex.ru, v_savitsky@yahoo.com

Abstract. The paper deals with development and application of automatic word clustering (AWC) tool aimed at processing Russian texts of various types, which should satisfy the requirements of flexibility and compatibility with other linguistic resources. The construction of AWC tool requires computer implementation of latent semantic analysis (LSA) combined with clustering algorithms. To meet the need, Python-based software has been developed. Major procedures performed by AWC tool are segmentation of input texts and context analysis, co-occurence matrix construction, agglomerative and *K*-means clustering. Special attention is drawn to experimental results on clustering words in raw texts with changing parameters.

1 Introduction

Recent advances in development of linguistic research tools encourage solution of the problems dealing with semantic data extraction from text corpora. One of the most relevant issues is automatic word clustering (AWC) – a procedure which provides data on hierarchical structure of the lexicon which are indispensable in construction of NLP-oriented lexicographic modules (dictionaries, thesauri and ontologies), word sense disambiguation, automatic text indexing, document clustering, information retrieval, etc. AWC procedures are widely used in various NLP systems, e.g.: COALS[1], InfoMap[2], Google Sets[3], DSM[4], SenseClusters[5], etc. However, some few AWC modules aimed at Russian corpora processing have been developed for research purposes (e.g.: [1], [2], [3], etc.) and very few are available on the Web (e.g.: SemClass[6]); thereby, the necessity of constructing such devices seems quite evident.

1 http://dlt4.mit.edu/~dr/COALS/
2 http://infomap.stanford.edu/
3 http://labs.google.com/sets
4 http://clg.wlv.ac.uk/demos/similarity/
5 http://senseclusters.sourceforge.net/
6 http://corpus.leeds.ac.uk/semclass/

The aim of the discussed project is elaboration and application of AWC tool for Russian. It is claimed that the proposed tool should allow processing texts of various types and size (raw and morphologically tagged texts, monolingual and multilingual parallel texts, texts of various genres, small and large corpora, etc.), it should be flexible and compatible with other linguistic resources.

The project is implemented in stages: first, the environment for processing raw texts is constructed; second, functions which enable operating on morphologically tagged texts are introduced. The paper presents results of AWC tool development achieved so far with regard to computer implementation, as well as experimental data on AWC procedures performed on raw monolingual and multilingual parallel texts.

2 AWC techniques

It is implied that AWC may be successfully performed on the basis of co-occurrence data obtained from corpora. Thus, AWC procedure requires realization of latent semantic analysis (LSA) (e.g.: [4], [5], [6]) and clustering algorithms (e.g.: [7], [8], [9]).

From a linguistic point of view, LSA is based on the possibility of detecting semantic similarity of words by comparing their syntagmatic properties (co-occurrence or distribution analysis). From a technical standpoint, LSA involves construction of vector-space models for processed texts; it means that the sets of contexts for each word are represented as distribution vectors in N-dimensional space. Varieties of LSA which take into account probable complexity of a corpus are discussed in [10], [11], etc.

It is possible to evaluate semantic similarity of words by measuring distances between their distribution representations. Such representations may be defined in terms of word distributions in contexts for raw texts or with respect to POS characteristics of context elements, i.e. in terms of POS tag distributions for tagged texts. Numerous metrics are used for the given purpose, e.g. Euclidean measure, Hamming measure, Chebyshev measure, cosine measure, etc. (cf. list of measures implemented in DSM). The selection of metrics often depends on qualitative parameters of processed texts. In our case, Euclidean measure was chosen as a basic metric. Results of measuring semantic distances are applied in clustering: words having similar distribution representations as a rule reveal similarity of meaning and should be included into the same cluster.

General approaches to clustering are exposed in hierarchical (agglomerative, divisive), partitioning (K-means, K-medoid, etc.), hybrid algorithms. Certain linguistic tasks require application of special clustering techniques, e.g. CBC [12], MajorClust [13]. The choice of a particular algorithm is determined by experimental conditions (corpora size, required speed of clustering, constraints for the number of resulting clusters, etc.). At the first stage of the project preference was given to basic clustering algorithms (agglomerative and K-means).

Data extracted from texts through AWC procedure admit cogent linguistic interpretation.

3 Computer implementation of AWC

Python-based AWC software developed and adjusted within the project framework maintains a set of conjoined modules performing text preprocessing, agglomerative clustering, K-means clustering.

Such parameters as names of input files (processed texts and sets of words subjected to clustering), context window size [$\pm s$], weight assignment for context items (yes/no), distance metric, clustering technique, ultimate number of clusters (C), etc. are determined by users.

The first module provides text preprocessing. Context segmentation is carried out in accordance with a particular context window size. Automatic weight assignment may be done for lexical items taking into account their positions in contexts. The given module is responsible for such operations as forming distribution representations of words, measuring semantic distances, building co-occurrence matrix. The second and the third modules support agglomerative and K-means clustering respectively. An output file contains co-occurrence data and clustering results.

4 Experimental results on AWC with various parameters

In order to determine research potential of AWC tool and to evaluate its effectiveness, a series of experiments on clustering words in raw texts was fulfilled, namely:

- automatic clustering of Russian most frequent and polysemous verbs in the experimental corpus of verbal contexts;
- automatic clustering of descriptors in scientific texts included into the corpus on Corpus Linguistics;
- automatic clustering of words in parallel texts: the original English text of the fairy story “Animal Farm” by G. Orwell and its translation into Russian.

4.1 Automatic clustering of verbs in experimental corpus

AWC proves to be quite productive with respect to distinguishing verbs of different semantic classes. Data on contextual neighbours of verbs and on verbal valency frames extracted from corpora play a crucial role in multilevel text analysis, therefore AWC tool was employed in trial processing of the experimental corpus of verbal contexts. The given corpus containing over

100 000 tokens was formed on the basis of a large Russian corpus *Bokr'onok* built in St.-Petersburg State University [1], [14]. Both raw and morphologically tagged versions of the experimental corpus are accessible.

Trial AWC procedures gave promising results. Major clustering parameters are as follows: context window size $s = \pm 5$, weight assignment for context items – *yes*, distance metric – Euclidean measure, clustering technique – agglomerative.

Experimental procedure was carried out for a set of the most frequent and polysemous verbs belonging to different semantic classes: verbs of intellectual activity: *dumať* (*think*), *ponimať* (*understand*), etc.; verbs of perception: *vidieť* (*see*), *smotreť* (*look*), etc.; verbs of transmission: *brať* (*take*), *dať* (*give*), etc.; verbs of functioning: *delať* (*do*), *rabotať* (*work*), etc.; verbs of management: *deržať* (*hold*), *brosať* (*throw*), etc.; verbs of movement: *idti* (*walk*), *jehať* (*drive*), etc.; verbs of location: *stojať* (*stand*), *ležať* (*lay*), etc.

Clustering of the given verbs was performed successfully, e.g.:

[*idti* (*walk*), *jehať* (*drive*) [*vidieť* (*see*), *smotreť* (*look*)]];
[*idti* (*walk*), *jehať* (*drive*) [*delať* (*do*), *rabotať* (*work*)]];
[*brať* (*take*), *dať* (*give*) [*vidieť* (*see*), *smotreť* (*look*)]];
[*deržať* (*hold*), *brosať* (*throw*) [*dumať* (*think*), *ponimať* (*understand*)]];
[*stojať* (*stand*), *ležať* (*lay*) [*dumať* (*think*), *ponimať* (*understand*)]].

Data on semantic distances for the given verbs seem to be reliable. Fluctuation of distance values indicates that diffrences in verb distributions are more or less significant.

The verbs belonging to the same class show similarity of distributions (their distance values are low), e.g.:

D (*dumať* (*think*), *ponimať* (*understand*)) = 0,107;
D (*delať* (*do*), *rabotať* (*work*)) = 0,117.

The verbs belonging to different classes reveal diverse distributions (their distance values are high), e.g.:

D (*ponimať* (*understand*), *ležať* (*lay*)) = 0,152;
D (*vidieť* (*see*), *idti* (*walk*)) = 0,151.

Difference in distributions is also registered for verbs belonging to the same semantic class but being in contrastive relations, e.g.:

D (*brať* (*take*), *dať* (*give*)) = 0,131.

However, such verbs are clustered correctly, so that the difference of their distributions seems to be less significant than the difference of distributions for verbs representing separate semantic classes.

In addition to clustering results achieved for raw texts, experiments on classification of verbal contexts extracted from tagged texts were also fulfilled. In this case a separate software module on similarity measurement was involved: it allows to compare POS tag distributions for a set of lexical items, cosine measure *Cos* ($0 \leq Cos \leq 1$) being used to compute similarity indices

within n-tuples of lexical items. Triplets of verbs sharing common semantic features (e.g., *ponimať* (*understand*) – *znať* (*know*) – *uznavať* (*learn*); *videť* (*see*) – *vosprinimať* (*perceive*) – *voobražať* (*imagine*); etc.) were under examination. Random samples of 200 contexts for each verb were processed with context window size $s = \pm 10$. The data obtained at this stage have proved rather encouraging: e.g.,

Cos (*ponimať* (*understand*), *znať* (*know*)) = 0,93;
Cos (*ponimať* (*understand*), *uznavať* (*learn*) = 0,90;
Cos (*znať* (*know*), *uznavať* (*learn*)) = 0,86;
Cos (*videť* (*see*), *vosprinimať* (*perceive*)) = 0,83;
Cos (*videť* (*see*), *voobražať* (*imagine*)) = 0,89;
Cos (*vosprinimať* (*perceive*), *voobražať* (*imagine*)) = 0,88.

Similarity indices calculated within verb triplets reflect differences in the structure of their lexical meaning and valency frames (although those differences may be rather subtle). Experimental data conform to the intuitive judgements made by native speakers of Russian. Relatively large dispersion of obtained *Cos* values supports the idea that the given approach can be efficiently used in disambiguation procedures.

4.2 Automatic clustering of descriptors in scientific texts

AWC shows considerable promise in processing terminological items and domain-restricted texts. In such cases clustering data contribute much to adequate domain modelling and ensure development of lexicographic and ontological systems. Linguistic resource involved in the experiment is the corpus on Corpus Linguistics being developed in St.-Petersburg State University and Institute of Linguistic Studies, RAS. The corpus contains texts of research papers in Russian [15], [16], [17], [18]; it includes over 105 articles (about 175 000 tokens) and abstracts (about 25 000 tokens). Each text of the corpus is preprocessed (special tags for tables, images, formulae, links, numbers, non-Russian text fragments, etc. are introduced); further the texts are supplied with metadata which include bibliographic passport and a set of 10 relevant descriptors (key words) indicating the topic of the paper. E.g., text № 2002_72_79 gets such a set of descriptors: [*arhiv* (*archive*), *bank* (*bank*), *dannyje* (*data*), *korpus* (*corpus*), *massiv* (*array*), *poisk* (*retrieval*), *razmetka* (*annotation*), *tekst* (*text*), *format* (*format*), *češskij* (*Czech*)], etc.

Major clustering parameters are as follows: context window size $s = \pm 5$, weight assignment for context items – *yes*, distance metric – Euclidean measure, clustering techniques – agglomerative and K-means, ultimate number of clusters $C = 3, 5, 7, 9$. It is worth noting that the results furnished by

agglomerative and *K*-means clustering differ distinctly with regard to central cluster size and filling, agglomerative clustering seems to be preferable, e.g.:

text № 2002_72_79, C = 5, agglomerative clustering:

> [*arhiv* (*archive*), *bank* (*bank*), *massiv* (*array*), *format* (*format*) [*razmetka* (*annotation*) [*češskij* (*Czech*) [*poisk* (*retrieval*) [[*tekst* (*text*), *korpus* (*corpus*)] *dannyje* (*data*)]]]]];

text № 2002_72_79, *C* = 5, *K*-means clustering:

> [[*arhiv* (*archive*)] [*bank* (*bank*)] [*razmetka* (*annotation*)] [*dannyje* (*data*), *korpus* (*corpus*), *poisk* (*retrieval*), *tekst* (*text*), *format* (*format*), *češskij* (*Czech*)] (*massiv* (*array*)]].

AWC procedure provided similar descriptions for other texts from the corpus, e.g.:

text № 2002_27_39:

> [*massiv* (*array*), *baza* (*base*), *dannyje* (*data*) [[*perevodčeskaja* (*translation*), *pam'at'* (*memory*)] [*sistema* (*system*) [*tekst* (*text*), *perevod* (*translation*)]]] [*korpus* (*corpus*), *parallel'nyj* (*parallel*)]];

text № 2006_5_15:

> [*poisk* (*search*), *internet* (*internet*), *zapros* (*query*), *pol'zovatel'* (*user*) [*servis* (*service*) [*častota* (*frequency*) [[*tekst* (*text*), *korpus* (*corpus*)] [*slovo* (*word*), *bigramma* (*bigram*)]]]]]]; etc.

It should be expected that resulting clusters obtained for each text of the corpus reveal semantic relations existing between lexical items. Sometimes those relations can be clearly defined as syntagmatic (e.g., *perevodčeskaja* (*translation*) – *pam'at'* (*memory*)) or paradigmatic (e.g., *massiv* (*array*) – *baza* (*base*)), but in most cases they merge (e.g., *tekst* (*text*) – *korpus* (*corpus*); *slovo* (*word*) – *bigramma* (*bigram*), etc.). Due to their intrinsic heterogeneity, linguistic relations between text items exposed in clusters may be smoothly converted into ontological relations between categories forming conceptual structure of the domain "Corpus Linguistics".

Alongside revealing semantic relations in the sets of terminological items, trial AWC procedure allows exposure of nuclear descriptors characteristic of Corpus Linguistics domain: *korpus* (*corpus*), *tekst* (*text*), *razmetka* (*annotation*), *poisk* (*retrieval*), *dannyje* (*data*), etc. Those descriptors are treated as exemplars of basic ontological categories expressing the essence of Corpus Linguistics.

Clustering results may be involved in the comparison of documents with partly coinciding sets of descriptors, e.g. text № 2002_72_79 and text № 2002_27_39 contain 4 similar descriptors, their sets being structured uniformly:

> [*massiv* (*array*) [*dannyje* (*data*) [*korpus* (*corpus*), *tekst* (*text*)]]];

at the same time, text № 2002_72_79 and text № 2006_16_24 are equally characterized by 5 common descriptors *korpus* (*corpus*), *tekst* (*text*), *format*

(*format*), *razmetka* (*annotation*), *poisk* (*retrieval*), although they are ordered in different ways:

text № 2002_72_79:

[*format* (*format*) [*razmetka* (*annotation*) [*poisk* (*retrieval*) [*tekst* (*text*), *korpus* (*corpus*)]]]];

text № 2006_16_24:

[*razmetka* (*annotation*) [[[*korpus* (*corpus*), *tekst* (*text*)] *format* (*format*)] [*poisk* (*retrieval*)]].

Given the texts which share common descriptors, similar clustering results prove adherence of the texts to the same topic, and conversely, differences in their clustering provide evidence on divergence of corresponding texts as regards their subject matter.

Various types of data provided by AWC procedure accompanied with expert descriptions (e.g., [19]) were used in the development of formal ontology on Corpus Linguistics maintained by ontoeditor Protégé [20]. Top hierarchy of ontological categories is as follows:

- *Domain «Corpus Linguistics»*
- Corpus data
  - Text corpus
  - Corpus type
    - Working with a corpus
      - Corpus developer
        - Data selection
        - Data digitalization
        - Annotation
        - Corpus-manager
          - User
            - Search
              - Query
                - Terminal string of symbols
                - Regular expression
                - Lemma
                - Tag
              - Result
                - Concordance
                - Context
                - Word index
                - Statistics

4.3 Automatic clustering of words in parallel texts

Processing multilingual parallel texts with the help of AWC tool enables us to compare relations between lexical items within semantic classes in contrasting languages and thus to verify adequacy of translation.

The original English text of the fairy-story "Animal Farm" by G. Orwell and its translation into Russian were involved in trial processing, Russian text size of about 24 000 tokens, English text size of about 30 000 tokens.

Clustering within a group of words denoting human beings, animals and birds (over 50 words occurring in both texts) was carried out with the following clustering parameters: context window size $s = \pm 5$, weight assignment for context items – *yes*, distance metric – Euclidean measure, clustering technique – agglomerative.

AWC tool proved to be quite efficient in distinguishing nouns within microgroups, e.g. in differentiating generic and specific names, in proper assignment of generic names, in revealing kinship hierarchy, e.g.:

[*voron* [*ovca*, *životnoje*]],	[*raven* [*sheep*, *animal*]];
[*cypľonok* [*košk*a, *životnoje*]],	[*chicken* [*cat*, *animal*]];
[*os'ol* [*utka*, *ptica*]],	[*donkey* [*duck*, *bird*]];
[*koza* [*uťata*, *ptica*]],	[*goat* [*ducklings*, *bird*]];
[*ptica* [*utka*, *golub'*]],	[*bird* [*duck*, *pigeon*]];
[*cypľonok* [*kurica*, *petuh*]],	[*chicken* [*hen*, *cockerel*]].

In most cases clustering results obtained for Russian and English items coincided, although some peculiar examples of divergent clustering output were found as well, e.g.:

[*kobyla (mare)* [*žereb'onok (foal)*, *lošaď(horse)*]],
[*foal* [*horse*, *mare*]];
[*ptica (bird)* [*čelovek (man)*, *životnoje (animal)*] *borov (boar)*],
[*bird* [*boar* [*man*, *animal*]]].

The difference in clustering indicates that relations within pairs "original text item vs. translation equivalent" are asymmetrical, possible explanations being different frequency of terms in the original and in translation as well as particular properties of the plot of the analyzed text.

It should be noted that at present AWC tool is successfully applied in the comparative study of 3 novels by V.O. Pelevin "Omon Ra", "Čapajev and Emptiness", "Generation P" (original Russian texts and English translations forming a parallel corpus, Russian part size of about 200 000 tokens and English part size of about 260 000 tokens) [21]. Special attention is drawn to words and word groups denoting culture-specific phenomena. It is claimed that clustering in sets of corresponding terms may be of great help in estimating stylistic and semantic proximity of original texts and translations.

5 Conclusions and work in progress

AWC tool developed for Russian ensures effective clustering of lexical items in raw texts. Trial procedures involving specialized software confirm reliability of implemented research techniques (LSA and clustering algorithms) forming actual basis of AWC tool. Experimental data show a wide range of possible applications of AWC in linguistic analysis and NLP systems.

Work in progress includes:

- software elaboration (user interface perfection, introduction of cluster visualization mode, implementation of additional distance metrics and clustering algorithms, e.g. MajorClust);
- linguistic experiments on AWC with regard to morphologically tagged texts of different size and various genres, mono- and multilingual texts (preliminary results of clustering Russian verbs in synsets taking into account POS tag distributions seem to be promising).

Further development of the project allows embedding of AWC tool into multi-level linguistic research environment equipped with a corpus manager and subsidiary modules.

Acknowledgments

The project is supported by the RF Presidential Grant № MK-9701.2006.6.

The authors would like to thank Prof. Viktor Zakharov, Prof. Mikhail Alexandrov, Prof. Irina Azarova, Prof. Larissa Beliaeva, Irina Larionova, Evguenia Malaia, Anna Marina, Anton Mukhin, Natalia Vinogradova for valuable advices and inspiring discussions concerning the project in question. The authors are grateful to the anonymous reviewer for helpful comments.

References

1. Azarova, I.V., Marina, A.S.: Avtomatizirovannaja klassifikacija kontekstov pri podgotovke dannyh dľa kompjuternogo tezaurusa RussNet. In: Kompjuternaja lingvistika i intellektuaľnyje tehnologii: Trudy meždunarodnoj konferencii “Dialog–2006”. Moscow (2006) 13–17
2. Baglej, S.G., Antonov, A.V., Meškov, V.S., Suhanov, A.V.: Klasterizacija dokumentov s ispoľzovanijem metainformacii. In: Kompjuternaja lingvistika i intellektuaľnyje tehnologii: Trudy meždunarodnoj konferencii “Dialog–2006”. Moscow (2006) 38–45
3. Križanovskij, A.A.: Avtomatizirovannoje postrojenije spiskov semantičeski blizkih slov na osnove rejtinga tekstov v korpuse s giperssylkami i kategorijami. In: Kompjuternaja lingvistika i intellektuaľnyje tehnologii: Trudy meždunarodnoj konferencii “Dialog–2006”. Moscow (2006) 297–302

4. Landauer, Th., Foltz, P.W., Laham, D.: Introduction to Latent Semantic Analysis. In: Discourse Processes 25 (1998) 259–284
5. Smrž, P., Rychlý, P.: Finding Semantically Related Words in Large Corpora. In: Text, Speech and Dialogue: Fourth International Conference TSD–2001. Lecture Notes in Artificial Intelligence, Vol. 2166. Springer-Verlag, Berlin Heidelberg New York (2001) 108–115
6. Pekar, V.: Linguistic Preprocessing for Distributional Classification of Words. In: Proceedings of the COLING–04 Workshop on Enhancing and Using Electronic Dictionaries. Geneva (2004) 15–21
7. Stein, B., Niggemann, O.: On the Nature of Structure and its Identification. In: Widmayer, P., Neyer, G., Eidenbenz, S. (eds.): Graph-Theoretic Concepts in Computer Science. Lecture Notes in Computer Science, Vol. 1665. Springer-Verlag, Berlin Heidelberg New York (1999) 122–134
8. Lin, D., Pantel, P.: Induction of Semantic Classes from Natural Language Text. In: Proceedings of ACM Conference on Knowledge Discovery and Data Mining KDD–01. San Francisco, CA (2001) 317–322
9. Shin, S.-I., Choi, K.-S.: Automatic Word Sense Clustering Using Collocation for Sense Adaptation. In: Proceedings of the Second International WordNet Conference GWC–2004. Brno, Czech Republic (2004) 320–325
10. Nekrestjanov, I.S.: Tematiko-orientirovannyje metody informacionnogo poiska. Dis. … kand. fiz.-mat. nauk. St.-Petersburg (2000)
11. Rohde, D.L., Gonnerman, L.M., Plaut, D.C.: An Improved Method for Deriving Word Meaning from Lexical Co-Occurrence.
`http://dlt4.mit.edu/~dr/COALS/Coals.pdf`
12. Pantel, P.: Clustering by Committee. Ph.D. Dissertation, Department of Computing Science, University of Alberta (2003)
`http://www.isi.edu/~pantel/Content/publications.htm`
13. Stein, B., Meyer zu Eissen, S.: Document Categorization with MajorClust. In: Proceedings of the 12th Workshop on Information Technology and Systems WITS–02. Barcelona, Spain (2002) 91–96
14. Azarova, I.V., Sinopaľnikova, A.A.: Ispolzovanije statistiko-kombinatornyh svojstv korpusa sovremennyh tekstov dľa formirovanija struktury kompjuternogo tezaurusa RussNet. In: Trudy meždunarodnoj konferencii "Korpusnaja lingvistika – 2004". St.-Petersburg (2004) 5–15
15. Doklady naučnoj konferencii "Korpusnaja lingvistika i lingvističeskije bazy dannyh – 2002". St.-Petersburg (2002)
16. Trudy meždunarodnoj konferencii "Korpusnaja lingvistika – 2004". St.-Petersburg (2004)
17. Trudy meždunarodnoj konferencii "MegaLing–2005": Prikladnaja lingvistika v poiske novyh putej. St.-Petersburg (2005)
18. Trudy meždunarodnoj konferencii "Korpusnaja lingvistika – 2006". St.-Petersburg (2006)
19. Zakharov, V.P.: Korpusnaja lingvistika. St.-Petersburg (2005)

20. Vinogradova, N.V., Mitrofanova, O.A., Paničeva, P.V.: Avtomatičeskaja klassifikacija terminov v russkojazyčnom korpuse tekstov po korpusnoj lingvistike. In: Trudy 9j Vserossijskoj konferencii "Elektronnyje biblioteki: perspektivnyje metody i technologii, elektronnyje kollekcii" RCDL–2007. Pereslavľ-Zalesskij, Russia (2007) in press
21. Beliaeva, L.N., Larionova, I.B.: Ispoľzovanije lingvističeskogo resursa dľa avtomatičeskoj klassifikacii leksiki kak osnovy dľa diagnostiki adekvatnosti perevoda (na materiale tekstov romanov V.O. Pelevina i ih perevodov na anglijskij jazyk). In: Tezisy meždunarodnoj konferencii "MegaLing–2007": Prikladnaja lingvistika v poiske novyh putej. Simpheropoľ, Ukraine (2007) in press

Corpus Analysis of Selectional Preferences in Russian

Olga Mitrofanova, Viktoria Belik and Vera Kadina

Department of Mathematical Linguistics
Faculty of Philology, St.-Petersburg State University
Universitetskaya emb. 11, 199034 St.-Petersburg, Russia
alkonost-om@yandex.ru, ogibbion14@pisem.net, veraiii@yandex.ru

Abstract. The paper presents results of a corpus-based study of selectional preferences in Russian word-groups. Much attention is drawn to the choice of context neighbours in verbal and nominal phrases, especially in noun-verb and adjective-noun word-groups. Research procedure requires analysis of co-occurrence data obtained from Russian texts. It is implied that selectional preferences of a lexical item may be defined through sorting its left/right neighbours in bigrams by *MI*-score values. Given an ordered set of neighbours for a lexical item, it is possible to induce its context patterns. Selectional preferences for frequent Russian lexemes in verbal and nominal groups are defined in terms of Optimality Theory rules. The rules are specified with respect to particular features of co-occurring lexical items.

1 Introduction

The principal goal of the project is the study of distributional properties of frequent Russian lexical items which implies fulfillment of the following tasks:

- collection and proper interpretation of co-occurrence data in Russian text corpora taking into account various parameters, e.g.: left/right positions of context neighbours, weights assigned to context elements according to their positions with regard to a query word, context window size, etc.;
- quantitative evaluation of relations of lexical items in collocations, exposure of relevant morphological, semantic and syntactic selectional preferences of the given words, formulation of selectional preferences in terms of Optimality Theory (OT) rules.

Theoretical basis of the study involves distributional approach to the meaning of lexical items. It is admitted that it is possible to identify the meaning of words while considering their syntagmatic features extracted from the adjacent context. Current linguistic research provides convincing evidence in favour of

the use of context analysis for determining word semantics (e.g., cf. papers presented at CONTEXT conference[1]).

The results of the study may be applied in several NLP spheres: e.g., in automatic word clustering (cf.: [1], [2]), word sense disambiguation (cf.: [3]), specification of valency frames and collocation models in lexical databases (cf: [4], [5]), contrastive research (cf.: [6]), etc.

2 Techniques of revealing selectional preferences

Selectional preferences of word X may be revealed given $N\{a, b, c,...\}$ – a set of its probable context neighbours. Set N is formed in course of processing of random contexts for word X extracted form a corpus. Ordering of set N is possible with regard to morphological, semantic, syntactic features of its elements. Quantitative criterion of preference of particular context neighbours for word X may be stated with regard to association measures *MI*, *Log*-likelihood, *t*-test, χ2-test, etc. (cf.: [7], [8]). We've chosen *MI*-score (mutual information coefficient) defined for bigrams of type yX / Xy where $y \ni N$ – left/right context neighbour of X. *MI*-score allows to estimate the force of association relations within collocations (e.g., between word X and its neighbour y in a bigram) as a correlation of collocation frequency *f(X*,y*)*, frequency of independent occurrencies of collocates *f(X), f(y)* and corpus size S:

$$MI = log_2 \frac{S \cdot f(X,y)}{f(X) \cdot f(y)}$$

MI-score allows to distinguish collocates with wide co-occurrence potential (they may be of high frequency but insignificant for X) and collocates tending to occur in combination with X, thus characterizing its selectional preferences (the higher is *MI*-score value for a bigram, the more neighbour y is preferable for word X).

It is proposed to formulate selectional preferences of words in terms of OT which seems to be of help in modelling competition of rules responsible for generation of linguistic units of various levels (from phonological up to semantic) (cf.: [9], [10]). Those rules can be ranged with regard to the degree of their significance. The rules which are of more importance get higher rank, their violation being more serious and resulting units being less correct, and vice versa. In other terms, enrichment of linguistic modelling with OT notions and instruments facilitates a shift from ideal language structures to optimal ones which may not be perfectly well-formed but must be acceptable. Thereby, hierarchy of priorities which controls the choice of context neighbours (revealing particular morphological, semantic, syntactic features) for word X in a context can be explained from OT standpoint. Traditional linguistics provides descriptions of rules responsible for dependency relations in word-

1 http://context-07.ruc.dk/CONTEXT07MainPage.html

groups, or rules for filling in valency frames, or rules of semantic agreement, etc. OT gives the possibility to introduce ordering into those classes of rules taking into account their natural priority. In case of co-occurrence study, it means ordering of context patterns of collocations based on their association force.

3 Linguistic resources and lexical data

Linguistic resources involved in the experiments include a lemmatized corpus of Russian texts based on M. Moshkov's electronic library (≈ 448 mln. tokens) and supplementary tools developed within AOT project[2]. Collocation extraction is performed with the help of AOT bigram search service[3] (cf.: [11]).

The study deals with the choice of the nearest left/right neighbours in verbal and nominal phrases, especially in noun-verb and adjective-noun word-groups. Russian verbs and adjectives of high frequency were chosen for experiments:

Verbs: *byť* (*be*), *vidieť* (*see*), *govoriť* (*speak*), *znať* (*know*), *jesť* (*eat*), *idti* (*walk*), *skazať* (*say*), *stať* (*become*), *hotieť* (*want*);

Adjectives: *blizkij* (*near*), *daľnij* (*remote*), *daľokij* (*far*), *dolgij* (*long*), *molodoj* (*young*), *pozdnij* (*late*), *sosednij* (*adjacent*), *starshyj* (*senoir*), *staryj* (*old*).

4 Experimental results

Bigram analysis required definition of *MI*-score threshold value for Russian and formulation of a decision rule which may help to reveal statistically relevant combinations of words.

It is argued that the threshold value $MI = 1$ holds true for Russian texts (due to the prevalence of syntactic groups and constructions with free order of elements), corresponding decision rule being as follows:

- if $MI > 1$, then a given combination of words is considered statistically relevant;
- if $MI \ni [0, 1]$, then a given combination of words is statistically neutral;
- if $MI < 0$, then words are in complementary distribution.

The given decision rule was derived from empirical evidence provided by corpus analysis. In course of bigram processing collocations characterized by $MI > 1$ were taken into account because relations between words in such combinations are nonrandom, the choice of collocates being predetermined by intrinsic

2 http://www.aot.ru/

3 http://aot.ru/demo/bigrams.html

linguistic factors. Collocations characterized by $MI \leq 1$ are beyond the scope of the study. Separate treatment was given to bigrams representing phraseological units: e.g., *reč + idti* (*reč iďot o … – the matter concerns …, MI* = 7,495) or *idti + vrazrez* (*to run counter to…, MI* = 9,466).

For each lexical item in question a set of bigrams with left/right context neighbours was formed. Those sets were ordered by *MI*-score value. Context neighbours extracted from bigrams were grouped according to their morphological features and common semantic components, where possible.

Fragments of bigram sets and groups of context neighbours are given below:

- bigrams containing the verb *idti* and its left context neighbours (*y + idti*):
  MI (*netoroplivo + idti*) = 4,668;
  MI (*smelo + idti*) = 4,467;
  MI (*pojezd + idti*) = 4,278;
  MI (*tropa + idti*) = 4,254;
  MI (*nado + idti*) = 3,154;
  MI (*rešiť + idti*) = 2,088;
  MI (*moč + idti*) = 1,770;
  etc.;
- groups of left context neighbours of *idti*:
  Adverbs: *toroplivo, netoroplivo, medlenno, bystro, dolgo,* etc. – common semantic features *speed* and *time*; *uverenno, smelo, uporno,* etc. – common semantic feature *emotional*; *krugom, sledom, vperedi, navstreču, dalee, daleko, kuda, kuda-to, nekuda,* etc. – common semantic feature *direction*; etc.
  Nouns: *karavan, pojezd, parohod,* etc. – common semantic feature *vehicle*; *doroga, tropa,* etc. – common semantic feature *path*; *dožď, sneg,* etc. – common semantic feature *natural phenomenon*; *boj, razgovor, torgovľa,* etc. – common semantic feature *complex dynamic event*; etc.
  Verbs and predicate adverbs: *moč, prodolžať, otkazyvať'a, molča, razrešiť, rešiť, sobirať'a, nado, pora,* etc.
- bigrams containing the verb *idti* and its right context neighbours (*idti + y*):
  MI (*idti + peškom*) = 7,240;
  MI (*idti + ožestočennyj*) = 5,475;
  MI (*idti + daľokij*) = 4,642;
  MI (*idti + spať*) = 3,860;
  MI (*idti + otdyhať*) = 3,670;
  MI (*idti + razgovor*) = 2,175;
  MI (*idti + volna*) = 1,218;
  etc.

– groups of right context neighbours of *idti*:

  Adverbs: *naprolom, naperekor,* etc. – common semantic feature *opposition*; *peškom, bosikom,* etc. – common semantic feature *means*; *r'adom, vperedi, sledom, navstreču, domoj, vdoľ, paralleľno, krugom, vper'od, pr'amo, napr'amik, naverh, mimo, s'uda, tuda,* etc. – common semantic feature *direction*; *normaľno, gladko,* etc. – common semantic feature *positive evaluation*; etc.

  Adjectives: (prepositional modifiers in dependent nominal phrases): *ožestočennyj, nepreryvnyj,* etc. – common semantic feature *intensity*; *medlennyj, bystryj,* etc. – common semantic feature *speed*; *daľokij,* etc.

  Verbs: *zavtrakať, guľať, spať, otdyhať, idti,* etc.

  Nouns: *dožď, par, sneg, volna,* etc. – common semantic feature *natural phenomenon*; *podgotovka, spor, boj, razgovor,* etc. – common semantic feature *complex dynamic event*; etc.

– bigrams containing the adjective *daľokij* and its left context neighbours (*y* + *daľokij*):

  MI (*beskonečno* + *daľokij*) = 6,631;
  MI (*donestis'* + *daľokij*) = 3,823;
  MI (*probiraťs'a* + *daľokij*) = 3,804;
  MI (*ves'ma* + *daľokij*) = 3,748;
  MI (*nemnogo* + *daľokij*) = 3,656;
  MI (*veršyna* + *daľokij*) = 2,279;
  MI (*čužoj* + *daľokij*) = 1,251;
  etc.

– groups of left context neighbours of *daľokij*:

  Adverbs: *beskonečno, strašno, ves'ma, nemnogo, stoľ, nastoľko, očen', sliškom, bolee, dovoľno,* etc. – common semantic feature *measure*; etc.

  Adjectives: *nevoobrazimyj, daľokij, čužoj, samyj, takoj, kakoj-nibuď,* etc.

  Verbs: *probiraťs'a, donestis', poslyšaťs'a, uslyšať,* etc.

  Nouns: *puť, doroga,* etc. – common semantic feature *path*; *strana, kraj, bereg, veršyna,* etc. – common semantic feature *place*; etc.

– bigrams containing the adjective *daľokij* and its right context neighbours (*daľokij* + *y*):

  MI (*daľokij* + *prošloje*) = 6,592;
  MI (*daľokij* + *predok*) = 6,181;
  MI (*daľokij* + *zvezda*) = 4,734;
  MI (*daľokij* + *okraina*) = 4,223;
  MI (*daľokij* + *galaktika*) = 4,111;

MI (*daľokij + daľokij*) = 3,334;
MI (*daľokij + južnyj*) = 2,291;
etc.

– groups of right context neighbours of *daľokij*:

Nouns: *prošloje, buduš'eje, predok, potomok, detstvo, junosť, drevnosť,* etc. – common semantic feature *time*; *rodina, daľ, kraj, okraina, strana, planeta, zvezda, galaktika, rasstojanije, gorizont, perspektiva, putešestvije, plavanije,* etc. – common semantic features *place* and *distance*; *raskat, grom, eho,* etc. – common semantic feature *sound*; etc.

Adjectives: *daľokij, prošlyj*, etc. – common semantic feature *remote*; *gornyj, severnyj, južnyj,* etc. – common semantic features *place* and *direction*; etc.

Processed data allowed to induce context patterns of type $POS + X$ / $X + POS$ for lexemes involved in the analysis. Those context patterns determining selectional preferences of words were ranked; ranking was performed with regard to *MI*-score values of context neighbours of corresponding groups, e.g.:

– selectional preferences for $X = idti$:

rank 1. $X + Adverb$
rank 2. $X + Adjective$
rank 3. $Adverb + X$
rank 4. $Noun + X$
rank 5. $X + Verb$
rank 6. $X + Noun$
rank 7. $Verb + X$

(3) selectional preferences for $X = daľokij$:

rank 1. $X + Noun$
rank 2. $Adverb + X$
rank 3. $X + Adjective$
rank 4. $Verb + X$
rank 5. $Adjective + X$

5 Conclusion

Data on selectional preferences of Russian frequent words obtained in course of experiments and exposed in context patterns admit more detailed linguistic interpretation; nevertheless, at this stage of the project they may be of great help in solution of several tasks dealing with automatic extraction of semantic information from Russian text corpora.

Acknowledgments

The project is supported by the RF Presidential Grant № MK-9701.2006.6.

The authors are grateful to Prof. Irina Azarova and Prof. Viktor Zakharov for inspiring discussions, and to the anonymous reviewer for helpful comments.

References

1. Pekar, V., Staab, S.: Word Classification Based on Combined Measures of Distributional and Semantic Similarity. In: Proceedings of European Chapter of ACL–03, Research Notes Session. Budapest, Hungary (2003) 147–150
2. Ananiadou, S., Spasic, I.: Using Automatically Learnt Verb Selectional Preferences for Classification of Biomedical Terms. In: Journal of Biomedical Informatics, Vol. 37 (2004) 483–497
3. Resnik, P. Selectional Preference and Sense Disambiguation. In: Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington, D.C., USA (1997) 52–57
4. Pekar, V.: Distributivnaja modeľ sočetajemostnyh ograničenij glagolov. In: Kompjuternaja lingvistika i intellektuaľnyje tehnologii: Trudy meždunarodnoj konferencii "Dialog–2004". Moscow (2004)
   `http://www.dialog-21.ru/Archive/2004/Pekar.htm`
5. Wagner, A.: Enriching a Lexical Semantic Net with Selectional Preferences by means of Statistical Corpus Analysis. In: Proceedings of the ECAI–2000 Workshop on Ontology Learning. Berlin, Germany (2000) 37–42
6. Agirre, E., Aldezabal, I., Pociello, E.: A Pilot Study of English Selectional Preferences and Their Cross-Lingual Compatibility with Basque. In: Text, Speech and Dialogue: 6th International Conference TSD–2003. Lecture Notes in Artificial Intelligence, Vol. 2807. Springer-Verlag, Berlin Heidelberg New York (2003) 12–19
7. Church, K.W., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. In: Computational Linguistics, Vol. 16 (1990) 22–29
8. Evert, S., Krenn, B.: Methods for the Qualitative Evaluation of Lexical Association Measures. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse, France (2001) 188–195
9. Blutner, R., de Hoop, H., Hendriks, P.: Optimal Communication. CSLI Lecture Notes, Vol. 177. Stanford, CSLI Publications (2006)
10. Hendriks, P., de Hoop, H.: Optimality Theoretic Semantics. In: Linguistics and Philosophy, Vol. 24 (2001) 1–32
11. Averin, A.N.: Razrabotka servisa poiska bigramm. In: Trudy meždunarodnoj konferencii "Korpusnaja lingvistika – 2006". St.-Petersburg (2006) 5–15

Lexterm, an Open Source Tool for Lexical Extraction

Joaquim Moré, Mercè Vázquez and Luis Villarejo

Linguistic Service – Universitat Oberta de Catalunya, Catalonia, Spain
{jmore,mvazquezga,lvillarejo}@uoc.edu

Abstract. This paper describes the development and exploitation of an open-source terminology-extraction tool named Lexterm. Lexterm was developed under the RESTAD project, whose aim was to introduce automatic processes in the translation of documents, especially those generated by the universities involved in the project. This tool was designed to be used not only by academic staff but also by any person interested in getting a raw bilingual glossary from a bulk of specific topic-domain documents. We will describe the tool, how it retrieves a list of term candidates with their translation equivalences from bilingual corpora, and how this list feeds machine translation and computer-assisted translation systems used in the generation flow of academic documents. We will also present how statistical methods are applied to get a list of term candidates as accurate as possible.

Keywords: *RESTAD project, terminology extraction, computer-assisted translation, machine translation, university document translation workflow.*

1 Introduction

Large institutions such as universities must tackle important translation requirements. This is especially evident in the case of Catalan universities set in a bilingual environment whose administrative and learning materials must be translated into Spanish for non-Catalan speakers and into Catalan in and out of Catalonia [1]. Apart from this, translations into English must also be performed not only for foreign students but also for the English versions of the university website or the English versions of university sponsored journals. The bulk of translated documents is so immense that the linguistic services of the Autonomous University of Barcelona (UAB), the University of Girona (UdG), the Technical University of Catalonia (UPC) and the Universitat Oberta de Catalunya (UOC)[1] decided to join forces in order to face their translation challenges. This joint effort was carried out under the RESTAD (Resources for computer-based translation applied to teaching) project, funded by the Ministry of Education and Universities of the Generalitat de Catalunya (Government of Catalonia) [2]. The RESTAD project consisted of developing

1 http://www.uoc.edu/web/eng/index.html

resources to improve the translation processes of administrative and learning materials at the universities involved in the project. These processes can be fully automatic [machine translations (MT) by a Catalan-Spanish, Spanish-Catalan system] or semi-automatic [computer-assisted translation (CAT) systems used by professional translators who translate into English].

One of the critical points in improving the translation of materials is the translation of terminology. Here 'terminology' should be understood in a broad sense. That is, the specific-domain lexical units (either one word or more than one word) which can be mistranslated because of two things. Firstly, these units are not stored in the bilingual lexicon of the machine translation system. Secondly, the human translator does not know the translation equivalent and has been unable to find the right denomination in the target language.

When translating a document, either with a MT system or a CAT system, it is very useful to have a terminological database available. Such a database establishes correspondences between terms in the working languages. And, a proper use of it should lead to an increase of terminological consistency in the translation. In order to build a terminological database, we have explored the lexical extraction over bilingual versions of the same documents. And, for such purpose, we have developed the tool Lexterm (Lexical Extractor for Terminology and Translation).

Traditionally, universities have developed and maintained bilingual glossaries for specific topics, but these glossaries do not cover all the terms that may appear in a document to be translated. However, despite the lack of resources where the translation equivalents in a second language are declared explicitly, the universities have large amounts of documents, along with their translations in this language, where the equivalents are declared implicitly. In order to automatically obtain a list of terms with their translation equivalents from bilingual corpora of specific-domain documents, we developed the Lexterm tool. The language independent tool retrieves a list of term candidates with their translation equivalences from a bilingual corpora. The output of Lexterm can then be used to feed the lexicon of the machine translation system and enrich terminological databases of computer-assisted translation systems.

2 Lexterm's functions

Lexterm performs two main functions. On the one hand, Lexterm extracts relevant lexicon from a text and, on the other hand, presents a translation equivalent of a relevant lexical unit by consulting a bilingual corpus.

The automatic relevant lexicon extraction uses a statistical technique which is based on calculating all word *n-grams* (normally from $n = 2$ up to an n specified by the user, with 3 as the default). Relevant lexical units will be found in the n-grams extracted, although many other non-relevant combinations will

also be found. So a filtering process is performed in order to display those n-grams which are likely to contain significant lexical units first.

Currently, the filtering process of Lexterm is based on the frequency of appearance of *n-grams* in the text and a significant word-combination criterion. The frequency filter consists of displaying a table of the *n-grams* whose frequency of appearance is over a threshold stated by the user. By default, the frequency threshold is 2. The word-combination filtering consists of rejecting those *n-grams* that begin or end with a stop word as candidates to be displayed in the table. Stop words are words that are typically not in the end position (the first or last) of a relevant lexical entry and are mainly functional words. This filtering is performed by consulting a list of stop words, which means that the results of the extraction process will depend on the quality of this list. The user can maintain and improve the list simply by editing it with a text editor. The list of *n-grams* that overcome these filters is a list of relevant lexicon candidates. This list will have to be revised as not all of these candidates will be really relevant. The person in charge of the revision ticks the *n-gram* candidates that are really significant lexical units and can export the *n-grams* ticked to a txt file, so that a monolingual glossary is automatically generated. As the revision is carried out, the list of stop words can be enriched with new words.

The second main function of Lexterm, which is the presentation of a translation equivalent of a relevant lexical unit, is performed by the *TonD* (Terminology on Demand) package that also uses statistical methods. Given a parallel corpus which consists of a list of pairs <s,t> where *s* is a segment in the source language and *t* is the version of *t* in the target language, *TonD* learns the most probable translation of a specific lexical unit by checking the appearances of the lexical unit in each source-language segment and the *n-grams* that appear in each corresponding segment in the target language. Those *n-grams* that appear in the target segments that co-occur with the lexical unit in the source segments are more likely to be its lexical equivalent. It should be remembered that this is a statistical process, and it does not always find the correct solution. For this reason, TonD offers more than one possible candidate so that the user can choose the correct one. The translation equivalent detection can be optimized by filtering the equivalent candidates with stop words. So, to have a list of stop words of the target language can be very helpful. Once the correct equivalences have been selected, the pairs <lexical unit in the source language, lexical unit in the target language> can be exported to a txt file and a bilingual glossary is automatically generated.

In order to help in the selection of a lexical unit and its denomination in the target language, Lexterm provides a *Search* function. This function shows the lexical candidate in its context, which may be very helpful for the user to make sure it is significant in the text. Moreover, the user can verify if the candidate

is a single lexical unit or is embedded in a larger one. The user can also apply this functionality in a bilingual corpus. In this case, not only is the lexical candidate shown in each of the source segments of the corpus but also the translation solution checked by the user is shown in each of the corresponding target segments. The user then verifies if the translation solution checked is the correct one. On the other hand, if none of the solutions are correct, the right equivalent can be found by reading the corresponding target segments.

3 Lexterm integrated into the workflow

Lexterm has been integrated into UOC's document translation workflow. The tool automatically creates bilingual glossaries to be used inside this workflow. In this way, we are setting the basis for a scalable structure in which we recycle every translation process outcome to be reused in future translations.

The workflow begins with a set of original documents belonging to a specific subject domain and their translations. These translations are either post edited versions of raw translations performed by the MT system or translations performed by a human translator with the help of a CAT system. The originals and the translations are aligned in order to obtain an aligned corpus where the source segments and the corresponding target segments are separated by a tabulator. The bilingual corpus is then transformed into a translation memory in the standard *tmx* format, so it can be used for both the MT system and by any CAT system that supports this standard.

The bilingual corpus in txt form is the source from which the process of bilingual terminology extraction is performed by using Lexterm. The extraction result is a file in plain text, in which the source lexical units and their target equivalents are separated by a tabulator. This format allows the file to be directly imported and transformed into a terminological database that can be used in a CAT system. Thus, the terminological database is helpful for a human translator when translating other documents with similar subject domain.

However, the tabulated term list is not directly imported into the lexicon of the machine translation system currently in use at UOC (Translendium[2]), as the linguistic information required by the system to translate the selected terms is not yet available at this point of the process. This information must be coded manually by using the *Lexshop* tool. After this codification, and before validating the lexicon with the new entries, the machine translation of the bulk of original documents is tested. When the test is finished, the lexicon of the system is updated with the new version. Now the system is ready to translate documents, with a similar subject domain, with a higher degree of translation accuracy.

2 `http://www.translendium.com/`

4 Test results

In this section, we present two useful cases in which Lexterm helps us to extract term candidates and their translation equivalents from the parallel corpus.

Firstly, we present the integrated workflow that allows us to extract terms from all kinds of UOC specialised domains, in this case, the institutional report. The corpus that we use in this case corresponds to the UOC's *Academic Year 2004-2005 Annual Report*; specifically, the corpus contains the principal sections of the *Annual Report*: organization, activity, annexes and inside back cover. The volume of words that constitute this corpus is 38,282 in Catalan and 40,619 in Spanish.

From this bilingual corpus, the result we obtained, is a file containing 3,054 aligned segments in Catalan and Spanish, equivalent to a total of 72,721 aligned words. From this aligned corpus, we carried out the automatic extraction of term candidates with the Lexterm tool and obtained a total of 1,271 candidates with their respective translation equivalents. We carried out this automatic selection of candidates using a *3-grams* limit and selecting a file of stop words to filter out words void of content or functional words (articles, prepositions, etc.), placed at the beginning or at the end of the term candidate. Term candidates cannot begin or end with one of these functional words. The program returns the list of term candidates ordered in terms of the number of times they appear in the set of documents. From the 1,271 term candidates extracted with Lexterm, if we take a sample of 10% of the results, or in other words, a total of 127 term candidates, we observe that there are 27 candidates that are denominations belonging to the academic field. Therefore, Lexterm's initial result gives us 21% of the terms belonging to the field.

We filtered this first result, obtained with the automatic extraction tool, by using statistical techniques over relative word frequencies from a general-purpose corpus [3]. These techniques eliminate candidates pertaining to general language so we get a list of more accurate term candidates. We carried out this filtering by means of a general language corpus made up of a million words and of the threshold parameter, which permits control of the level of exigency in considering whether a term candidate belongs to general language or to a specific vocabulary. Upon carrying out this semiautomatic filtering, the number of candidates decreased from 1,271 to 471, that is to say, we reduced the term candidates by 63% using a threshold of 7,000. At the end of the process, this list of term candidates should be revised by a linguist in order to determine which candidates are suitable to be reused in the MT system and by any CAT system. From the 471 filtered term candidates, if we take a sample of 10% of the results, or in other words, 47 term candidates, we observe that there are 27

candidates that are denominations belonging to the academic field. Therefore, when the initial Lexterm result is subsequently filtered it gives us 57% of the terms belonging to the field. As the results that we achieve are satisfactory, we are planning to develop a statistical filter into the Lexterm which works with a general language corpus. This filter will help us to obtain a final list of term candidates that are relevant or belonging to a specialist area.

5 Conclusions and future work

Lexterm is a helpful tool to get relevant lexical units from domain-specific texts when bilingual glossaries of these domains are not available or they are not helpful enough for certain translation requirements. The tool presents a list of candidates with their translation equivalents that saves the effort of doing this task manually, as is still the case in the translation workflow of some universities and institutions concerned with terminology. Moreover, the tool results can be integrated into a workflow where machine and computer-assisted translation are involved. However, the main drawback is the huge number of candidates to be revised when only the *n-gram* frequency and the stop word filters are performed. Furthermore, if we want to reduce the number of candidates by increasing the frequency threshold, we must assume that relevant units whose frequency in the text is below the threshold can be very frequent in future texts. Anyway, regardless of the threshold set, the number of rejected candidates overwhelms the number of accepted ones, so our priority is to reduce this noise. We are working on the integration of the linguistic filtering of *n-grams* that cannot appear in a glossary because they contain finite verbs, conjunctions, adverbials and pronouns.

We have already developed the filtering of *n-grams* that are embedded in larger *n-grams* with the same frequency and in this case we have the following candidates with the frequency in brackets: “echo control” (10), “control device” (8) and “echo control device” (8). Only “echo control” and “echo control device” overcome the filtering. This method has not been integrated into Lexterm yet, but it will be operative in a new version where the statistical calculation of the relevance of *n-grams* in a specific-domain text will also be taken account [3]. This calculation will be carried out by calculating the *tf-idf* value of an *n-gram*, which works on the following assumption: if an *n- gram* with a frequency which is over the threshold in the domain-specific text is rarely found in a large number of documents with varied domains, it is a relevant lexical candidate.

Lexterm is an open-source tool which is freely distributed under a GNU/GPL license. It can be freely downloaded from http://www.linguoc.cat/. Lexterm is intended to be used by any person in any working environment (translation, correction, document generation, etc.) We think this is an interesting approach in the development of free software for translators [4] and for other professionals.

6 Acknowledgements

We would like to thank the linguistic services of the universities involved in the RESTAD project, the Ministry of Education and Universities of the Generalitat de Catalunya that funded it, and Antoni Oliver Gonzàlez, teacher at the Culture and Language Department at UOC, who initiated and participated in the development of the tool above described.

References

1. Climent, S.; Moré, J.; Oliver, A.; Salvatierra, M.; Sànchez, I.; Vàzquez, M.: "Tecnologies de la traducció per a la gestió de la doble oferta docent en català i castellà de la UOC". Zeitschrift für Katalanistik [on line publication]. N^{a}. 18, pp. 31–57 ISSN 0932-2221 (2005)
2. RESTAD Project: `http://www.uoc.edu/serveilinguistic/home/restad/restad.html`
3. Peñas, A., Verdejo, F., Gonzalo, J.: "Corpus-based terminology extraction applied to information access", In Proceedings of the Corpus Linguistics 2001 conference, pp. 458–465. ISBN 1 86220 1 07 2. (2001)
4. McKay, C. Open Source Update: A guide to free and open source software for translators, `http://www.lulu.com/content/178586` (2005)

Tools for Working with Corpus Evidence in the Lexical Database LEXIKON 21 (Program PRAMAT and the Exemplification Tool)[1]

Zdeňka Opavská and Barbora Štěpánková

Institute of the Czech Language of the ASCR, v. v. i.
{opavska,stepankova}@ujc.cas.cz

Abstract. Within the Department of Lexicography and Terminology of the ICL ASCR, v. v. i. an independent program named PRAMAT has been developed for sorting the evidence found in the corpora and possibly other texts. This program serves as a lexicographer's 'desktop' for working with the examples (especially with the corpora evidence) when creating the entries in the lexical database LEXIKON 21. With regard to the complicated structure of the lexical database, a special Exemplification tool for storing the selected contexts has been developed as part of PRALED which allows for more extensive segmentation of the examples than is usual in traditional monolingual dictionaries. This paper will focus on describing and demonstrating the functions of these tools (inserting the evidence from a corpus, amending it with comments, segmenting it in PRAMAT and creating and saving the chosen examples for the future treatment of entries in LEXIKON 21).

1 Introduction

An integral component of the future treatment of entries in LEXIKON 21[2] (hereinafter L 21) is – like it was with the monolingual dictionaries *Příruční slovník jazyka českého* (Reference Dictionary of the Czech Language) [8], hereinafter only as PSJČ, *Slovník spisovného jazyka českého* (Dictionary of the Standard Czech Language) [9], hereinafter as SSJČ, and *Slovník spisovné*

1 This paper was created within the research plan of the ICL of the ASCR, *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (AV0Z90610521).

2 *Lexikon 21 (L 21)* – a Czech lexical database being developed at the Department of Lexicography and Terminology (DLT) of the Institute of the Czech Language (ICL) of the ASCR, v. v. i., within the research plan *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (AV0Z90610521); *PRALED (Prague Lexical Database)* – a program for L 21 being jointly developed by the DLT of the ICL of the ASCR, v. v. i., and the Centre for Natural Language Processing of the FI MU, Brno; *PRAMAT* – an independent program intended for treating evidence (mainly corpus evidence), being developed at the DLT. For more information, see the paper by J. Světlá: *The Possibilities and Limits of Lexicographical Description of the Czech Lexicon in a Database Form.*

češtiny pro školu a veřejnost (Dictionary of Standard Czech for Schools and the General Public) [10], hereinafter as SSČ[3] – work with the lexical material and exemplification evidence. The basic approach when treating the entries is the same: it proceeds from lexical material, from contexts related to the lexical item in question, from which it progresses to the meaning/meanings of the given lexical item; the individual meanings and the area of their usage are then illustrated in the exemplification section of the entry. However, the material resources as well as means and tools which can be used when describing the Czech lexis are different. In many respects, also the principles of the compilation of L 21 are different; this arises from the very nature of this new lexicographic project.

2 Treating lexical material in L 21

L 21 is different in its material resources from the existing monolingual dictionaries. Unlike PSJČ [8], SSJČ [9] and SSČ [10], which took advantage of lexical card archives, the basic material resource for L 21 is a language corpus, namely SYN2000; in 2006, it was joined by further synchronic corpora, i.e. SYN2005 and SYN2006PUB[4]. Nonetheless, like the Czech pre-computer lexicography relied on elaborated processes for treating excerpted lexical material, it has been necessary also for the treatment of entries within L 21 to gradually develop and verify lexicographic procedures for processing corpus material. Such a method has been proposed on the basis of the current knowledge when processing a sample entry.

In brief it may be said that it is a procedure combining a concordance analysis (should there be a high number of concordances, a random sample of a size of 300 examples is used[5]) and an analysis of collocations acquired through the Word Sketch Engine or through the statistical functions of the corpus Most Frequent Collocations and Frequency Distribution. The SYN2000 corpus is the resource for the concordance analysis; if the exemplification is not sufficient for a lexical item or lexical meaning, the

3 For synopsis of the history and conception of Czech monolingual dictionaries, see J. Filipec 1975 [3], Z. Hladká 2005 [5]. Briefly on the conception of the dictionaries in question, see also the treatises on the compilation of a dictionary in PSJČ (1935–1937), vol. 1 [7], SSJČ (1960), vol. 1, [11] and in SSČ (1978, 2003) [12], [13].

4 These resources are expected to be complemented in future by oral corpora and possible other written synchronic corpora made available. Another resource of lexical material will be partial excerption carried out at the DLT of the ICL of the ASCR, v. v. i.

5 When and under what circumstances it will be necessary to augment this sample by additional examples requires further testing.

SYN2000 concordances will be complemented by the SYN2005 or SYN2006PUB concordances. In terms of corpus parameters, items doc.txtype, doc.temp, doc.opus are selected from the data on the source for SYN2000, whereas for SYN2005 and SYN2006PUB they are the items opus.txtype, opus.rokvydání, opus.id. The recommended (however not binding) length of the concordances is a clause to several sentences[6]. When determining collocations through Word Sketch, all three corpora are used, i.e. SYN2000, SYN2005 and SYN2006PUB.

The mentioned work procedures will be additionally tested, complemented and adapted during further lexicographical treatment.

3 The PRAMAT program

During the work with the corpus, it turned out that although the Bonito program allows the concordances to be categorised and placed into individual groups by on the basis of numbers assigned to the individual concordances, for the needs of L 21 it will be necessary to develop a more specific software tool, which would serve as a lexicographer's desktop for the work with examples. Our work experience yielded the proposal of the so-called 'Materiál' (Material) card, which was to function as a component of PRALED. The proposed design subsequently developed into the independent PRAMAT program, whose programmer is P. Žikovský. The program functions independently of PRALED and the files created in it are saved outside the environment of PRALED. It was, however, suggested that selected documents be transferred (copied) from PRAMAT to the Exemplification tool in PRALED in a pre-determined format. PRAMAT is being further developed and modified.

The PRAMAT program consists of two parts: a table above and commentaries below (see Figure 1). *The table part* is intended primarily for treating examples (with the maximum number of examples in one file being 1,000) and makes it possible for examples to be inputted, commented upon, sorted, deleted, copied and transferred to Exemplification in PRALED. The table part of PRAMAT consists of: ID (the identification number of the example), Tag1, Tag2, Tag3 – fields intended for comments on the example (numbers, letters or words can be used here as needed by the compiler), 'Zdroj1' (Source1) – a field for the information on the general source of the example (e.g. SYN2000, SYN2005), 'Zdroj2' (Source2) – a field for the data on the specific source of the example and 'Text' – a field for the text itself of the example. The examples can be sorted on the basis of all fields with the

6 This length has been recommended with respect to the transfer of selected concordances into PRAMAT as well as with regard to the fact that quotational examples (in the length of one or more sentences) will be transferred into the Exemplification in PRALED.

exception of the text of the example. An important function of the tabular part is the function 'Vložit z Bonita' (Paste from Bonito). This function makes it possible to paste one or more pieces of corpus evidence from the files in Bonito in such a manner that the evidence from the corpus automatically separates into fields for 'Zdroj2' and 'Text', and the evidence is also assigned the relevant information on the general source as has been set in 'Zdroj1'. *The part for the commentaries* can be used by the compiler of the entry to record whatever he/she needs for his/her work, e.g. the working explanations of the meaning, working notes, etc. The area for the table or commentary sections can be enlarged/reduced on the screen by pulling the red line separating the table or commentary parts. The PRAMAT program will also make it possible to save selected examples into the Exemplification tool in PRALED in the given format (for L 21 they will be in the form of quotational examples where the lemma will be marked in red while information on the source in green). In the proposal of PRAMAT it is also taken into account that upon clicking on the brief data on the source, its full wording will be displayed. This decoding of source data should function both in PRAMAT and when transferring the examples into Exemplification in PRALED.

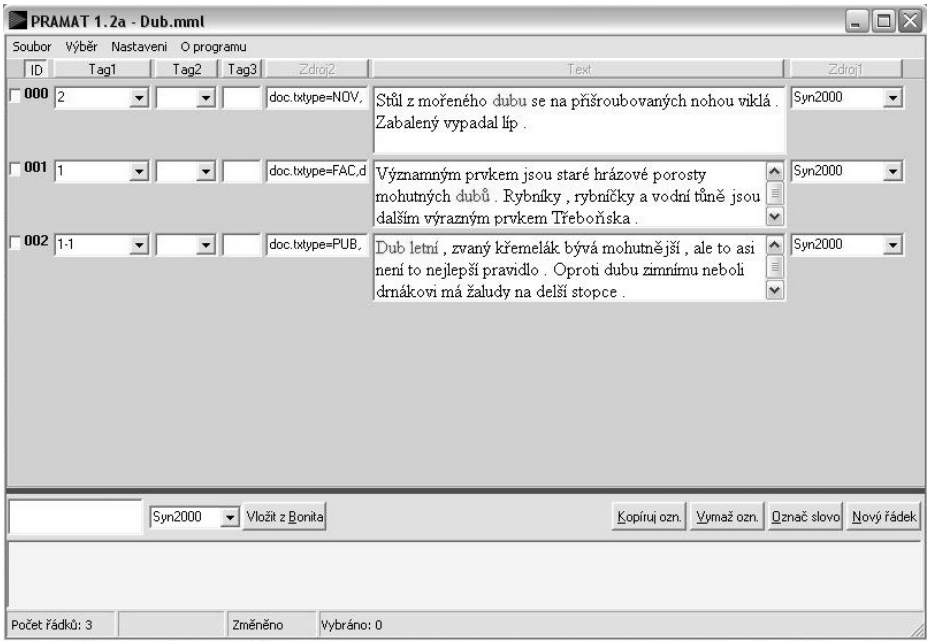


Fig. 1. Pramat

4 Exemplification

The exemplification in monolingual dictionaries is used to illustrate the usage of a lexeme in context, to verify its meaning and to provide information on its lexical and grammatical collocability.[7]

We suggest that two main types of examples be used when treating exemplification in L 21[8]. These are quotational examples on the one hand and non-quotational/illustrative examples on the other. In the Czech lexicographic tradition, both types of exemplification have been used in monolingual dictionaries, namely depending on the size of the given dictionary[9]. An advantage of the database is the possibility to enter a great number of examples, the plan is to use both types of exemplification in a L 21 dictionary entry if possible, because it has turned out that e.g. for the exemplification of synsemantic (functional) words, interjections or some types of adverbs, it is necessary, due to their function in the text (and e.g. also because of homonymy), to provide mainly the quotational exemplification.

Quotational exemplification in L 21 is in the form of contexts of one or more sentences with the resource given, thus formally building on PSJČ [8]. Unlike PSJČ, however, L 21 will contain not only quotes from fiction, but suitable examples in the corpora will be selected from across the genres. An attempt is hence reflected in the quotational exemplification to record the current usage – at least when written published texts are concerned. The corpus name, text type, year of publication and shortened name of the resource are displayed as information on the source of the example in the exemplification.

Ex.: *SYN2000|doc.txtype=PUB, doc.temp=1996, doc.opus=hnk6|V kursu přednášejí učitelé katedry logistiky a obchodního podnikání.*[10]

The term '*non-quotational exemplification*' basically corresponds to the exemplification by illustrative phrases in the form of syntagmas. These are primarily collocations[11]. In this fashion mainly typical and common collocations will be recorded, and their ordering will correspond to the

7 Cf. eg. J. Filipec, 1995, pp. 37–40 [4], F. Čermák, 1995, pp. 107–108 [1].

8 The L 21 exemplification conception is being developed as a joint product of the collective of the DLT of the ICL of the ASCR, v. v. i.

9 The problems of treating the exemplification in PSJČ [8], SSJČ [9] and SSČ [10] were devoted a detailed manuscript by J. Machač (undated) [6]. For more information on the form of exemplification, see also the treatment principles in the individual dictionaries.

10 This is a version of the preview for compilers; the interface for the end user may look different.

11 A collocation may be understood as 'a syntagma of lang. features of a lexical nature' (F. Čermák, Z. Hladká in ESČ, p. 218 [2]) and not as merely 'typical collocations'.

formal-semantic model, the specific collocations will then be selected by the author on the basis of statistical and collocation tools (frequency, cohesion of collocations, word sketches etc.); naturally, also here the author and subsequently the editors have to rely also on their knowledge of the language.

For the ordering of typical and common collocations with nouns in L 21, for instance, this basic model was proposed: 1. adjective+noun, 2. noun+noun, 3. noun+preposition+noun, 4. verb+noun, 5. noun+verb, 6. coordinative phrases[12] (see Figure 2). The model only serves as a basis for authors, because it is not assumed that all the items of the model would exist in the case of all nouns. If it is possible, lexemes occur in the base form with real usage simultaneously being taken into account, e.g. if some collocations occur exclusively in plural, they will also be given in this form in the exemplification.

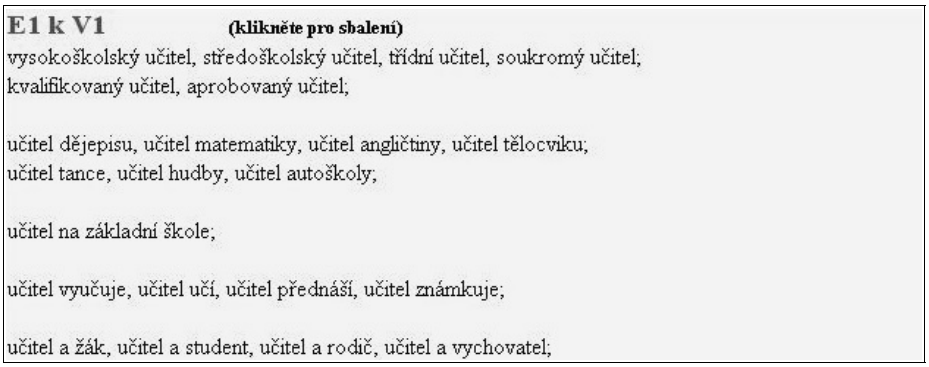


Fig. 2. Exemplification

The usage of the so-called longer adapted examples remains an open question – these sentence structures illustrating an important feature of the a lexical item would be possible to use e.g. for exemplifying the valence of a verb. These would thus be examples whose primary purpose would not be to show real usage in texts but to emphasise a certain phenomenon.

The Exemplification tool in PRALED is structured into the so-called exemplification blocks. Each block contains examples of a similar nature recording the individual types of usage. As far as the order of the blocks is concerned, it proceeds from unmarked examples to marked examples. First

12 We plan for coordinative phrases to appear in L 21 also in the form of quotational exemplification, with their classification under collocations being rather a processing aid.

come independent blocks with non-quotational and quotational examples whose main function is to indicate a typical and common usage of the word, hence to illustrate the collocability, and further to provide an explanation of the meaning. The following separate blocks provide examples requiring some specific complementation: examples with marked usage, examples with figurative usage, examples illustrating a specific grammatical or semantic feature.

All exemplification blocks have had the same form designed for them (see Fig. 3).

A special type of exemplification is the so-called 'miniheslo' (mini-entry). This exemplification block has two different functions. It first aim is to show certain types of multi-word expressions, for instance those that in the hierarchical ordering of the given field specify a concrete species (e.g. the mini-entries *dub letní* and *dub zimní* are placed under the lexeme *dub* in the terminological meaning 'botanický rod stromů Quercus'). Secondly, it has a so-called registration function. Although our current priority is to process one-word lexemes, a proposal for the processing of multi-word items is being simultaneously prepared. Therefore, also collocations, which are a transition to a separate entry (i.e. to an idiom or another multi-word lexical item), are recorded in the mini-entry. Owing to the possibilities of the form, a mini-entry can be provided with explanation, a sufficient number of examples, grammatical information as well as qualifiers; it is thus easily transformable into the form of a separate entry.

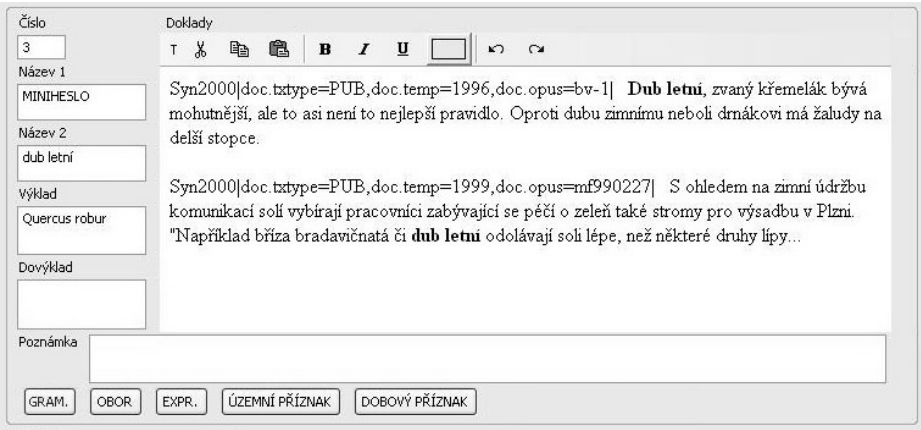


Fig. 3. Exemplification – Mini-entry

Exemplification, as one of the main outcomes of a lexicographer's work with language material, is an important information component of a dictionary entry in L 21. An indisputable advantage of its database form is that it is structured and unrestricted in space. Moreover, the very architecture itself of the database is flexible enough even for the further needs of the compilers which will arise during further elaboration of the conception of L 21.

References

1. Čermák, F.: Paradigmatika a syntagmatika slovníku: možnosti a výhledy. In: F. Čermák, R. Blatná (eds.) Manuál lexikografie. H&H, Jinočany (1995) 90–115
2. Encyklopedický slovník češtiny. Nakladatelství Lidové noviny, Praha, 2002
3. Filipec, J.: Cesta k českému jednosvazkovému slovníku. Naše řeč 58, No. 5 (1975) 225–233
4. Filipec, J.: Teorie a praxe jednojazyčného slovníku výkladového. In: F. Čermák, R. Blatná (eds.) Manuál lexikografie Jinočany, H&H (1995) 14–49
5. Hladká, Z.: České slovníkářství na cestě k jednojazyčnému výkladovému slovníku. Naše řeč 88, No. 3 (2005) 140–159
6. Machač, J.: Slovní spojení (v jednojazyčném výkladovém slovníku). manuscript (undated) 10 pages
7. Předmluva. In: Příruční slovník jazyka českého. A–J. Státní nakladatelství, Praha (1935–1937) VII-XI
8. Příruční slovník jazyka českého. Státní nakladatelství, Praha (1935–1957)
9. Slovník spisovného jazyka českého. Nakladatelství Československé akademie věd/Academia, Praha (1960–1971) 2nd ed. (1989)
10. Slovník spisovné češtiny pro školu a veřejnost. Academia, Praha (1978), 2nd ed. (1994), 3rd ed. (2003)
11. Výklad o uspořádání slovníku. In: Slovník spisovného jazyka českého. A–M. Nakladatelství Československé akademie věd, Praha (1960) VII–XVIII
12. Zásady zpracování slovníku. In: Slovník spisovné češtiny pro školu a veřejnost. Academia, Praha (1978) 779–799
13. Zásady zpracování slovníku. In: Slovník spisovné češtiny pro školu a veřejnost. Academia, Praha (2003) 641–646

Corpora and Programs

14. *Český národní korpus – SYN2000.* Ústav Českého národního korpusu FF UK, Praha 2000. Available online at: `http://ucnk.ff.cuni.cz`
15. *Český národní korpus – SYN2005.* Ústav Českého národního korpusu FF UK, Praha 2005. Available online at: `http://ucnk.ff.cuni.cz`
16. *Český národní korpus – SYN2006PUB.* Ústav Českého národního korpusu FF UK, Praha 2006. Available online at: `http://ucnk.ff.cuni.cz`
17. *Word Sketch Engine.* Available online at: h`ttp://ucnk.ff.cuni.cz/corpora`

Computer Processing Derivational Relations in Czech

Karel Pala and Dana Hlaváčková

Faculty of Informatics, Masaryk University Brno
Czech Republic
{pala,hlavack}@fi.muni.cz

Abstract. In the paper we deal with the derivational relations in Czech that form typical derivational nests (or subnets). Derivational relations are mostly of semantic nature and their regularity in Czech allows us to describe them in a way suitable for computer processing and then add them to the electronic databases such as WordNet almost automatically. For this purpose we have used the derivational version of morphological analyzer Ajka that is able to handle the basic and most productive derivational relations in Czech. A special derivational interface has been developed in our NLP Lab at FI MU by means of which we have explored the semantic nature of the selected noun derivational suffixes (22) as well as verb prefixes and established a set of the semantically labeled derivational relations – presently 14. With regard to the verbs we have paid attention to the selected verb semantic classes in connection with the derivational relations between selected prefixes (4) and corresponding Czech verbs. As an application we have added the selected derivational relations to the Czech WordNet and in this way enriched it with approx. 30 000 new Czech synsets.

1 Introduction

In the highly inflectional languages like Czech the derivational relations represent a system of the semantic relations that definitely reflects cognitive structures that are used by the language users and should serve as base for developing a linguistic ontology. Such ontology undoubtedly exists in the minds of the language users but according to our knowledge it has not been written down yet. Obviously, for language users derivational affixes (prefixes, infixes, suffixes) function as formal means that express semantic relations necessary for using language as a vehicle of communication. In our view, for those reasons the derivational relations should be considered as having semantic nature though a question may be asked what kind of semantics we are dealing with (see Sect. 4). In any case it has to be remarked that grammatical categories such as gender or number display a clear semantic nature. Thus our motivation is to explore these issues and formulate derivational rules (D-rules) allowing to generate automatically as many word forms as possible. Having these rules would mean that it would not be necessary to compile and keep large lists of stems (word forms).

2 Formal and derivational morphology in Czech

In Czech words are regularly inflected (declined, conjugated) as they express different grammatical categories (gender, number, case, person, tense, aspect etc.) using affixes. This is what is called *formal* morphology in Czech grammars and its description mostly deals with the system of the inflectional paradigms. Then there is a *derivational* morphology which deals with deriving words from other words, e.g. nouns from verbs, adjectives from nouns or verbs etc., using affixes again. The derivations are closely related to the inflectional paradigms in a specific way: we can speak about derivational paradigms as well, they are described, for instance, in Pala, Sedláček, Veber, 2003.

For Czech inflectional morphology there are automatic tools – morphological analyzers exploiting the formal description of the inflection paradigms – we work with the analyzer called Ajka (cf. Sedláček, Smrž, 2003) and developed in NLP Lab at FI MU. Its list of stems contains approx. 400 000 items, up to 1600 inflectional paradigms and it is able to generate approx. 6 mil. Czech word forms. The dictionary of stems has served as a starting collection of data for the study of the Czech derivational relations.

We are using Ajka for several purposes: lemmatization and tagging, as well as a module for Czech syntactic analyzer, etc. We have also developed a derivational version of Ajka (D-Ajka) that is able to work with the main regular derivational relations in Czech – it can generate new word forms derived from the stems. The present version of D-Ajka deals reasonably with suffixes, the processing of prefixes is the task we are working on.

2.1 Derivational relations in Czech

The D-relations cover a large part of the Czech word stock (up to 70 %). Here we are interested in describing derivational processes (see examples) by which new words are formed from the corresponding word bases (stems or roots). In Czech grammars (Mluvnice češtiny, 1986) we can find at least the following main types (presently 14) of the derivational processes:

1. **mutation** noun → noun derivation,
   e.g. *kniha* → *knih-ovna* (*book* → *library*), semantic relation expresses location – between an object and its typical location. It also includes verb – verb derivation which is realized by means of prefixation, see below Sect. 2.2.1.
2. **transposition** (the relation existing between different POS):
   noun → adjective derivation, e.g. *noc* → *noč-ní* (a night → night), semantically the relation expresses property,

3. **agentive relation** (existing between different POS):
   verb → noun e.g. *myslit → mysli-tel* (think → thinker), semantically the relation exists between action and its agent,
4. **patient relation** verb → noun,
   e.g. *povolat → povolan-ec* (*to conscript → a conscript*), semantically it expresses a relation between an action and the object (person) impacted by it,
5. **instrument (means) relation** verb → noun,
   e.g. *hořet → hoř-ák* (*burn → burner*), semantically it expresses a tool (means) used when performing an action,
6. **action relation** (existing between different POS): verb → noun,
   e.g. psát → psa-n-í (write → writing), usually the derived nouns are characterized as deverbatives, semantically both members of the relation denote action (process),
7. **property-va relation** (existing between different POS):
   verb → adjective, e.g. *prodat → prod-aný* (sell → sold), usually the derived adjectives are labelled as de-adjectives, semantically it is a relation between action and its property,
8. **property-aad relation** (existing between different POS):
   adjective → adverb, e.g. *plný → pln-ě* (*full → fully*), semantically we can speak about property,
9. **property-an** (existing between different POS): adjective → noun,
   e.g. *laskavý → laskav-ost* (*kind → kindness*), semantically the relation expresses property in both cases,
10. **gender change relation** noun → noun,
    e.g. *řidič → řidič-ka* (*driver → she driver*), semantically the only difference is in sex of the persons denoted by these nouns (male – female),
11. **diminutive relation** noun → noun → noun,
    e.g. *květ → kvít-ek → kvít-eček* (*blossom → small blossom → very little blossom* or more typically a *blossom* to which a speaker has an emotional attitude), in Czech the diminutive relation can be binary or ternary,
12. **augmentative relation** noun → noun,
    e.g. *chlap → chlap-isko* (*guy → big, strong guy*), semantically it expresses different emotional attitudes to a person.
13. there are two more relations that are sometimes regarded inflectional but in our view they belong to the derivational ones as well:
    gerund relation verb → adjective: *mluvit → mluvící* (*speak → speaking*) and **passive relation** verb → adjective (passive participle):
    (*koupit → koup-en*, *buy → bought*)

14. relations expressed by prefixation hold mainly between a base verb → derived verb, e.g *nést → od-nést* (*carry → carry away*) or *jít → při-jít (go → come).* The inventory of prefixes in Czech consists of approx. 241 items and they form a grouping which consists of the core of about 19 primary prefixes, i. e.: *do-, na-, nad-, o-, ob-, od-, po-, pod-, pro-, pře-, při-, roz-, s-, se-. u-, v-, vy-, z-, za-.* Then there is a group of double prefixes like *do-pře, do-vy-, od-na-, po-vy-, po-na-, po-po-*, etc., whose number is quite large, about 190. Further, there are prefixes of foreign origin coming mainly from Latin and Greek, e. g. *anti-, dis-, des-, hyper-, hypo-, inter-, intra, meta-, sub-, super-*, etc.

Semantically, prefixes in Czech denote a number of different relations such as various properties of motion, location, time, distribution, measure and some others. The complete description of the semantics of Czech prefixes, however, would be rather laborious, thus here we will show just the possible solutions for few selected verbs of motion – see below Sect. 3.2.

3 Web interface for derivation relations

As we have already hinted the formal means epressing D-relations are affixes, i.e. prefixes, stem-forming infixes and suffixes. For their analysis we have developed a web interface that allows us to find combinations of the nouns with respective suffixes and verbs with corresponding prefixes. Infixes or intersegments can also be handled through the web interface but they are basically covered by the list of stems – instead writing rules for changes in stems we prefer to use more variants of one stem (mainly for technical reasons). We could try to perform a deeper root analysis but this is a topic for a separate paper.

As starting data we have used a list of noun stems taken from the stem dictionary of the D-Ajka analyzer – it includes approx. 126 000 items. The similar list of verb stems contains 42 745 items from which 4763 items are classified as iteratives. The procedure of deriving both nouns and verbs consists of the three basic steps:

1. a set of words is defined by means of the prefix, suffix and morphological tag;
2. defining a derivational rule – typically it is a substitution of morphemes (suffixes) at the end of the word (noun) and prefixes at the beginning of the word (verb);
3. manual processing of the obtained results (lists) – usually correcting or deleting cases that cannot be regarded as properly derived forms though they may follow the given rule.

3.1 Nouns

An example of the derivational analysis for suffix *–ík*:: it occurs with the nouns denoting an agent or instrument (means), e.g. *zed-n-ík* (*bricklayer*) or *kapes-ník* (*hankerchief*).

First, we wanted to derive agentive nouns: so we have entered the suffix *-ík* and tag k1gM (noun, masculine animate) and generated the list of all words ending with *-ík*. The output is a list of 1210 nouns including proper names (from the original list of 126 000 Czech nouns). To obtain instrument nouns we input the tag k1gI (noun, masculine inanimate). As an output result we get a list of 715 nouns including proper names. The number of all words ending with suffix *-ík* (disregarding the grammatical tag) in stem dictionary of Ajka is 1830. The difference in the given numbers follows from the homonymy, for instance, some nouns can be both masculine animate and masculine inanimate (e.g. the noun *náčelník* can denote – *chief* as well as *čelenka – headband*). Such cases have been processed manually.

In this way we processed 22 Czech derivational suffixes and as a result we obtained a detailed classification of the indicated derivations capturing agentive, instrumental, location and also resultative relation, for instance *spálit* → *spálenina* (*to burn* → *a burn*) which should be mentioned as well. At the same time the complete lists of all stems with the indicated suffixes together with labeling their semantic relations between the stems and respective suffixes were obtained as well. For the processed suffixes the coverage is complete (with regard to the list of 126 000 Czech noun stems and).

3.2 Verbs

The web interface works with verb prefixes in a similar fashion. For this example we have selected four verb prefixes: *do-* and *od-* (they are semantically symmetrical*), při-* and *před-* (from the 42 745 items).

1. The derivational rule consists of the prefix *do-* and morphological tag k5 – the web interface found 1239 verbs, 173 are verbs of motion. They denote motion to a point (destination) (up, down), using legs or means of transport (typically a vehicle).
2. The derivational rule consists of prefix *od-* and tag k5 – the web interface gives 1598 verbs, the number of the verbs of motion with prefix *od-* is 107. They denote motion from a starting point (up, down), using legs or means of transport (typically a vehicle).
3. The derivational rule consists of prefix *při-* and tag k5 – the web interface yields 1318 verbs, from them 170 are verbs of motion. They denote motion to a target (point, place, space).
4. The derivational rule consists of prefix *pře-* and tag k5 – the web interface offers 1305 verbs, from them 207 are verbs of motion. They denote a motion through an environment or over a barrier.

Thus using the described procedure we are able to find pairs of the word forms in which the first one is considered basic and the second one derived. The direction of the derivations is not always unambiguous but the most important goal is to establish the relation itself – to decide about its direction is not so relevant. There are also cases when changes in stem take place – they have to be checked and added manually.

3.2.1 Verb classes and the selected prefixes

If we have a look at the rules defined for the derivational web interface and at the resulting list of the verbs obtained by their application we can see that they contain some semantic groupings.

In other words, we obtain classes that can be understood as verb semantic classes, e. g. verbs of motion, verbs of motion with vehicles, verbs denoting time properties of actions (apart from the aspect).

In NLP Lab at FI MU we are working on a lexical database VerbaLex containing complex valency frames for approx. 12 000 Czech verbs (Horák, Hlaváčková, 2006). Within VerbaLex we work with verb semantic classes that were originally adopted from the Levin's list of English verb classes (Levin, 1993) and the list of Martha Palmer's VerbNet project (Palmer et al, 1998, 395 classes). These verb classes have been translated and adapted for Czech language. Presently, we work with approximately 100 semantic verb classes in the VerbaLex.

It is necessary to stress that we initially started with Levin/Palmer's classes but within VerbaLex they were modified with regard to the predicate-argument structures of Czech verbs, i. e. our classes are based also on the inventory of the semantic roles denoting verb arguments. Thus, in building the semantic classes we prefer semantic criteria against the alternations used by Levin. As a result we get verb classes that are semantically more consistent than Levin's. This can be demonstrated, for instance, with the classes containing various verbs of motion (see below). Then it becomes obvious that Levin/Palmer's classes mix up arbitrarily the verbs that should be kept apart and put together the ones that are semantically different. This is a consequence of the fact that the Levin's criteria are more syntactic (alternations) than semantic.

When we started working with verb prefixes we realized that they also allow us to make the verb classes more consistent and less arbitrary as well as closer to the real lexical (corpus) data. This follows from the fact that in Czech prefixes determine and modify verb meanings in a quite regular way. The derivational interface makes it possible to generate all verbs with the selected prefix from the list of the verbs (morphological database). In this way the prefixes help us to find semantically homogenous groups – classes – of verbs that can be confronted with the existing Levin/Palmer's classes.

Regularity of the prefixes can be demonstrated on the following group of the verbs of motion – they denote typically motion performed by humans or animals using their legs – those verbs can occurr with all following four prefixes: *do-, od-, při-, pře-*:

batolit se (*toddle*), *-běhnout (run), belhat se* (*walk with a limp*), *capat* (*toddle*), *capkat* (*totter*), *cárat* (*saunter*), *cupat* (*patter*), *cupitat* (*scurry*), *cupkat* (*patter*), *courat se* (*stroll*), *dupat* (*stomp*), *dusat* (*stamp*), *fičet* (*zoom*), *hnát se* (*rush*), *hopkat (trip), hopsat* (*skip, jig*), *hupkat* (*frisk*), *hupsat* (*caper about*), *jít* (*walk, go*), *klopýtat* (*stumble*), *kolébat se* (*waddle*), *kulhat* (*limp*), *kulit se* (*trundle*), *kutálet se* (*roll*), *lézt* (*crawl*), *pajdat* (*hobble*), *pelášit* (*scamper*), *plahočit se* (*plod*), *plazit se* (*creep*), *plížit se* (sneak), *plouhat se* (*trudge*), *ploužit se* (*struggle along*), *pochodovat* (*march*), *potácet se* (*stagger*), *řítit se* (*dash*), *skočit (jump), skotačit* (*frolic*), *škobrtat* (*trip up*), *šlapat* (*step, cycle*), *šmajdat* (*drag along*), *šmatlat* (*drag along*), *šourat se* (*shamble*), *štrachat se* (*drag oneself*), *ťapat* (*patter*), *valit se* (*roll, draw near*), *vandrovat* (*ramble, hike*).

Thus we have *do-běhnout* (*reach, get to*), *od-běhnout* (*run away from*), *při-běhnout* (*dash in*), *pře-běhnout* (*run across, run over*) and similar derivations for all the verbs in the list.

Some English equivalents above may be rather near synonyms, we give here only the senses related to a motion. The existing dictionaries are not reliable in this respect except for Fronek (2000).

4 What is the nature of the D-relations?

In sect. 2.1 we have introduced the labeling of the Czech D-relations. The question may be asked what is the real nature of D-relations, whether it is semantic or rather morphological (formal). The D-relations exist between morphemes, typically between stems and corresponding suffixes or prefixes. This formal feature makes them different from the relations between sentence constituents, as e.g. between verbs and their arguments. However, the main criterion is whether the particular relation affects meaning irrespective of its formal realization.

If we apply this criterion to the D-relations discussed above, such as deriv-ag, deriv-loc, deriv-instr, deriv-g, deriv-dem, deriv-pos, deriv-pro, we definitely come to the conclusion that their nature is semantic.

D-relations like deriv-an, deriv-na, deriv-dvrb, deriv-ger, deriv-aad, deriv-pas are sometimes characterized as morphological only and their semantics is left aside. The first two relations hold between nouns and adjectives and both denote properties (e.g. deriv-an: *nový* → *novost* (*new* → *newness*)), but we have to take into account that there is something that may be called semantics of the parts of speech, i.e. in one case property is expressed by the adjective and then by the noun which is derived from the adjective. Deriv-na denotes property as well but here the adjective is derived from noun as in *boj* → *bo-*

jovný (*fight → combative*). The relation deriv-dvrb exists between a verb and noun, e.g. číst → čtení (*read → reading*), and it denotes action which is first expressed by the verb and then by the deverbative noun. We can say that in these cases the only difference lies in the optics of the individual parts of speech but this difference should be understood as semantic as well. However, it should be remarked once more that quite often the differences in the semantics of the parts of speech are not treated as truly semantic.

If we look at what standard Czech grammars (see e.g. Karlík et al, 1995) say about the semantics of the parts of speech we find the formulations such as: nouns denote independent entities, i.e. persons, animals and things and also properties and actions. Verbs then denote states and their changes and processes (actions) and their mutations. These descriptions certainly refer to the semantics of the nouns and verbs. They are usually followed by the explanations about morphological processes, i. e. usually derivations by which some parts of speech are formed from the others, as we have described them above. What is relevant and what is missing in the standard grammars are more detailed and extensive semantic classifications of nouns, verbs, as well as adjectives and numerals. They are beginning to appear only recently and have the form of ontologies – the standard grammars do not use this term at all.

4.1 D-relations and verb prefixation

The situation is slightly different for the relation that was labeled as prefixation above. Here prefix is a formal means that denotes the meaning of the derived verb, thus it is possible to speak about the meanings of the prefixes. Their descriptions can be found in literature (Šlosar, Rusínová, 1982) but they take a form of the examples without considering more complete data. What we try to show here is how the picture may look like with larger data. However, the task is quite complicated and we can offer just an outline how the prefix relations could be processed.

If we take prefixes *do-* and *od-* the relations induced by them are the following:

a) *do-* denotes motion **to** a point or place (space) – *do-jet* (*arrive*), *do-jít* (*come to*)

b) time – finishing an action – *do-číst* (*finish reading*)

c) additivity – *do-lít* (*fill up*).

Prefixes *do-* and *od-* are semantically symmetric with regard to a), so *do-* denotes:

a) motion **from** a point or place (space) – *ode-jít* (*go away*), *od-jet* (*leave by a vehicle*)
b) fulfilling an obligation – *od-pracovat* (*work-off*)
c) time – completing, finishing an action – *ode-hrát* (*play out*).

Two remarks have to be made:

first, verbs of motion in Czech differ depending on what means of transport they imply, e. g. legs or vehicles – this difference is not expressed systematically in English (see e. g. class escape-51.1-1 in Levin/Palmer's classification),

second, one group of the motion verbs occurs with an Agent (human, animal, vehicle) only (*při-jít, come*) while other verbs of motion have two obligatory actants, i. e. moving Agent and moved Object (*nést knihu – od-nést knihu*, *carry a book – carry away a book*). In our view, these distinctions should be reflected in the different semantic classes, which, however, is not always a case, see, for instance, Levin/Palmer's class run-51.3.2-1.

5 Results – an application

A possible application of the above mentioned processing D-relations by the derivational Ajka (apart from prefixation) is adding derived literals (lemmas) to the Czech WordNet. The final result – the number of the literals generated from the individual D-relations is given below together with their semantic labels:

deriv-na	641 (property, noun → adj)
deriv-ger	1951 (property, verb → adj)
deriv-dvrb	5041 (action, verb → noun)
deriv-pos	4073 (possessive, noun → adj)
deriv-pas	9801 (passive, verb → adj)
deriv-aad	1416 (property, adj → adverb)
deriv-an	1930 (property, adj → noun)
deriv-g	2695 (gender, noun → noun)
deriv-ag	186 (agentive, verb → noun)
deriv-dem	3695 (diminutive, noun → noun)
Total	31429 literals

These numbers also tell us how productive the particular relations are. Note that the most frequent is passive relation which is followed by the deverbative (action) relation. The third most frequent relation is a possessive one. It would be interesting to examine what these facts can tell us about semantic structure of texts.

Enriching Czech WordNet with the D-relation makes it possible to capture other semantic relations than the basic one (synonymy, hypero/hyponymy, antonymy, etc.) in this electronic lexical database (Pala, Hlaváčková, 2007).

6 Conclusions

In the paper we present the first results of the computational analysis of basic and most regular D-relations in Czech using derivational web interface and derivational version of the morphological analyzer Ajka.

Though the analysis is far from complete at the moment the number of the generated items has led us to the decision to include them in Czech WordNet and enrich it considerably with the derivational nests (subnets). In our view, this kind of enrichment makes Czech WordNet more suitable for some applications, namely for searching.

The second and even more important reason for doing all this is a belief that the derivational relations and derivational subnets created by them reflect basic cognitive structures existing in natural language. More effort is needed to explore them from the point of view of now so popular ontologies – they certainly offer an empirical ground (on the formal level they are expressed by the individual morphemes) for natural language based ontologies.

Acknowledgements

The research was supported by the grant projects GA 201/05/2871, 1ET100300419, LC536 and NPVII 2C06009.

References

1. Fronek, J., 2000. Comprehensive Czech-English Dictionary, Leda, Prague.
2. Horák A., Hlaváčková, D. VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In Computer Treatment of Slavic and East European Languages, Third International Seminar. Bratislava: VEDA, 2005, p. 107–115.
3. Horák A., Pala K., Rambousek A., and Povolný M. 2006. First version of new client-server wordnet browsing and editing tool. In Proceedings of the Third International WordNet Conference – GWC 2006, p. 325–328, Jeju, South Korea, Masaryk University, Brno.
4. Horák A., Smrž P. 2004. Visdic – WordNet Editing and Browsing Tool, Proceedings of the 2nd GWC, Brno, Masaryk University.
5. Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: The University of Chicago Press.

6. Karlík P. et al. 1995. Příruční mluvnice češtiny (Every day Czech Grammar), Nakladatelství Lidové Noviny, Prague, pp. 229, 310.
7. Pala K., Hlaváčková D. Derivational Relations in Czech WordNet. In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. Praha: ACL, 2007, p. 75-81.
8. Pala K., Sedláček R., Veber M. 2003. Relations between Inflectional and Derivation Patterns, Proceedings of EACL, Budapest.
9. Palmer, M., Rosenzweig, J., Dang, H. T. et al. Investigating regular sense extensions based on intersective Levin classes. In *Coling/ACL-98, 36th Association of Computational Linguistics Conference.* Montreal 1998, p. 293–300.
10. Petr J. et al. 1986. Mluvnice češtiny 1 (Grammar of Czech), Praha: Academia.
11. Sedláček R., Smrž P. 2001. A New Czech Morphological Analyser Ajka. Proceedings of the 4th International Conference on Text, Speech and Dialogue, Springer Verlag, Berlin, p. 100–107.
12. Šlosar D., Rusínová Z. 1982. Čeština pro cizince, skriptum, FF MU, Brno, p. 95–97.
13. Vossen P. 2003. EuroWordNet General Document, Version 3, University of Amsterdam.
14. Web address of the Princeton WordNet 3.0:
`http://wordnet.princeton.edu/perl/webwn`.

Wider Framework of the Research Plan *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century*[1]

Albena Rangelova and Jan Králík

Institute of the Czech Language of the ASCR, v. v. i.
{rangelova,kralik}@ujc.cas.cz

Abstract. This paper informs about the wider framework of work activities within the research plan *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century*, where the central position is given to LEXIKON 21 and PRALED. Further goals of the Department of Lexicography and Terminology of the Institute of the Czech Language of the ASCR, v. v. i. are the consolidation of existing material collections and descriptive databases in cooperation with the Department of Data Electronisation: digitisation of collections of excerption slips, creation of new excerption databases as well as digitised versions of dictionaries. The final objective is to present our results to both the professional and general public.

The strategic goal of the institutional research plan *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (2005–2010), implemented at the Institute of the Czech Language of the ASCR, v. v. i. (hereinafter only ICL), is a comprehensive preparation for the creation of a modern monolingual dictionary. Especially the Department of Lexicography and Terminology (DLT) in cooperation with the Department of Data Electronisation (DDE) have participated in the realisation of this singular project. The individual aspects of the main stream of the work – the design of software tools, a number of conceptual and realisation questions – are dealt with in other papers of the members of the DLT in these proceedings. Here we would like to inform briefly on the wider framework, which comprises miscellaneous activities focused especially on the creation of the material, technical and personnel prerequisites for lexicographic work, including the presentation of lexicological and lexicographic research for the wider public using modern information technologies.

The research plan contributes to creating an integrated database system consisting of a number of component parts which have so far existed either in

1 This paper was created within the research plan of the ICL of the ASCR, v. v. i. *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (AV0Z90610521) and within the research project of the ASCR *Resources and Tools for Information Systems* (1ET101120413).

isolation or only in printed form. The component wholes will be interconnected into a uniform search environment, which will facilitate effective searching both in the database as a whole and in the individual component files. The data made accessible will additionally be possible to use for various linguistic purposes, e.g. for further lexicographic elaboration or any general linguistic studies in the area of lexis.

When fulfilling the tasks of the research plan, we plan to build on the scientific results of a number of cooperating workplaces. Extensive cooperation has been established chiefly with the Centre for Natural Language Processing of the Faculty of Informatics of Masaryk University in Brno. We further cooperate with the Institute of Theoretical and Computational Linguistics at the Philosophical Faculty of Charles University (PF CU) and with the Institute of Formal and Applied Linguistics at the Mathematical-Physical Faculty of Charles University. An especially significant partner is also the Institute of the Czech National Corpus at the PF CU, administering the extensive text corpora SYN2000, SYN2005, SYN2006PUB and others.

The existing research plan is in accord with the overall strategy of the ICL to build a database of the rich lexis of the Czech language and progressively make it accessible in such a way that it would be possible to both extend it further and use it optimally. In terms of the strategic goals of the workplace, this also involves the preparation of existing primary and secondary resources of lexical material for their use at a new technological level – specifically the scanning and description of the lexical collections, the digitisation of the dictionaries created at our workplace, the transfer of our electronic collections to a higher technological platform, etc. The aim is the above-mentioned uniform user environment joining an entire range of component databases, descriptive as well as material.

The central place among the descriptive databases will be taken by the very detailed, contemporarily-conceived lexicographic description of Czech lexis named LEXIKON 21 (hereinafter L 21), see the paper by J. Světlá in these proceedings. The main content of the research since the beginning of the work on the research plan has been the creation of a lexicographic workstation customised for our goal: on the one hand the development of the software for the description of lexical items (PRALED) and for work with material (PRAMAT), see the paper by Z. Opavská and B. Štěpánková here, and on the other the elaboration of the conceptual questions of the future multifaceted description of the lexis (see also the papers of J. Světlá, M. Voborská, E. Birkhahnová and V. Chudomelová in these proceedings).

We further strive for the creation of a number of dictionary databases of already-published dictionaries transferred into electronic form. The following works are already accessible now in various applications and in different modes: *Příruční slovník jazyka českého* [Reference Dictionary of the Czech Language], hereinafter only as PSJČ [9], *Slovník spisovného jazyka českého* [Dictionary of the Standard Czech Language], hereinafter as SSJČ, [12], *Slovník spisovné*

češtiny pro školu a veřejnost [Dictionary of Standard Czech for Schools and the General Public], hereinafter as SSČ [10], [11], *Akademický slovník cizích slov* [The Academic Dictionary of Loanwords in Czech] [1], cf. also [17]. These publications have been made available as electronically-stored texts for the internal purposes of the research team: *Retrográdní slovník současné češtiny* [Retrograde Dictionary of Contemporary Czech] (on the basis of the Czech Academic Corpus) [16], *Slovník slovesných, substantivních a adjektivních vazeb a spojení* [Dictionary of Verbal, Nominal and Adjectival Phrases and Collocations] [15], *Nová slova v češtině. Slovník neologizmů 1* [New Words in Czech: A Dictionary of Neologisms 1], hereinafter as SN 1 [7], *Nová slova v češtině. Slovník neologizmů 2* [New Words in Czech: A Dictionary of Neologisms 2], hereinafter as SN 2 [8], *Český jazykový atlas 1* [Czech Linguistic Atlas] [2], the work *Slovesa pro praxi. Valenční slovník nejčastějších českých sloves* [Verbs In Practice: A Valency Dictionary of the Most Frequent Czech Verbs] [14] is being electronised, we plan to scan the publication *Co v slovnících nenajdete. Novinky v současné slovní zásobě* [Not Found in Dictionaries: New Words in Contemporary Czech Word Stock] [13] in the future. When developing this segment of the lexical database, it is necessary to respect not only intellectual property rights but also copyright and the interests of the publishers, therefore there will be certain limitations in the accessibility to individual works and distinctions will be made between them. It will be used mainly for searching component information for scientific or other non-commercial purposes.

The electronisation of historical dictionaries is greatly to the credit of the DDE. This task converges with research projects on language development and partially overlaps them[2]. The reason for this is that the conception of processing the main multi-volume works (the dictionaries of Jungmann [5], Gebauer [4], Kott [6]) offers important reference information also for the processing of the contemporary language and the same may be the case with smaller dictionaries, e.g. subject-based, but also older until the so-called middle period. For the sake of the optimisation of uniform searching, it will therefore be necessary to unify the user environments (search engines) created up to now of various electronic forms of historical as well as more recent dictionaries and progressively incorporate them into a new, unifying system. The vision of a universal search engine which would go through the indexes has the working title 'lemmarium'. The gradual sharpening of the contours of its conception will preserve the specifics of individual dictionaries; however, it cannot be excluded that experience and technical possibilities will also provide this outcome of the project with solutions which have been beyond the current horizon. The

2 See the website Vokabulář webový [Web Vocabulary], `http://vokabular.ujc.cas.cz`

connection of the lexical archive of the ICL with the electronic form of PSJČ [9] (see below) just being implemented is a step in this direction.

There are two types of material databases which the DLT of the ICL has at its disposal – one part has been designed and developed as software products with various purposes, the other part is being created by the gradual electronisation of the lexical collections of the workplace. The neological database of the DLT was already established within the project *Popis nové slovní zásoby s využitím počítačové techniky* [Description of New Vocabulary Using Computer Technology] (1994–1996) and progressively complemented with new data in the course of the subsequent grant projects (*Systémotvorné procesy neologizmů v současné češtině* [System-Forming Processes of Neologisms in Contemporary Czech], 1998–2000, *Internacionalizmy v nové slovní zásobě češtiny* [Internationalisms in the Present-Day Word Stock of Czech], 2001–2003). Its first part with the working title 'Archiv 1' contains 203,000 electronically stored excerpts (see Fig. 1).

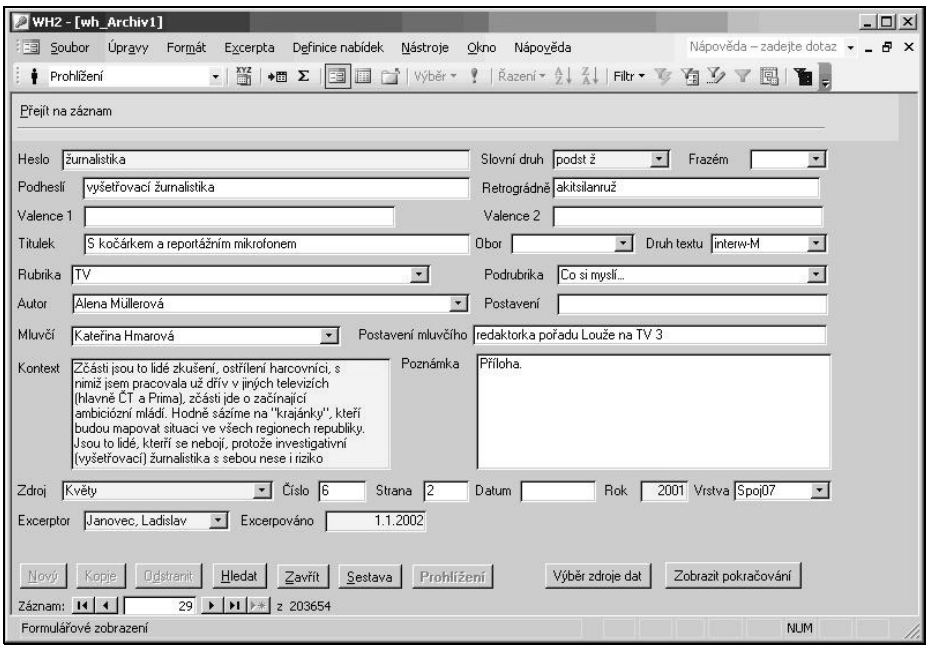


Fig. 1. WinHesla2 Program – an excerption card

This collection of neologic lexical material gave rise to two dictionaries of neologisms – SN 1 [7] and SN 2 [8]. In accordance with the new tasks of the DLT, the excerption was expanded to the phenomena of synchronic dynamics not captured in the existing dictionary works. The methodology of work was also updated – much more actively taking advantage of the electronic text archives

(Newton) [21] and internet sources. A specific task is to ensure the compatibility of the database of neologic material with the new software of the workplace. The Excerption program WinHesla2 (programmer B. Lehečka) works in the database environment of MS Access 2003, which has its limitations, and therefore we expect to transfer the data and user interface to a more flexible platform. We are also planning to augment the excerption by the terminological level in a separate database.

The database created on the basis of the lexical archive of the ICL, containing 9.5 million card excerpts of general (Modern Czech) lexis has enormous significance. It is priceless material, which made it possible for exemplary monolingual dictionaries of the Czech language, PSJČ [9], SSJČ [12], SSČ [10], to be created. An electronic form of the archive is now being developed – the complete collection of excerpts has already been scanned and the images of the excerption cards are currently being annotated and inserted into a database (programmer M. Spousta). More than 4 million cards have been prepared in electronic form in the database so far. The interconnection of the database of the lexical archive with the electronic form of PSJČ [9] (programmer P. Květoň), allowing simultaneous work with both collections (see Fig. 2), is unique.

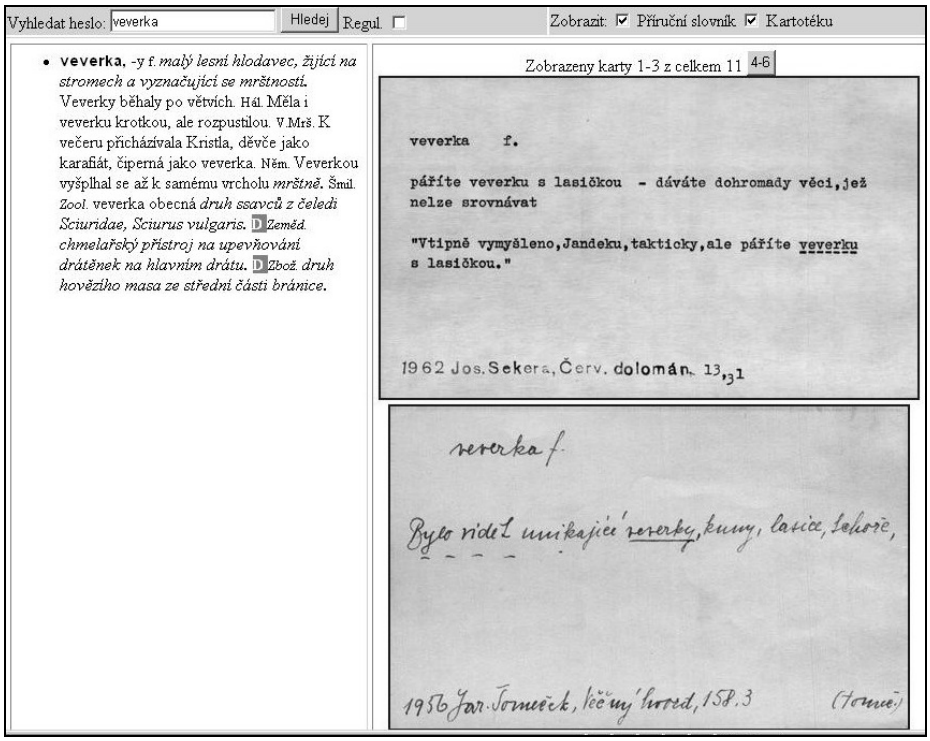


Fig. 2. Word found simultaneously in PSJČ [9] and in the database of the lexical archive

Smaller, collections of linguistic and technical terminology significant on their own (almost 300,000 electronically documented and glossed excerpts) have also been transferred into electronic form, and their database version will be available on the internet. These will gradually be joined by further material collections in electronic form, e.g. an extensive dialectological archive, lists of place-names, a collection of personal names, etc., and also smaller useful catalogues prepared at the Department of Language Culture as well as in individual research projects.

A specific area of work, which will require proper attention, is the presentation of scientific results for the wider public. Alongside the already-existent webpages of the ICL, we plan to put into operation a web nest of the DLT, where joint applications with graduated user rights will be available (some of the content components will continue to have an internal, purely working character also in the future). Currently, the applications Database of Indexes and Bibliographic Database have already been prepared. The Database of Indexes presents a union list of entries of PSJČ [9], SSJČ [12], SSČ [10] and *Frekvenční slovník češtiny* [Frequency Dictionary of Czech] [3]. For the needs of analytical work, also some unpublished integral wholes of words, e.g. a proposal of a list of entries of the lexical standard, were incorporated into the database of indexes[3]. In this database, a lemma can be searched for by entering a character string while selecting its placement (at the beginning or end of a word), see Fig. 3.

Fig. 3. Database of indexes with the displayed chain at the end of a word

3 This index was created in the ICL in the 1970s as internal working material for a task which was discontinued; it is preserved as a manuscript without the authors listed and undated.

Using this application, it is possible to acquire information quickly on in which dictionary the searched-for expression appears as a head word. So far, it has been possible by clicking on the abbreviation of the title of a dictionary to gain information on the dictionary work in question, in the future we plan to make it possible to jump directly to the searched-for head word in the text of the dictionary. The database of indexes will gradually be complemented by further lists of entries and in future will become the starting point for the so-called 'lemmarium' as an integral search engine including the lists of entries of material collections, digitalised dictionaries and other, not only lexicographic databases.

The bibliographic database of the DLT, which will also be accessible from the web nest being developed, contains structured bibliographic records related to lexicographic work. It is already functional and fully available for the internal needs of the department (see Fig. 4). Prospectively papers and studies by members of our department will be incorporated into these databases, especially texts related to the research plan, which have been published in less accessible sources (with bibliographic data shown).

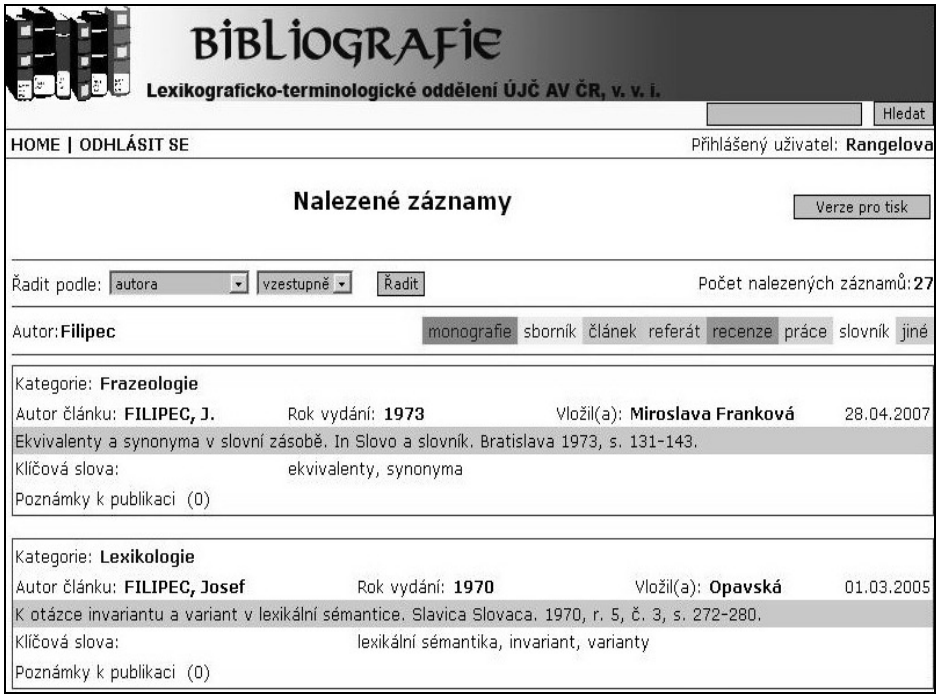


Fig. 4. Bibliographic database with searched-for sample of the records

It is further planned that until it is possible to publish all the material collections of the ICL, samples of the material, both scanned and newly excerpted, will be published at the web nest. These samples will be accompanied by information on the lexical archive of the ICL, or information on the individual collections. It will also be possible to place examples of the new lexicographic description here: in the final phase of the work (in 2010), selected sample entries from L 21 will be placed here which will be prepared for the public and will fulfil the functions of earlier sample issues with the possibility of feedback (by using the section "Write us").

In conclusion, it is still necessary to mention that such a demanding task as the presented research plan is presupposes the creation of appropriate technical and personnel conditions for long-term research activity, therefore the important organisational facets of our work are the building of a research team (engagement and scientific preparation of new colleagues) and the optimal usage of new technological opportunities (the provision of appropriate workstations and server to be used only for the needs of the research plan). Owing to the new technological conditions, it is possible to increase significantly the efficiency of communication and discussion within the working team: the internal communication environment with the working name 'Fórum' is being developed, which will also be used for the recording and resolving of component problems of a conceptual and realisation character.

The institutional research plan *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (2005–2010) is in its complexity directed towards the creation of a comprehensive collection of linguistic, especially lexical data, whose further usage will be of a fundamental scientific-recognition, documentary as well as nationally and culturally prestigious consequence. Within the plan, the methodical, methodological and technological prerequisites of modern research work are being established in the areas of lexicology and lexicography, focused on the preparation of a new monolingual dictionary of the Czech language. The implementation of this research plan will not only bring specialised scientific results (extensive information on Czech lexis) but will also significantly contribute to better familiarity of the public here and abroad on the Czech lexicographic tradition as well as the current research on lexis.

References

1. Akademický slovník cizích slov. Academia, Praha (1995)
2. Český jazykový atlas 1. Academia, Praha (1992)
3. Frekvenční slovník češtiny. Nakladatelství Lidové noviny, Praha (2004)
4. Gebauer, J.: Slovník staročeský. Česká grafická akc. společnost Unie, Praha (1903–1916)
5. Jungmann, J.: Slovník česko-německý. 2nd, unchanged edition. Academia, Praha (1989–1990)

6. Kott, F. Št.: Česko-německý slovník zvláště grammaticko-fraseologický. J. Kolář, V Praze (1878–1893)
7. Martincová, O. et al.: Nová slova v češtině. Slovník neologizmů 1. Academia, Praha (1998) (SN 1)
8. Martincová, O. et al.: Nová slova v češtině. Slovník neologizmů 2. Academia, Praha (2004) (SN 2)
9. Příruční slovník jazyka českého. Státní nakladatelství, Praha (1935–1957) (PSJČ)
10. Slovník spisovné češtiny pro školu a veřejnost. Academia, Praha (1978), 2nd ed. (1994), 3rd ed. 2003 (SSČ)
11. Slovník spisovné češtiny pro školu a veřejnost. Elektronická verze, LEDA spol. s r. o., Praha (2004)
12. Slovník spisovného jazyka českého. Nakladatelství Československé akademie věd, Praha (1960–1971), 2nd ed. Academia, Praha (1989) (SSJČ)
13. Sochová, Zd., Poštolková, B.: Co v slovnících nenajdete. Novinky v současné slovní zásobě. Portál, Praha (1994)
14. Svozilová, N., Prouzová, H., Jirsová, A.: Slovesa pro praxi. Valenční slovník nejčastějších českých sloves. Academia, Praha (1997)
15. Svozilová, N., Prouzová, H., Jirsová, A.: Slovník slovesných, substantivních a adjektivních vazeb a spojení. Academia, Praha (2005)
16. Těšitelová, M., Petr, J., Králík, J.: Retrográdní slovník současné češtiny. Academia, Praha (1986)
17. Velký slovník cizích slov. LEDA spol. s r. o., Praha (2005)

Corpora and text archives

18. *Český národní korpus – SYN2000.* Ústav Českého národního korpusu FF UK, Praha 2000. Available online at: `http://ucnk.ff.cuni.cz`.
19. *Český národní korpus – SYN2005.* Ústav Českého národního korpusu FF UK, Praha 2005. Available online at: `http://ucnk.ff.cuni.cz`.
20. *Český národní korpus – SYN2006PUB.* Ústav Českého národního korpusu FF UK, Praha 2006. Available online at: `http://ucnk.ff.cuni.cz`.
21. Textový archiv Newton Information Technology, s.r.o., `http://www.newtonit.cz` (Newton)

Optimization of Russian Bilingual Dictionaries

Elizaveta Rumyantseva

Moscow State Linguistic University
Moscow, Russia
knabino@gmail.com

Bilingual dictionaries are essential for the work of everybody who has to do with foreign languages – students, teachers, translators, interpreters, travellers and many others. It is clear that a dictionary serves here as a helper in the intercultural dialogue of two or more personalities. Communication is the basis of every human interaction and it consists of several components, according to the theory proposed by B. Gorodetsky. He talks about the following aspects of any "communicative act": communicants, circumstances of communication, system of communicative intentions, communicative processes, and communicative texts [1]. When the interlocutors speak different languages, they will be probably also involved into the communicative process of translation, but the scheme of the interaction in the main remains the same.

Thus, a bilingual dictionary, being a communicative tool, is undoubtedly based on the theory of translation. As the prominent translator V. Komissarov said, translation is "a means to assure a possibility of communication between people speaking different languages" [2]. The possibility of translation roots in the notion "equivalence". It goes without saying that a complete equivalence of linguistic units is impossible, it can be only partial. This fact was pointed out by many prominent linguists and lexicographers, in particular, by Sherba [5]. So the aim of a lexicographer (here he acts as a translator) is to find the closest possible equivalent to a given lexeme.

In our work we do not consider all dictionaries in general, but mainly so-called "active" dictionaries. As we know, there exist active and passive dictionaries (depending on the native language of the users), as well as bidirectional ones which incorporate both types. Active dictionaries are meant for translating from the native language into a foreign one, passive dictionaries are for the opposite task – they give an equivalent in our native language for a lexical unit in a foreign language. The former reflect the active language skills (speaking/writing), the latter – the skills of reading and listening.

Our scientific interest focuses on active dictionaries because it is usually more difficult to find an appropriate equivalent in a foreign language than vice versa. Moreover, an imperceptible mistake in the translation to a foreign language can have serious consequences – so-called "communicative failures". Due to the tangible lack of high-quality active bilingual dictionaries, very few ones meet the requirements of translators and interpreters who seek exact equivalents appropriate to a given context in the broad sense of the word. As in most Russian-foreign dictionaries translation equivalents are simply listed without any differential information,

the choice of an adequate word or expression proves to be a complicated task. We are convinced that a good bilingual dictionary should distinguish among several quasi-synonymic translation versions in the target language.

Nowadays most monolingual English dictionaries are compiled according to a user-friendly conception (i.e. it is easy to use them) and are based on large text corpora (e.g. Longman Dictionary of Contemporary English (LDCE) and Collins COBUILD English Language Dictionary (CCELD), Oxford Advanced Learners' Dictionary (OALD)). We would be happy to implement some of their principles of synonymic differentiation in Russian bilingual dictionaries as well.

We propose a rather detailed description of dictionary compilation process which consists of three stages [4]:

- The elaboration of the dictionary parameters and structure;
- The study of lexemes and compilation of the dictionary;
- The experimental verification of the dictionary.

It is worth mentioning that it is a work scheme and can be modified later.

First of all, a comprehensive study of several contemporary Russian-English dictionaries has been carried out (among them well-known paper-back dictionaries under the editorship of A. Smirnitsky, D. Ermolovich and the latest e-dictionary Lingvo). This analysis served as a base for the elaboration of the Russian-English e-dictionary entry structure paying special attention to the types of information presented there. We maintain that the difference between quasi-synonymic equivalents can be accounted for by the sublanguage to which the units belong. Nevertheless, other factors, such as collocations, style etc., play an important role in their differentiation as well.

We resort to computer interface because it has the advantage of hierarchical information presentation within the dictionary entry, which facilitates the search of the appropriate equivalent. Hilary Nesi, a British lexicographer, recently observed that "computer dictionaries offer three great potential benefits for users: they are quick and easy to use, they can provide access to large amounts of data, and they are interactive" [3].

Having accumulated all the necessary evidence, we started to compile experimental e-dictionary entries for some Russian lexemes using our new principles of information presentation. As an example of an entry for a Russian-English learners' dictionary the most frequent headword in our database of quasi-synonyms "становиться/стать" (meaning 'become') was chosen, and a suggested dictionary entry was compiled in three formats: extended, normal and compact. Further we present the "compact" version of the pilot entry:

ПРЕДЛАГАТЬ, ПРЕДЛОЖИТЬ предложение

= делать, сделать предложение

= вносить, внести предложение

фразы для ситуации ПРЕДЛАГАТЬ

1a. (ДАТЬ ЧТО-ТО) to **offer** sth. (to sb.)/sb. sth. ⇒ an offer

см. также: = *давать, предоставлять, преподносить, протягивать, подавать, вручать, дарить, показывать, предъявлять;*

⇔ *брать, принимать / отказываться*

1б. *(ТОВАРЫ ИЛИ УСЛУГИ)* to **offer** sth. to sb. /sb. sth. ⇒ an offer

см. также: = *продавать, отдавать,*

⇔ *брать, покупать / отказываться*

1в. *(СВОИ УСЛУГИ, СДЕЛАТЬ ЧТО-ТО)* to **offer** to do sth. ⇒ an offer

см. также: = *вызываться, заявлять*

2а. *(ПОПРОСИТЬ)* to **suggest** that sb. do sth. ⇒ a suggestion

to **propose** (a toast, marriage) to sb. *всегда с «to»*

см. также: = *приглашать, просить, уговаривать, агитировать, ходатайствовать, молить, звать,*

⇔ *соглашаться/отказываться*

2б. *(ПРЕДПИСАТЬ)* to **order** sb. (to do sth.) ⇒ an order

см. также: = *требовать, предписывать, приказывать, наказывать, командовать, распоряжаться, велеть,*

⇔ *выполнять / отказываться, не выполнять*

3а. *(НЕМАТЕРИАЛЬНОЕ)* to **propose** sth./that… *офиц.*

(ИДЕЮ, ПЛАН) to **suggest** sth./(that)…/doing sth./wh- ⇒ a suggestion

(ВМЕСТЕ СДЕЛАТЬ ЧТО-ТО)

to **suggest** sth/(that)…/doing sth./wh- ⇒ a suggestion

см. также: = *сообщать, высказывать, вносить, ставить, подавать, намекать, наталкивать, советовать, рекомендовать, убеждать, внушать,*

⇔ *принимать, соглашаться / отклонять, возражать*

3б. *(КАНДИДАТУРУ)* to **suggest** sb. for sth. ⇒ a suggestion

см. также: = *выдвигать, выставлять, рекомендовать, номинировать*

⇔ *выбирать, принять / отклонять*

УПРАЖНЕНИЯ

As you can notice, this entry was meant to serve as a thesaurus for English learners who probably would like to study other synonyms, antonyms, hyponyms, conversives and other related words for the given Russian headword. In a computer dictionary this supplementary information can be hidden so that only those who need it could see it. The electronic format present many other interesting possibilities which we plan to explore in future.

The present research is our first attempt at the optimization of a bilingual dictionary entry and it may have further implications in both theoretical and practical lexicography. It can help lexicographers to compile bilingual dictionaries appropriate to their users for any language pair.

References

1. Gorodetsky B.Y. Kommunikativnye osnovy teorii iazyka. // Metody sovremennoi kommunikacii. – Moscow, 2003: page 88.
2. Komissarov V.N. The Theory of Translation: Linguistic Aspects. – Moscow, 1990: page 37.
3. Nesi H. Dictionaries on Computer: How Different Markets Have Created Different Products
   http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/llc/Conference1998/Papers/Nesi.htm
4. Rumyantseva E.A. Aspects of the compilation of bilingual learners' e-dictionaries.
   http://dialog-21.ru/dialog2006/materials/html/Rumyanceva.htm
5. Shcherba L.V. Slovari s medoticheskoi tochki zrenia // Prepodavanie iazykov v srednei shkole. Obshie voprosy metodiki. – Moscow, 1947: page 91.

Corpus of Spoken Slovak Language

Milan Rusko[1] and Radovan Garabík[2]

[1] Department of Speech analysis and Synthesis, Institute of Informatics,
Slovak Academy of Sciences, Bratislava, Slovakia
milan.rusko@savba.sk
[2] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
http://korpus.juls.savba.sk/

Abstract. In this paper a short description of activities towards building a general speech corpus of spoken Slovak language is given. Different rôles and specific features of text corpus and speech corpus are investigated as well as the most frequent mistakes and misunderstandings of the concept of a speech corpus are mentioned. The concept of a big representative corpus of spoken language and its desired properties are presented. The paper gives an overview of the current state of the art in speech corpora all over the world. It explains the need for a national speech corpus and indicates some of the typical areas of research and applications taking advantage of the existence of such a corpus. The speech databases currently available in Slovakia are listed and the particularities of annotation structures of these databases are pointed out. The authors search for a general annotation structure suitable for the kind of speech corpus envisaged. Some of the basic concepts and technical solutions used in recording and computer aided annotation used for the existing speech corpora are described. The most significant problems standing in the way of building a big speech corpus are pointed out. Furthermore, a pilot version of a speech corpus is presented, containing several recordings and their orthographic transcription.

Keywords: *speech corpus, database, spoken speech, Slovak.*

1 Introduction

Speech corpora play an irreplaceable rôle in present-day automatic speech processing research and development. The information obtained from speech corpora and databases is used for building acoustic models for speech recognition, language models for natural language processing, dialogue models for dialogue management in human-machine interaction and many other purposes. Special speech databases are being built for "unit selection" or "corpus based" speech synthesizers. Every database is built for its particular purpose and is therefore application specific with regards to the choice of speech material and annotation aimed at covering the needs of the actual application.

It would certainly be helpful to have a general speech corpus available for the Slovak language that would allow for broad research in many scientific

areas ranging from linguistics, stylistic analysis, research of dialects, phonetics, phonology, from speech communication to extralinguistics, vocalics and speech acoustics. A pilot version of such a speech corpus, which could be considered as a statistically representative sample of the spoken speech communication in Slovakia is being prepared at the Slovak National Corpus department[1] of the Ľ. Štúr Institute of Linguistics, in collaboration with the Department of Speech analysis and Synthesis at the Institute of Informatics of the Slovak Academy of Sciences. The aim of the pilot version is to investigate the principal ways of building a spoken corpus, consider different possibilities for a transcription and query mechanism and prepare the way for a big, representative corpus. According to its expected volume and diversity of speech material the final corpus has to be collected with the mutual cooperation of several institutions. The benefit of having such a corpus available would be extraordinarily big not only for theoretical research, but also for commercial application development as well. The cultural consequences are not negligible either, since language represents a substantial part of national culture.

2 "Corpus" versus "database"

In principle, any collection of more than one text can be called a corpus – corpus being the Latin expression for "body", hence a corpus is any body of text. But the term "corpus" when used in the context of modern linguistics most frequently tends to have more specific connotations than this simple definition.

According to McEnery and Wilson [2]

> "the following list describes the four main characteristics of the modern corpus: sampling and representativeness, finite size, machine-readable form, and standard reference". Scientists are therefore interested in creating a corpus which is maximally representative of the variety under examination, that is, which provides them with an as accurate a picture as possible of the tendencies of that variety, as well as their proportions. The corpus should contain a broad range of speakers and genres which, when taken together, may be considered to "average out" and provide a reasonably accurate picture of the entire language population.
>
> The term "corpus" also implies a body of text of finite size, but this property does not have universal validity – it is possible to create a monitor corpus. This "collection of texts" is an open-ended entity – texts are constantly being added to it, so it gets bigger and bigger. The main advantages of monitor corpora are: dynamic nature – new texts can always be added, unlike the synchronic "snapshot" provided by finite corpora; and wider scope – they provide for a large and broad sample of language.

> Their main disadvantage is that they are not such a reliable source of quantitative data (as opposed to qualitative data) because they are constantly changing in size and are less rigorously sampled than finite corpora. [2]

(We prefer a national speech corpus to be open as to reflect the newest tendencies in Slovak speech communication.)

According to Sinclair [3] a (text) corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. A computer corpus is a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance. A corpus can be divided into subcorpora. A subcorpus has all the properties of a corpus but happens to be part of a larger corpus. Corpora and subcorpora are divided into components. A component is not necessarily an adequate sample of a language and in that way is distinct from a corpus and a subcorpus. It is a collection of pieces of language that are selected and ordered according to a set of linguistic criteria that serve to characterize its linguistic homogeneity. While a corpus may illustrate heterogeneity, and also a subcorpus to some extent, the component illustrates a particular type of language.

The term annotated corpus is used for any corpus which includes codes that record extra information. (We think that according to this definition the existing Slovak speech databases can be considered as specialized satellite components of the future general speech corpus.)

Campbell has published a practical definition (coming out of several older definitions) explaining the difference between a database and a corpus [4] :

> A "database" is an organized collection of information, typically designed for ease of retrieval by computerized methods; a "corpus", on the other hand, is a collection of naturally-occurring spoken or written material in machine-readable form, that are in themselves more-or-less representative of a language for the systematic study of authentic examples of language in use. The important difference is that while both comprise an accumulation or assemblage of texts or recordings which can be considered as representative of a genre, the former is usually "constructed", and the latter "obtained". More specifically, a database is purpose-built; a store of information which is structured from the beginning, while a corpus is a body of information from which knowledge can be derived.

3 "Text corpus" versus "speech corpus"

In some countries the first attempt to build a general spoken language corpus was made by linguists who had experience in collecting and text corpora or by people from the speech processing community who had been involved in speech database construction. Therefore in some cases the speech corpus was treated very similarly to a text corpus supplemented with an "audio version" of the

text included in the corpus. The non-verbal cues or even prosody and other important information were omitted. The annotation then consists only of an orthographic transcription, some basic data about the identity of speaker and the situation when the speech was recorded.

Exaggerating a bit, one could say that a user of such a corpus finds himself in a position similar to that of patient with aprosodia – an inability to comprehend (or articulate) emotional voice tones and miss the affective or "feeling" content of speech. But the speech corpus offers a wide scale of information on different aspects of human communication, which should not be restricted to the textual and linguistic content.

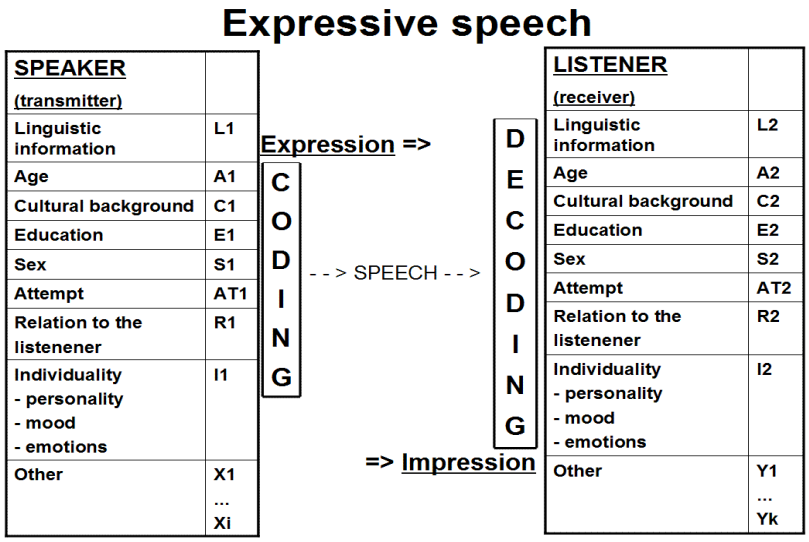


Fig. 1. Simplified scheme of transmission of various information from a speaker to a listener. Every part of the information carried by the speech signal can be an object of research and can be important for applications and should be therefore (at least partly) annotated in the speech corpus.

The corpus should be open to a broad scientific and public community, to allow for the novelty of previously unconsidered usage of the data. As Bird & Liberman say "Once created, a linguistic database may subsequently be used for a variety of unforeseen purposes, both inside and outside the community that created it." [5]

From an acoustical point of view, speech uses only several acoustic quantities (fundamental frequency, time duration of phonetic elements and pauses, intensity of acoustic pressure and frequency spectrum) to carry diverse information not only on the linguistic content, but also on the speaker and communication situation.

Pointing out bad practices in speech corpora building Campbell says [4] "when designing speech databases, care is usually taken to exclude all inarticulate prosody, since it is associated with "ill-formed" speech". (We agree, that the speech is not ill-formed, but our knowledge is still insufficient and the models we have developed are not able to model the natural speech communication correctly.)

A segment in spoken language is an individual consonant, vowel, tone, or stress that makes up a word. An utterance is made up of both segments and supra-segmental features. These are broadly divided up into prosody and paralinguistics. Prosody refers to pitch, loudness, duration, intonation and tempo. Paralinguistics, which is much more difficult to measure, refers to the expression of speaker characteristics, individuality (personality, mood and emotion) – the speaker's attempt and his relationship to the listener. These nonverbal or suprasegmental elements of a speech utterance constitute a significant part of its meaning. The nonverbal cues of the voice are the object of study of vocalics.

The speech corpus should therefore contain different information and various levels of annotation, such as:

- sound file properties (name, description, format, recording conditions, copyright, etc.)
- linguistic information (various transcriptions, linguistic annotation – morphological tags, part of speech tags, syntax, semantic annotation, prosody annotation, etc.)
- extralinguistic information (dialogue and communicative acts annotation, voice quality, pauses, fillers, disfluences, elements specifying background noise and signal quality etc.)

4 General and representative corpus of spoken language

Several attempts have been made to design a relatively general and representative corpus for many terrestrial (and even extraterrestrial[6]) languages – mainly for the "big ones", like English, American, Chinese, Japanese, Spanish, French, Korean, but also for Polish, Irish, Scottish (Gaelic), Czech, Croatian and others. For illustration we will mention some details on some of them.

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20^{th} century, both spoken and written. The spoken part includes a large amount of unscripted informal conversation, recorded by volunteers selected from different age, region and social classes in a demographically balanced way, together with spoken language collected in all kinds of different contexts, ranging from formal business or government meetings to radio shows and phone-ins to broadcast news and conversational telephone speech [7].

There are two parts to the 10-million word spoken corpus: a demographic part and a context-governed part.

The Demographic part of the Spoken Corpus was recorded by 124 volunteers from different social groups. They were male and female volunteers from a wide range of ages, and they lived at 38 different locations across the UK. Recruits used a personal stereo to record all their conversations unobtrusively over two or three days, and logged details of each conversation in a special notebook. Those who took part in the recordings were asked after the conversation to give permission for their speech to be included in the corpus. Information about the participants, such as age, sex, accent and occupation, was recorded when available.

The Context-Governed part of the Spoken Corpus was created with the intention to collect roughly equal quantities of speech recorded in each of the following four broad categories of social context:

- Educational and informative events (lectures, news broadcasts, classroom discussion, tutorials)
- Business events (sales demonstrations, trades union meetings, consultations, interviews)
- Institutional and public events (sermons, political speeches, council meetings, parliamentary proceedings)
- Leisure events (sports commentaries, after-dinner speeches, club meetings, radio phone-ins.)

The Spoken Language Corpus of Swedish at Göteborg University, which is general and covers the whole of Sweden (although it is not called "national"), is an incrementally growing corpus of spoken language samples from several languages which presently consists of 1.26 million words from about 25 different social activities. Because spoken language varies considerably in different social activities with regard to pronunciation, vocabulary, grammar and communicative functions, the goal of the corpus is to include spoken language from as many social activities as possible in order to facilitate research that will provide a more complete understanding of the rôle of language and communication in human social life [8].

The recording facilities covered are: auctions, bus driver/passenger conversation, court, dinner, discussion, factory conversation, formal meeting, hotel, informal conversation, information, service (phone), interview, lecture, market, medical consultation, religious service, retelling of article, rôle play, shop, task-oriented dialogue, therapy, trade fair, travel agency.

The Czech National Corpus has several projects of spoken corpora available [9] – the Prague Spoken Corpus (PMK), the Brno Spoken Corpus (BMK) and ORAL2006.

The PMK was collected during the years 1988–1996 and was the first available corpus of spoken Czech language. The audio recordings were taken in the city of Prague and surroundings, and the corpus was designed to contain four main sociolinguistic variables – speaker's sex, age, education and discourse type, and for simplicity all divided into two sets (man/woman; under 35/over 35 years; less than university/university education; formal speech/informal speech). The corpus contains 674 992 words and is available only in the form of transcribed text. The BMK was collected during the years 1994–1999 in the city of Brno, following the same structure as the PMK.

The most recent ORAL2006 tries to get recordings from the whole area of Bohemia, divided into four main regions. The sociolinguistic distribution of the recordings is kept balanced according to the speaker's age, sex and education, and less to the region of origin. The corpus contains recordings of 754 persons, amounting to 1 312 282 tokens of transcribed text.

4.1 Available speech databases in Slovak

The first professional speech database in Slovak was *SpeechDat-E SK* [10], following the SpeechDat specification [11] and having recordings from 1000 speakers. In spite of the fact that this database is specialized for training and testing speech recognizers in teleservices, it contains phonetically rich sentences which can be used for some purposes in speech research [12].

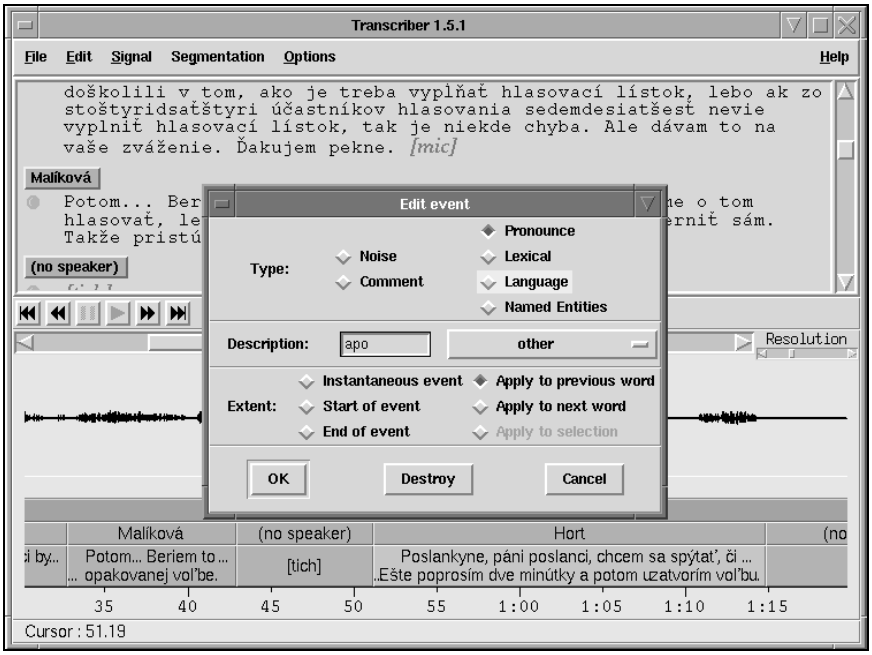


Fig. 2. Annotation of parliament proceedings in Transcriber [15].

MobilDat-SK, which was developed in a frame of the IRKR project [13] is a mobile phone counterpart of SpeechDat with 1100 speakers. Moreover this database contains an unprompted item, where every speaker answers to one of a set of simple questions (How do you get from your house to the closest supermarket? How do you cook scrambled eggs? etc.)

The *TV news audiovisual database* is being built at Technical University Košice for the purpose of experiments in speech recognition, which should have an application in automatic TV news subtitling [14].

The *TV debates* (e.g. "Pod lampou") *audiovisual database* is being built at Technical University Košice for the purpose of experiments in dialogue modeling and expressive speech recognition, which should have an application in automatic TV program subtitling.

The *Parliament proceedings audiovisual database* is being built at the Institute of Informatics, Slovak Academy of Sciences for the purpose of experiments in speech recognition, which should have an application in automatic Parliament proceedings transcription.

SyntDat – a speech synthesis database designed for unit selection speech synthesis (used in Kempelen 2.0 to 2.2 synthesizers) [16].

5 Some controversies

Discourse markers, that have more or less generally accepted transcription in English e.g. sounds representing backchannels and minimal positive feedback (yes, yeah, yah, okay, mhm, hm, aha, uhu), negative minimal feedback (no, n-n, uh-uh), hesitation (er, erm), exclamations – joy/enthusiasm (yay, yippee, whoohoo, mm:), questioning/doubt/disbelief (haeh), astonishment/surprise (a:h, o:h. wow, poah), apology (oops), disregard/dismissal/contempt (ts, pf), exhaustion (ooph), pain (ouch, ow), requesting silence (sh, psh), anticipating trouble (oh-o:h) etc. are still waiting to be get a standardised transcription in Slovak.

We have no experience with transcription of onomatopoeic noises. Intonation modelling needs a generally accepted annotation scheme which still does not exist although the first attempt towards the definition of Slovak ToBI has already been made [17].

We have no annotation scheme for many supralinguistic and extralinguistic phenomena (e.g. emphasis, voice quality and many others).

If we accept a grapheme to be the smallest element of written text, it would be reasonable to define a phoneme to be the smallest element annotated in a speech corpus. This means that speech recognition technology in Slovak capable of finding phoneme boundaries with acceptable reliability would be needed. For pitch contour and voice quality measurement we often need pitch

marks. Their determination is not language dependent, but reliable pitch marking is still a difficult task.

6 Obtaining Slovak speech recordings

Apart from recordings originating in the specialized databases mentioned earlier, a large part of our proposed corpus will consist of recordings obtained on purpose. The main sociolinguistic data observed will be speaker's sex, age, education, discourse type, conformance with the standard language and region of origin (inspired by the Czech spoken corpora). Although there is a huge potential in spoken corpora for dialect studies, our corpus will focus (at least from the beginning) on standard Slovak. Therefore the recordings will be made primarily in urban areas.

7 Corpus manager

There are several requirements for the corpus query possibilities, each targeting a different end-user group. On one hand, we want a powerful tool for working on the transcribed text, for statistical analysis on the various aspects of the data. This is easily achieved by a standard corpus manager interface, offering all the usual functions for the transcribed text. However, the existing text corpus managers offer no easy possibilities of linking with the specific sound data – this is not necessarily an insurmountable disadvantage per se, because any serious research on the acoustic level will be supposedly performed with rather specialized tools and for specific purposes, and it is not quite feasible trying to accommodate all the possible uses.

The corpus also has to be usable for casual users, without the need to install specialized client software and to study the (often complicated) program controls. Following the ubiquity of web applications, it is obvious that the corpus should be accessible through a simple WWW interface, with a possibility for the user to directly access the relevant sound sample. These two approaches are not exclusive, there is no reason not to provide both possibilities (in fact, a similar system was deployed in the (text) Slovak National Corpus).

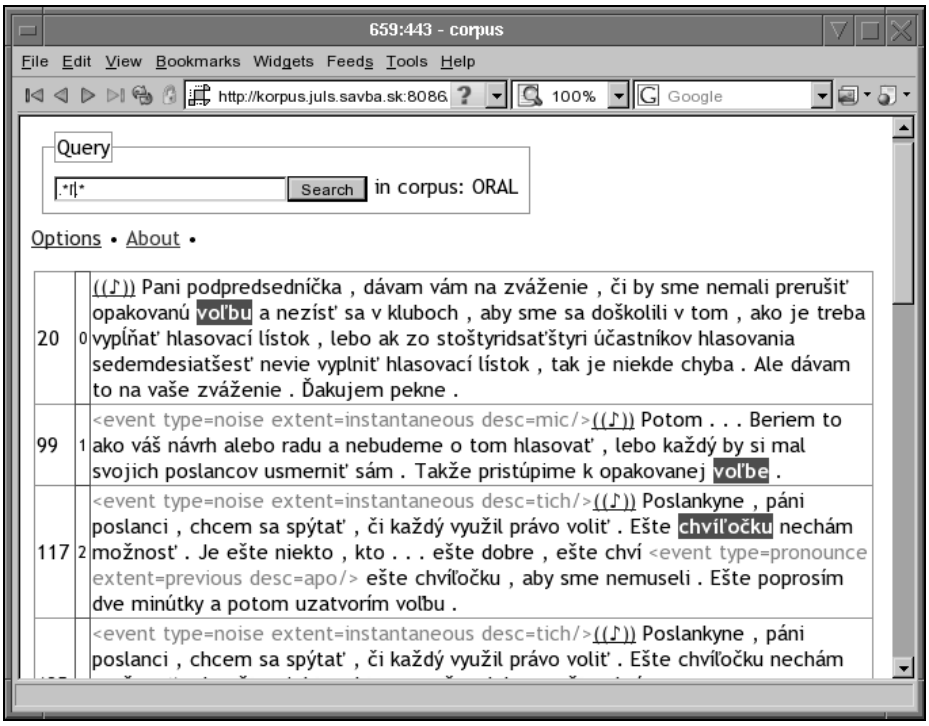


Fig. 3. Speech corpus web interface

8 File formats

For transcription, we are using the *transcriber* [15] software, which allows the annotators to define speakers' identities, define various types of extralingual events and speech phenomena and seamlessly integrate the audio and text data. The transcriber stores its data in a native XML format with links to the audio files and timestamps at synchronization points. We take advantage of this format and use the corpus manager to index the (postprocessed) XML files directly.

There are two conflicting requirements for the audio file format – the first is to maximize the sound quality, the second is to minimize the file size. Given the expected longevity of the spoken corpus and the ever-decreasing cost of storage media, sacrificing quality for the sake of saved disk space is not applicable anymore. This holds even in the case of inaudible quality degradation (using a high bitrate lossy compression protocol). Therefore it is desirable to

archive the original audio data in either an original format, or by using only lossless compression. On the other hand, there are uses for the corpus requiring only access to speech without very noticeable distortion, e.g. demonstration to casual users or as a part of a foreign language instruction process. For web-based services, the size of the transmitted files is important, as well as use of a common multi-platform format, not requiring installation of specialised software.

For the original format, we decided to use the FLAC lossless codec[18], giving a compression ratio of about 50 per cent compared to uncompressed PCM data (more for stereo input). Unfortunately, most modern budget dictaphones use proprietary WMA[1] or DSS formats, which are already lossy compressed. Therefore we expect some of the audio records in the corpus obtained from external sources to be in the WMA format (there is a lack of relevant software and tools needed for DSS format processing and conversion), which can potentially preclude the usage of the data for some specialised purposes, since the sound is already mapped to a psychoacoustic model – primarily, the corpus would not be usable for the development of new psychoacoustic models. However, when keeping the quality at a sufficiently high level, even frequency analysis as required by phoneticians is applicable.

For the format presented to users, we decided to use the lowest compression quality (bitrate) that gives only slight perception of quality distortion.

We used primarily the SPEEX codec[19]. SPEEX was designed specifically for speech encoding at lower bitrates, and gives an excellent compression ratio. Another advantage is a special decoder mode enhancing perceived sound quality (we found that sometimes the SPEEX encoded data sound subjectively better than the original). Before encoding, the sound samples were downmixed to one mono channel and downsampled to ultra-wideband frequency (32 000 Hz, one of the recommended sampling rates for the SPEEX codec). The files were encoded using variable bitrate encoding, encoding complexity 10, at quality 6, which gives an average bitrate of 23 kb/s.

Because of a rather lesser SPEEX penetration to the usual desktop PC systems, we decided to offer Ogg/Vorbis[20] as an alternative (downmixing to single channel, but without resampling, since the Vorbis codec does not have strict recommendation as per the sampling frequency, and resampling often makes the audio sound subjectively worse compared to SPEEX). We used the experimental aoTuV encoder[21] optimized for lower bitrates. Encoding was done at quality -1, giving an average bitrate of 40 kb/s.

Users can therefore choose between SPEEX, Ogg/Vorbis and original (or FLAC) format. There is also a Java applet available, playing SPEEX format for users unable or unwilling to install the required codecs.

---

1 We are using the general name WMA here, although technically WMA can mean several different incompatible codecs (WMA, WMA Pro, WMA Lossless, or WMA Voice).

9 Levels of transcription

Different levels of transcription are possible, each of them putting different strain on the annotation process. In our corpus project, we selected three different levels – orthographic, phonetic/phonemic and suprasegmental transcription.

9.1 Orthographic transcription

Orthographic transcription is the most straightforward, and the basic type of annotation that distinguishes a simple collection of recordings from a speech corpus. We decided the orthographic transcription in our corpus should follow the standard Slovak orthography, transcribing only the differences from standard Slovak pronunciation as an additional word attribute. This both makes the transcription easier to read as well as allowing us to deploy usual NLP tools (e.g. morphology analysis, lemmatization). In some areas, we follow standard Slovak pronunciation, as opposed to the prescribed official one. In particular, the pronunciation of letter *ä* as /ɛ/ does not warrant specific transcription, but its pronunciation as /æ/ does. Similarly, pronouncing the syllables *le*, *li* and *lí* as /lɛ/, /li/ and /li:/ is not marked, but palatalized pronunciation /ʎɛ/, /ʎi/ and /ʎi:/ is. Even though officially correct, it has for all practical purposes disappeared from standard Slovak.

Although tempting, we have chosen not to use the standard punctuation symbols to denote extralingual information (such as pauses and hiatus in speech), since human annotators are prone to unconsciously deploying such marks where orthography rules require, not where the phenomena really occur. We are using specific annotation software possibilities instead, with usual punctuation marks (comma, colon, exclamation mark etc.) being at the annotator's discretion. For the same reasons, we are not using capital letters for any special purpose, the annotators can capitalize words as they feel natural. We recommend putting the dot at the end of sentences as dictated by the logical flow of the document (not by pauses in discourse), the sole purpose of this is to help the automatized analysis tools (in particular morphology analyzer), where marking the end of sentences sometimes improves the processing accuracy.

9.2 Phonemic/phonetic transcription

Phonetic transcription is useful for speech recognition, speech synthesis and basic linguistic research. However, making a correct phonetic transcription requires trained annotators with a good knowledge of language phonetics and is rather time consuming and sometimes controversial. Therefore we decided to include phonemic transcription, with just some phonetic features

(distinguishing several most frequent allophones). This requires designing a general model of phonemic analysis of the Slovak language usable for the transcription process – to our knowledge, no such analysis universally accepted among Slovak linguists exists so far. Only a part of the corpus will be manually transcribed phonetically (in addition to the orthographic transcription). For he rest of the corpus, an automatic grapheme-to-phoneme conversion will be available.

9.3 Suprasegmental annotation

A suprasegmental annotation scheme must provide a mechanism for indicating suprasegmental structure such as word/syllable boundaries and stress markings. The specification may address other types of suprasegmental structure.
A different phonological intonation annotation scheme is needed for every particular language. Inspired by the successful ToBI (Tones and Break Indices) for American English [22] the intonation annotation scheme Sk-ToBI was introduced for Slovak [17]. ToBI annotation by hand is extremely time consuming, therefore only a limited part of the corpus will be annotated manually. This can later serve for training automatic annotation algorithms.

10 Copyright issues

It can be argued that recorded "natural" speech is not protected by the Slovak Republic copyright law (the law is not very clear about the issue). However, the recordings cannot be distributed without consent from the author, as long as there are any data from which the author's identity can be inferred, and according to the current laws it is nearly impossible to legally record somebody without informing him in advance. This means that we are unable to get recordings of really natural speech, and the representative part of the corpus has to be recorded in other ways – e.g. masking the recording as sociological research or public opinion poll, so that the recorded subjects are not aware of the linguistic nature of the recordings. Even so, we cannot expect to obtain spontaneous natural speech.

11 Conclusion

In spite of the fact that we aware of the complexity and resource cost of building a general and representative speech corpus in Slovak we believe that Slovak linguists and speech researchers will proceed in a common effort towards a Slovak speech corpus that could be included in the Slovak National Corpus, as it is common in the leading corpora in the world.

References

1. Slovak National Corpus. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2006. Available from WWW: http://korpus.juls.savba.sk/
2. McEnery, T., Wilson, A., "Part Two: What is a Corpus, and what is in it?" Web pages to be used to supplement the book "Corpus Linguistics" published by Edinburgh University Press, ISBN: 0-7486-0808-7. http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fra1.htm
3. Sinclair, J., School of English, University of Birmingham http://www.ilc.cnr.it/EAGLES96/corpustyp/node5.html
4. Campbell N.: Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language, Journal Language Resources and Evaluation, Issue Volume 39, Number 1 / February, 2005, Publisher Springer Netherlands, pp. 109-118.
5. Bird St. and Liberman M., A Formal Framework for Linguistic Annotation. Speech Communication, 33 (1,2), pp 23–60, 2001.
6. Corpus of Spoken Martian. http://www.elsnet.org/nps/0014.html
7. British National Corpus, http://www.natcorp.ox.ac.uk/
8. Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E. and Ottesjö, C. (2000, December). The Spoken Language Corpus at the Department of Linguistics, Göteborg University [55 paragraphs]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [Online Journal], 1(3). Available at: http://www.qualitative-research.net/fqs-texte/3-00/3-00allwoodetal-e.htm [Date of Access: 20th of May 2007].
9. Waclawičová, M.: Mluvené korpusy v ČNK: několik poznámek k mluveným projevům a polyfunkčním výrazům. In: Korpusová lingvistika: Stav a modelové přístupy. Studie z korpusové lingvistiky, sv. 1. Eds. F. Čermák, R. Blatná. NLN a ÚČNK, Praha, 2006. P. 347–358.
10. Heuvel, H., Boudy, J., Bakcsi, Z., Černocký, J., Galunov, V., Kochanina, J., Majewski, W., Pollak, P., Rusko, M., Sadowski, J., Staroniewicz, P., Tropf, H. S.: Five Eastern Europen Speech Databases for Voice-Operated Teleservices Completed. In: Eurospeech 2001 – Aalborg, Denmark, 2001.
11. http://www.speechdat.org
12. Rusko, M., Daržagín, S., Trnka, M.: SpeechDat-E telephone speech database as an important source for basic acoustic-phonetic research in Slovak. In: Proceedings of the International Congress on Acoustics, ICA 2004, Kyoto, Japan, part I. p. II-1676 – II-1682. ISBN 4-9901915-6-0.
13. Juhár J., Ondáš S., Čizmár A., Rusko M., Rozinaj G., Jarina R.: "Development of Slovak GALAXY/VoiceXML Based Spoken Language Dialogue System to Retrieve Information from the Internet", Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh, Pennsylvania, USA, 2006. ISSN 1990-9772, pp. 485–488.

14. Pleva,M., Juhár, J., Čižmár, A.: Vývoj a evaluácia multilingválnej databázy pre systémy automatickej transkripcie správ elektronických médií. About development and evaluation of multilingual database for automatic broadcast news transcription systems. Acta Electrotechnica et Informatica, Vol.4, No.2, 2004, ISSN 1335-8243, pp.56-59.
15. Barras, C., Geoffrois, E., Wu, Z., Liberman, M., 1998. Transcriber: a free Tool for Segmenting, Labeling and Transcribing Speech. In: Proc. First Int. Conf. on Language Resources and Evaluation (LREC 98), Granada, Spain, pp.1373–1376.
16. Rusko, M., Daržagín, S., Trnka, M., Cerňak M.: Slovak Speech Database for Experiments and Application Building in Unit-Selection Speech Synthesis. In: Proceedings of Text, Speech and Dialogue, TSD 2004, Brno, Czech Republic, pp. 457 – 464.
17. Rusko, M., Sabo, R., Dzúr, M.,: Sk.ToBI Scheme for Phonological Prosody Annotation in Slovak, in: Lecture Notes in Artificial Intelligence 4629, Springer Verlag, 2007, pp. 334–341, ISBN 978-3-540-74627-0.
18. http://flac.sourceforge.net/
19. http://www.speex.org/
20. http://xiph.org/vorbis/
21. http://www.geocities.jp/aoyoume/aotuv/
22. Silverman, M. et al.: ToBI: A standard for labeling English prosody, Proceedings of the 2nd International Conference of Spoken Language Processing, Banff (1992), 867–870.

Prosody Annotation in Slovak Using Sk-ToBI

Milan Rusko[1], Róbert Sabo[1] and Martin Dzúr[2]

[1] Department of Speech analysis and Synthesis of Institute of Informatics of Slovak Academy of Science, Dúbravská cesta 9, 845 07 Bratislava, Slovakia
[2] Departmant of Slovak Literature and Literary Science of Philosophical faculty of Comenius University, Gondova 2, 818 01 Bratislava, Slovakia
[1]{milan.rusko,robert.sabo}@savba.sk, [2]martindzur@gmail.com

Abstract. Research and development in speech synthesis and recognition calls for a phonological intonation annotation scheme for the particular language. Inspired by the successful ToBI (Tones and Break Indices) for American English [1] and GToBI [2] for German, this paper introduces a new intonation annotation scheme for Slovak, Sk-ToBI. In spite of the fact that Slovak prosodic rules differ from those of English or German, we decided to follow the main principals of ToBI and to define a special Slovak version of Tones and Break Indices annotation scheme. The suitability of the proposed annotation system was tested on the studio-recorded intonation speech database as well as on live recordings of a puppet theatre actor. The resulting conclusions show a way to further improvement of the Sk-ToBI annotation conventions. This paper is also meant as the first introduction of this new model to the international linguist community.

Keywords: *ToBI, Slovak, prosody, speech, intonation, pitch accent, annotation.*

1 Motivation

Prosody analysis and processing represent an inevitable part of current automatic speech processing systems. A phonological intonation annotation scheme is needed for this purpose. As there was no such scheme available for Slovak, we decided to create a system for intonation labeling based on the ideas of the ToBI annotation.

The most common convention for prosody labeling is ToBI (Tones and Break Indices) that was set up by a team of American researchers on the basis of the Pierrehumbert's model of intonation and presented in 1992 [3]. Although Slovak does not have the same prosodic features as English, we have drawn inspiration from this convention as far as the use of its labels and tiers is concerned but we have adapted it to Slovak prosody. We had to determine basic features and rules of the Slovak intonation, which enabled us to create a set of essential types of pitch accents and their combinations that can be found in the spoken form of this language [4].

Slovak is a Slavic language. It is a stress language with fixed stress on the first syllable. As far as we know this is the first attempt to create a phonological prosody annotation scheme for this language.

There are various tonal realizations of utterances in Slovak. We will outline especially the ones belonging to the neutral style of standard Slovak, and thus to create a basis for setting up intonation labeling conventions that we have called Slovak Tones and Break Indices (Sk-ToBI).

2 Sk-ToBI annotation scheme

2.1 Break index tier

Break indices are used to rate the degree of juncture between words and between the final word and the silence. Similarly to TOBI we use in the Slovak system indices 0-4 and signs „?“ and „-“ for questionable parts [9,10,4].

2.2 Tone tier

To mark the tones and pitch accents we have adopted the signs used in English ToBI. We have excluded bitonal accents H+!H* and L+H* and also H+L* (used in GToBI) from the Sk-ToBI system:

Single tone accents

*H** high pitch accent,

*L** low pitch accent,

*!H** an accent pitched approximately in the middle of the range of the melodic contour (this accent can follow after the same accent (*!H**) or after the high accent (*H**).

Bitonal accent

L+H* low pitch accent with raise to high target after accented syllable.

Boundary tones

%H onset at the beginning of the speaker's utterance with a very high pitch,

H- ending of the intermediate phrase with a high low pitch (before a break of type „3“or with combination with final boundary (H-H% or H-L%) tone before „4“) ,

L- ending of the intermediate phrase with a pitch (before a break of type „3“ or with combination with final boundary (L-L% or L-H%) tone before „4“),

!H- ending of the intermediate phrase approximately in the middle of the utterance pitch contour boundaries (before a break index „3“)

H% ending of an intonation phrase with high pitch – anticadence (break index „4“),

L% ending of the intonation phrase with low pitch – conclusive cadence (before a break index „4“).

Uncertanties. Similarly to the break index tier, the annotator can use in this tier the mark „?“. When he cannot exactly determine the tone „H” or „L“, the tone should be marked as „X?*“, „X?-“, „X?%“ or *? as the transcriber is not certain even there is a pitch accent.

2.3 Possible realizations of pitch contours in Slovak

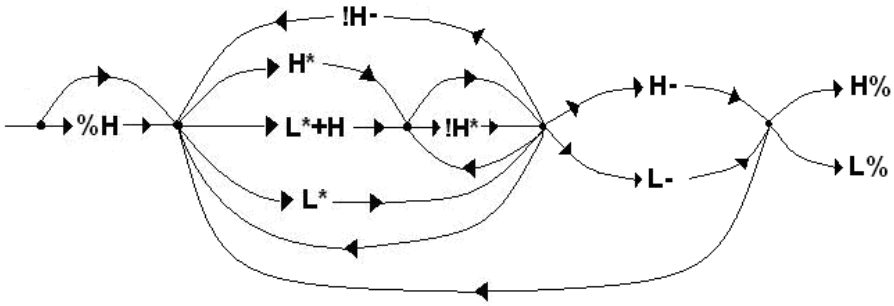


Fig. 1. The diagram of possible sequences of tones and pitch accents in the Slovak sentence

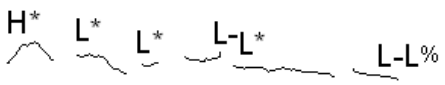
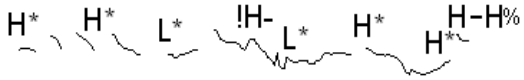
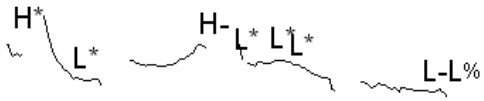
L-	Low pitched semicadence – typical for paratactic clauses.	*Ležím s knihou v posteli a nezunuje sa mi.*  (I lie in a bed with a book and it doesn't bother me.)
!H-	Semicadence placed within the middle pitch range – typical for the most of Slovak hypotactic clauses but also within longer simple sentences.	*Chcete sa s ním stretnúť a dať mu ten dar?*  (Do you want to meet him and give him that gift?)
H-	High pitched semicadence – typically occurs in complex questions and in some types of hypotactic clauses (eg. disjunctive clauses).	*Už si ho nevšímala, ani mu neodpovedala.*  (She was no longer noticing him nor she was answering to him.)

Table 1. Three possible types of semicadence at the end of intermediate phrase

Ends of intermediate phrases are traditionally labelled within English, German and other ToBI systems by a set of symbols (L-, H-) and ends of intonation phrases by (L%, H%). The first group is applied to label the tone pitch immediately before the end of the intermediate phrase and the second one indicates the end of utterance with a conclusive cadence (falling intonation) or anticadence (rising intonation).

The course of intonation contour at the end of the intermediate phrase (index type „3“) is called semicadence, and the three typical boundary tones it can reach are shown in the examples in Table 1.

As far as the end of intonation phrase is concerned, the course of intonation contour at its end is cadence or anticadence and the four typical boundary tone combinations it can reach are shown in the examples in Table 2.

Cases such as H-H% and L-H% should refer to anticadence and L-L% and H-L% should refer to conclusive cadence.

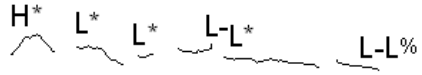
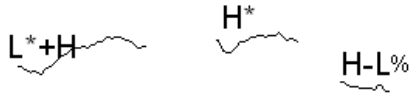
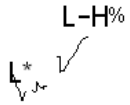
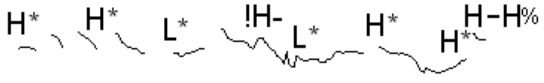
L-L%	End of the utterance by conclusive cadence (typical for declarative sentences).	Ležím s knihou v posteli a nezunuje sa mi.  (I lie in a bed with a book and it doesn't bother me.)
H-L%	High pitched end of the utterance with conclusive cadence (typical for declarative sentences, or exclamatory sentences).	*V y n e s s m e t i!*  (Empty the trash bin!)
L-H%	End of the utterance with low placed pitch with anticadence. (e.g. yes – no question).	Neprišiel?  (Hasn't he come?)
H-H%	End of the utterance with high pitch and anticadence (e.g. yes–no question).	*Chcete sa s ním stretnúť a dať mu ten dar?*  (Do you want to meet him and give him that gift?)

Table 2. Four types of realization of the end of the intonation phrase.

3 Speech material

As a basic source of knowledge for our study we used the theoretical works on the Slovak phonology (e.g. [5] and [6]). This study was then followed by a research on the recorded speech material. Our speech sources consisted of several speech databases:

1. „intonation part" of the speech synthesis database which was designed to cover all the basic types of sentence intonation contours in Slovak [7] (artificial, read material),
2. database of recorded puppet plays of a traditional puppeteer Bohuslav Anderle (artistic style with very expressive speech).

Speech material from each of the databases was listened through by two researchers with degree in linguistics. Every utterance, that seemed to contain new unseen features, was analysed in detail and checked for suitability of Sk-ToBI for its annotation.

3.1 Analysis of the speech material using Sk-ToBI

The set of marks developed for neutral speaking style was shown as not completely sufficient for expressive speech used in recordings of a puppeteer theatre actor. A wide pitch range is typical for such recordings. Read sentences had range approximately 100 Hz (see Fig. 2) and the puppeteer approximately 300 Hz (see. Fig 3).

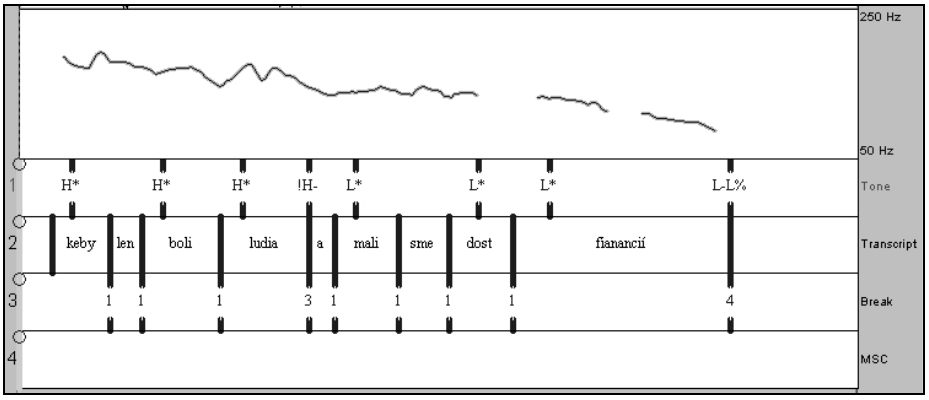


Fig. 2. Utterance from corpus of read sentences
"*Keby len boli ľudia a mali sme dosť financií.*"

Figure 1 shows the pitch contour for the utterance of Gašparko (Mr. Punch): „*...slovo. Hi! Čertíčku nie, prosím ťa nie, nie! Juj! Čert, ja sa zabijem!*". In this example we can see the high difference in pitch between the end of previous utterance (...slovo) at 351Hz and the last utterance (Čert, ja sa zabijem) at 632

Hz. By standard marks L* and H* we are not able to record this extreme difference.

Expressive speech of characters (e.g. rising fear of Gašparko) causes pitch changes in wide range and some additional information is needed in such cases. This information can be included in miscellaneous "MSC" tier. Therefore it is necessary to define a set of marks describing type of event which influences the prosody. Besides typical marks as laugh, breath, cough we propose to mark events of an emotional state like anger, fear, surprise, enjoyment, etc. Unlike in English ToBI, we allow using nested labels[1] in our conception.

Also we recommend labeling the stressed words because the semantic emphasis markedly influences prosody. This information could be very useful for further processing of annotated speech material.

The set of marks of this tier should be open for possible broadening in a case of need.

In expressive utterances it is often appropriate to use the mark %H which labels high placed pitch at the beginning of the speaker's utterance.

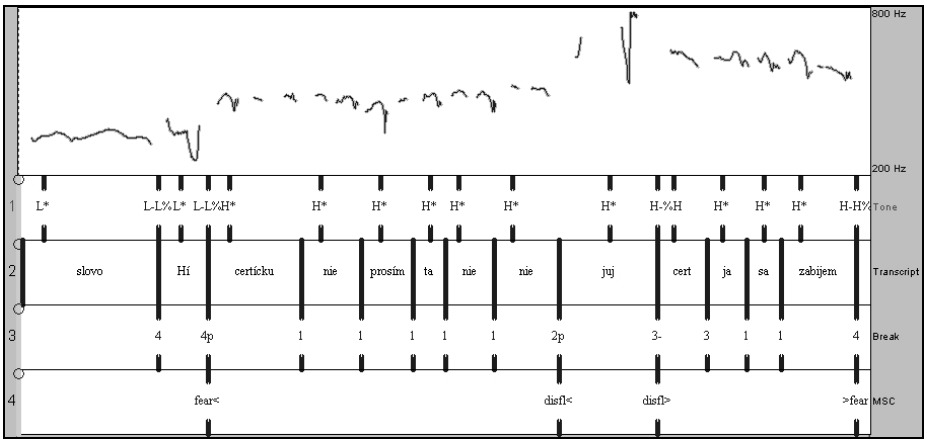


Fig. 3. Pitch range of silly billy utterances *„...slovo. Hí! Čertíčku nie, prosím ťa nie, nie! Juj! Čert, ja sa zabijem!*".

4 Conclusion

Our paper presents a phonological prosody annotation scheme for Slovak, Sk-ToBI. We have analyzed a speech material of two different speaking styles: read sentences and recorded plays of a puppeteer, representing artistic style

1 Example: anger< … cough< … cough> … anger> Cough can occur inside the utterance pronounced in anger.

with very expressive speech. We have allocated concrete events and design additional marks in miscellaneous tier which should abroad the current set of marks for Sk-ToBI. It is needless to say that the set of marks could be changed and completed after a further annotation of another speech styles. We believe that the definition of Sk-ToBI gives a serious basis for prosody research and opens a possibility to start a development of prosody-driven and expressive speech oriented applications in Slovak, mainly in the area of analysis, recognition [11] and synthesis of human speech.

References

1. Silverman, M. et al.: ToBI: A standard for labeling English prosody, Proceedings of the 2nd International Conference of Spoken Language Processing, Banff (1992), 867–870.
2. Baumann, S., Grice, M., Benzmüller, R.: GToBI – a phonological system for the transcription of German intonation, Proceedings Prosody 2000, Speech Recognition and Synthesis Workshop, Cracow (2000) 21–28.
3. Pierrehumbert, J. B.: The Phonology and Phonetics of English Intonation. PhD dissertation, MIT. IULC edition, (1980).
4. Dzur M., Sabo R., Rusko M.: Sk-ToBI Scheme for Phonological Prosody Annotation in Slovak, Proceedings of TSD 2007, Pilsen 2007, Czech republic. (Accepted)
5. Kráľ, Á.: Pravidlá slovenskej výslovnosti, SPN, Bratislava (1988).
6. Sabol, J., Kráľ, Á.: Fonetika a fonológia, SPN, Bratislava (1989).
7. Rusko, M., Darjaa, S., Trnka, M., Cerňak, M.: Slovak Speech Database for Experiments and Application Building in Unit-Selection Speech Synthesis. In: Proceedings of Text, Speech and Dialogue, TSD 2004, Brno, Czech Republic (2004) 457–464.
8. Zibert, J. Mihelic, F. et al.: The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation – Overview, Methodology, Systems, Results. In: Proceedingds of the 9th European Conference on Speech Communication and Technology Interspeech 2005, Lisbon (2005) 629-632.
9. Beckman, M.E., Gayle M. A.: Guidelines for ToBI labelling. Version 2.0, February 1994,
   `http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/`
10. Hirschberg, J., Beckman M.: The ToBI annotation conventions (1994) `http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html`
11. Cernak, M., Wellekens, Ch. J.: Emotional aspects of intrinsic speech variabilities in automatic speech recognition. In: SPECOM 2006, 11th International Conference Speech and Computer, Saint-Petersburg (2006), Russia, 405–408

The Possibilities and Limits of Lexicographical Description of the Czech Lexicon in Database Form[1]

Jindra Světlá

Institute of the Czech Language of the ASCR, v. v. i.
svetla@ujc.cas.cz

Abstract. Since 2005, the Department of Lexicography and Terminology (DLT) of the Institute of the Czech Language (ICL) of the ASCR, has focused on the creation of the material and technical preconditions for using modern information technologies when investigating and describing the vocabulary, and mainly on the design and development of our own lexicographic software, ranging from the initial specification of requirements, developing and testing a special lexicographic program called PRALED, to launching the lexical database designed by us at our own server. The programming part involves the development of relevant software, the empty shell of the database and the interface for data input, which should after 2010 also serve for the development of a new monolingual dictionary named LEXIKON 21. Our department has simultaneously focused on conceptional work linked with future detailed treatment of Czech vocabulary in the lexical database.

1 Introduction

The lexical database (hereinafter LDB) can implement various functions. It can be programmed so as to be used by language users when searching for requested language data in text corpora[2] or in the existing dictionaries transferred into electronic form[3], or it can be conceived from the very beginning so as to serve for the preparation of a new monolingual dictionary[4]. Another possibility is to combine these functions in such a way that the LDB of Czech being newly created would become not only a resource for a new lexicographic

1 This paper was created within the research plan of the ICL of the ASCR, v. v. i. *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (AV0Z90610521).

2 Cf. F. Čermák, Manuál lexikografie, p. 55 [1]: 'The lexical database of the relational type tends to be the highest and most abstract achieved level of corpus'.

3 Cf. Slovník spisovné češtiny [5], electronic version based on the 2nd ed., publ. LEDA; cf. also K. Pala 2001, p. 155 [4].

4 Cf. point C8 of the research plan above: 'This database will be designed in such a way so that its content could become a starting point for creation of a dictionary of current Czech language targeted at both common users and also at the needs of processing natural language in information technologies'.

description of the language, more detailed than in a traditional monolingual dictionary, but also in such a way that it would allow for searching in available electronic monolingual dictionaries or encyclopaedias. We selected the last alternative, which is the most advantageous for future users but most challenging for the programmers, authors of the conception and future compilers of the new electronic dictionary in the form of a lexical database: after its completion, the user will be able to confront a new dictionary version with previous dictionaries in an electronic form, in which it is possible to search using the integrated DEBDict engine; furthermore, morphological information will be generated upon request with the help of the morphological analyser AJKA, in which it is possible to display the entire paradigm (both of these tools have been developed at the FI MU in Brno).

2 Advantages and disadvantages of the database treatment

The processing of vocabulary by computers presents immense possibilities and removes the various limitations which the publishing of printed dictionaries had to take into account (the size, alphabetical ordering, clustering, etc.). Nevertheless, in order for lexicographic work to be able to utilise all possibilities of informatics, it was first necessary to analyse the feasible possibilities of computer support (electronisation of existing dictionaries and the extensive material resources of the ICL, the possibilities of making use of corpus texts, the available lexicographic programs and other supportive tools), to define the conception of the future LDB, its total size and the detail of the treatment of an entry, and on that basis to propose a suitable technical solution (specification of the database structure and software requirements) as well as a new methodology of lexicographic work. Specifically, it was necessary to conceive an apt manner for structuring lexical data for Czech (design of individual fields for different types of information) and their inserting in the LDB as well as new principles for lexicographic description of Czech vocabulary arising from it, because 'the development of a lexical database entails the electronic coding of information categories while simultaneously applying a standard lexicographical analysis'[5]. In connection with computer work, also an entirely new form of an entry emerges, the structure of which needed to be designed first, which resulted in the creation of the form for data entry.

Since computers started to be used, the forms and technology of lexicographic work have been changing, but it is at the same time desirable to keep building on traditional principles and methods of lexicographic description,

5 L. Kralčák 2001, p. 154 [3]

because technology can never completely replace a lexicographer's work in the phases of the analysis of exemplification material and the actual treatment of entries, it only makes it easier.

Among the main advantages of database processing are the new possibilities related to the work methodology, whereas a disadvantage for the compiler is that the relevant requirements have to be first incorporated in the program and subsequently developed conceptually in the Manual for Authors so that they could be really utilised. Among the new possibilities are for example:

- the possibility of a more detailed lexicographic description than in a traditional monolingual dictionary: the principle of maximum conciseness does not apply here, and therefore it is not necessary to save space, frequent words of the description metalanguage do not need to be truncated, the explanation of the meaning can be more explicit, the entries can exhibit a higher information density, some information can be hierarchised and presented in a different way at separate levels of description while employing hypertext links, lists, etc.;
- the possibility of division of work in our team: the emphasis put on the systemic approach to vocabulary and the separation of subsystems will ensure that component problems, e.g. morphology, synonymy and antonymy, phraseology, terminology, etc., be monitored in a uniform manner; continuing observation of the individual word classes and component lexical-semantic groups will make analogous treatment of definitions in the case of the same type of entries, uniform description metalanguage, gradual specification of the types of explanation and unification of the data from the individual parts of the entry easier;
- quicker searching in reference resources (i.e. in electronic dictionaries);
- a great advantage will be various possibilities of searching in the whole database: alphabetic, full-text, retrograde, searching based on the individual information fields, qualifiers, etc.; for that it is, however, necessary to propose and specify at the beginning of the work the requirements for the access into the database through various parameters (filters) and simultaneously to propose various output formats (displaying the sought data based on the entered parameters, the printing of selected configurations, the display and printing of the entire dictionary article in a form similar to a printed dictionary);
- the possibility of defining synonyms through not only their meaning but also their distribution;
- the possibility of providing a copious, internally segmented and classified exemplification;
- the possibility of illustrating better the syntactic and semantic collocabilities of the lexical unit;
- the possibility of using an elaborated note apparatus, which can also be hidden, etc.

3 On the conception of the lexical database LEXIKON 21 and the development of the PRALED program

So as to make it possible to render the Czech lexis in the form of a lexical database, it was necessary to design and develop such a database first, which means to define all the requirements concerning the software needed for its creation, to design the entire format of the entry, to specify the individual items which will be filled in for each entry, to define their content, to determine their order in the form, the manner of inserting them, to propose the possibilities of interlinking various entries, to set the requirements for the display and search of inputted data, the design of the output formats, etc.

The first version of the PRALED program, including the form for data entry and the empty shell of the new Czech lexical database, was created based on our requirements in cooperation with the workplace of the FI MU in Brno at the end of 2005, and in the 1st half of 2006 it was gradually being further specified and tested. When testing PRALED, however, it became apparent that the originally planned arrangement of the form with a fixed order of the individual sections might cause problems during the transition from one individual part to another within longer entries and that this version of the form's structure is not suitable for our purposes, because we need to be able to move in all directions in the entry form, we often need to go between the individual meanings, to see them concurrently and gradually to specify the respective explanatory definitions, continually to insert examples for various meanings, we need to have the possibility to come back to the individual sections of the entry form and to complement or revise the data at different levels of the hierarchy and in various stages of treatment.

On the basis of our experience acquired when testing the individual sections (tools) of the entry form and the subsequent treatment of a testing sample, we have been developing a new user interface called PRALED 2 since the middle of 2006, whose aim it was to eliminate the shortcomings of the first version when working with the tools in the entry form, and moreover it was necessary to design such a display of the data entered that would better suit our needs and correspond more to the traditional printed form of dictionary entries. Therefore, the entire design of the entry form and the principle for using the individual tools contained in it has been gradually changing.

This new user interface currently already makes it possible to enter data into the individual tools separately in any order and immediately to display them in the output format on the screen and allows the search for required data using filters (i.e. according to various parameters or attributes, e.g. according to individual word classes, according to date of inserting or changing, according to compiler etc., according to a combination of the data required). In

our work, we will exploit the various possibilities offered to us by the new technologies: hypertext links, displaying information of the same type in the uniform formatting and using various colours, fonts, etc., arrangement of some information into lists, printing of selected configurations, etc.

The program designed by us will from the very beginning serve – in accordance with the research plan – mainly for inserting and storing lexicographic data in the LDB and searching for required information of various types using the entered attributes. After 2010, we should smoothly switch to another project – the preparation of a new monolingual dictionary in an electronic form. In the final stage of the work on the new dictionary, thus after the complete filling of the lexical database LEXIKON 21 with the processed entries of an expected number of ca 100,000 – 120,000 lexical units, an interface will be added for the users. After the LDB has been made available to the public, it will be possible to provide the future users of the database with a complex display of every entry (as an dictionary article) as well as various groups of entries in an electronic form (e.g. via the internet or on a CD-ROM), and the user interface designed by us will additionally allow for the concurrent publication of an entire monolingual dictionary also in its traditional form, which today's users are accustomed to; this will then make it possible for various other outcomes to arise in printed and electronic forms (not only a monolingual explanatory dictionary but also component specialised dictionaries, e.g. a dictionary of synonyms, antonyms, homonyms, dialecticisms, most common terms, etc.).

Within the current research plan, we will test the principles of treatment for the needs of a future monolingual dictionary for the time being on selected entries; at the same time a detailed conception of a complex lexicographic description will be prepared on the one hand, which will meet both the requirements of the normal users of a monolingual dictionary and the requirements of professional users – linguists, and on the other a detailed conception of a subsequent computer processing of the lexicographic data for the purposes of natural language processing in information technologies. Already from the beginning, the emphasis will be on the arrangement of each entry in terms of meaning, on the explanations of meanings and their exemplification by a sufficient number of authentic examples; from the viewpoint of further language research, the entry form will also contain tools which will in cooperation with other scientific workplaces make the future treatment of various data suitable for professional users possible.

After the testing and trial operation have been completed, the empty shell of the database will gradually begin to be filled under the name LEXIKON 21 (hereinafter L 21). The actual treatment of the lexical units included in the index according to the new lexicographic conception will not be implemented until the next task, i.e. during the creation of a new monolingual dictionary.

4 Overall structure of the database

4.1 Entry window of the application – 'Heslář' (List of entries)

The following information automatically displays in individual columns of the Index for each entry: the lemma, word class, designation of its homonym, variants, explanation of the meaning, compiler, date of entry/changes. Moreover, the entries are colour-differentiated according to type (one-word entry – multi-word entry – sub-word unit – abbreviation/symbol). This window is intended for searching and displaying the list of entries according to the set requirement.

Selection of entries for the List. In view of the various criteria which could be used when selecting entries for a monolingual dictionary, it was decided for the first phase of work that an interim starting point would be a purely formal selection of lexical units on the basis of their frequency in the SYN2000 corpus. A list of the 50 thousand most frequent lemmas included in *Frekvenční slovník češtiny* (Frequency Dictionary of Czech) [2], hereinafter FSČ, which had been compiled on the basis of the above-mentioned corpus, was therefore for the time being designated as an initial, working index for the first stage of work. This List of entries will be revised in the next phase and gradually complemented through a deliberate selection so that it will be ready in 2010 for the processing of a new mid-sized to large monolingual dictionary.

The new lexicographic description will be based on one-word lexemes, nothing will be clustered, all the word classes will be treated anew. Also homonymous lexemes, variant forms of an entry, selected abbreviations, symbols and proper names will be treated as separate entries. In the future, also an independent treatment of selected multi-word lexemes (including terminological phrases and phrasemes) is planned.

4.2 Form for data entry

We have developed a working form for data inserting for each type of headword; according to type of entry, we distinguish between four different entry forms (for one-word entries, multi-word entries, sub-word units, for abbreviations/symbols – they differ in the menu in the tools and are colour-differentiated in the List of entries). So far, the form for data entry for one-word entries has been almost completed, which will be for entries contained in the FSČ automatically generated from the source index with the lemma already filled in and with the frequency in the SYN2000 corpus shown. The other types of forms are being designed.

The entry form is in accordance with the requirements of the lexicographic description subdivided hierarchically into several levels: 'Heslo' – 'Význam' – 'Podvýznam' (Entry – Meaning – Submeaning). The initial working desktop for treating a new entry is the card 'Heslo' (Entry), from which we will get to the card 'Význam' (Meaning). It is possible to switch between the cards (Entry – Meaning 1 – Meaning 2, etc.) using the headings (tabs) placed on the top toolbar; the cards for the individual meanings are numbered automatically.

The data entry into the LDB is done in the entry form using the individual tools. The data referring to the entire entry are entered using relevant tools in the card 'Entry', whereas the data referring only to the individual meanings have their tools in the card 'Meaning'. All tools are opened as windows by buttons placed on the bottom toolbar, but the sets of the tools for the card 'Entry' and the card 'Meaning' are not identical.

The program makes it possible to open more tools (windows) simultaneously and place them on the desktop as needed. Each tool remembers its last position and the format (size) of the window – all can be controlled and changed with the mouse. Inside each tool, there are several items (boxes) to be filled. Each item in the given tool has its name (title) and relevant field to be filled in, into which the individual data of the lexicographic description may be inserted in various ways:

1) Automatically. With some selected items, such as the name of compiler, date of creation, date of the last change, frequency in the SYN 2000 corpus, the entry of the required data is pre-defined in the program;

2) By selection from a set offer (menu). A part of the data assigned to all lexemes will be created as a result of classification and categorisation into the individual categories, it is therefore desirable to apply a uniform approach when determining and entering the relevant category in the entry form by the method of a selection from a set offer in the menu. It was necessary to provide relevant range of possibilities for each offer concurrently with the program. This theoretical elaboration of the LDB conception all the way to the level of specific offers in the individual tools has gradually been worked on and is continuously tested. At the end, a Manual for Authors will be prepared, which will summarise the overall conception, the list of all the offers to select from and the elementary work procedures for data entry in various entry forms according to the type of entry.

For the required outcome to be as 'user-friendly' as possible, it was necessary to decide what all of the formulations which are selected or inserted in the tools of the entry form would look like, because the offer must contain a complete enumeration of the possibilities which could be relevant for the given information. Through the selection from such an offer in the menu for each item (where it is possible), we want to ensure that a great part of the data which will become part of the LDB have a formalised form.

Also the terms of all of the items or boxes in the form must be well thought-out, because they are used not only by the compilers as instructions for filling, but they will also be displayed as a component of the description metalanguage, and in addition they will be utilised by future users as a filter (attribute) for searching. We were therefore trying to decide between Czech and international, shortened and full terms (i.e. titles of the individual items), etc.;

3) By typing in a text field. A part of the information will be entered by typing in a text field, i.e. in a metalanguage description in the entry form of a verbal commentary. Also in the case of these descriptive data, it is desirable that the used formulations be unified within the bounds of possibility. This applies both to the explanation or definition of the meaning according to individual word classes and lexical-semantic groups on the one hand and to the formulations in various notes, etc. as this would consequently make formalised searching for various data throughout the database easier;

4) By entering in a list. Lists that can be complemented have been prepared e.g. for the creation of a series of synonyms, for recording phrasemes and other multi-word lexical items, which will be treated separately and hypertext-linked in the future.

The procedure of entering data in an entry form. After the 'ZÁHLAVÍ' (HEADING OF THE ENTRY) has been filled in (i.e. after the lemma has been entered and the type of the entry and the word class have been determined, which is crucial because it generates the selection of the individual entry forms), it is possible to insert all data for the specific entry in question in any order. The data are entered (by choosing or typing) into respective items in a selected tool and saved in the LDB. Immediately after saving, all the entered data are displayed in the card 'Entry'.

The order of the tool buttons on the bottom toolbar as well as the order of the items within the tools are designed so as to correspond to the logical structuring of the dictionary article (and for the most part are consistent with the Czech lexicographical tradition). The filled-in data will display in the same order (with some exceptions, e.g. in the case of notes) also in 'NÁHLED' (PREVIEW). The compiler, however, can enter the individual data in any order – i.e. having completed the required fields in the Heading tool, he/she can continue in any card and with any item in whichever tool according to his/her actual needs, and the data saved will be automatically placed in the correct position and displayed in the pre-determined format in Preview. The author (compiler) of an entry only has to be careful about at which level of the hierarchy of the entry he/she currently is.

4.3 The display of an entry in Preview

As the completed items will be visible in each tool only upon opening the window concerned, we simultaneously had to design a further program for the separate display of all the data entered in Preview, which immediately after the data has been entered and saved into the database appears in the upper part of the card 'Entry' in the output format including different colours and styles of printing type for the individual tools (e.g. the explanation of the meaning appears in bold and cursive inside a red frame). Information of a similar type is grouped into colour-differentiated zones; when the entered data are displayed, the information is assigned to a pre-determined position within each zone (e.g. all grammatical characteristics are displayed in green).

This display form in Preview resembles the traditional appearance of an entry in monolingual dictionaries but at the same time will take into account the possibilities of electronic database processing (it will also contain e.g. hypertext links, various lists, etc.). In order to be able to print, it is necessary to transfer this hierarchical structure of the entry into a linear form (e.g. expand all lists before printing).

4.4 Access to material resources

In spite of the fact (or precisely thanks to it) that the modern technologies of electronic processing of extensive material files are constantly being perfected, research on the current lexis still remains a time-consuming matter, because along with the new possibilities of searching for linguistic units the extent and content of the studied material are also expanding.

At the time when the research plan was being submitted, only one fully tagged and lemmatised corpus – SYN2000 (100 million words) was available for research purposes. Already in 2006, another corpus of written texts –SYN2005 (100 million) was also made available, which will make it possible to search in further texts but it can at the same time double the time needed for the analysis of the searched concordances, because in some cases it is necessary to perform the same procedures twice on different material – so far it has not been technically possible to merge the two corpora. It is still an enormous benefit for us, because it has become apparent mainly in the case of words which can be found in SYN2000 with only low frequency that thanks to a different selection of sources in the new corpus further evidence is offered, which often helps elucidate and concretise the meaning concerned. The third corpus made available is SYN2006PUB (300 million), chiefly oriented on journalistic texts. Our LDB will remain separate from the corpora of the Institute of the Czech National Corpus, i.e. it will not be possible to access individual corpora directly from the database in order thus to prevent our server from being unnecessarily overloaded.

References

1. Čermák, F., Blatná, R. (eds.) Manuál lexikografie. H&H, Jinočany (1995)
2. Frekvenční slovník češtiny. Nakladatelství Lidové noviny, Praha (2004)
3. L. Kralčák, Projekt slovníka štúrovskej slovenčiny a jeho počítačová podpora. In: Slovenčina a čeština v počítačovom spracovaní, Veda, Bratislava (2001) 150–154
4. Pala, K.: Návrh české lexikální databáze. In: Slovenčina a čeština v počítačovom spracovaní, Veda, Bratislava (2001) 155–167
5. Slovník spisovné češtiny pro školu a veřejnost. Academia, Praha (1978), 2nd ed. (1994), 3rd ed. (2003)

Corpora

6. *Český národní korpus – SYN2000.* Ústav Českého národního korpusu FF UK, Praha 2000. Available online at: `http://ucnk.ff.cuni.cz`
7. *Český národní korpus – SYN2005.* Ústav Českého národního korpusu FF UK, Praha 2005. Available online at: `http://ucnk.ff.cuni.cz`
8. *Český národní korpus – SYN2006PUB.* Ústav Českého národního korpusu FF UK, Praha 2006. Available online at: `http://ucnk.ff.cuni.cz`

Automatic Term Recognition in Polish Texts

Dominika Urbanska and Dariusz Piechocinski

Polish-Japanese Institute of Information Technology,
Warsaw, Koszykowa 86, Poland
dominika.urbanska@pjwstk.edu.pl,dpiechocinski@gmail.com

Abstract. Automatic Term Recognition (ATR) is not a simple task, it is however an extremely important one. Methods for ATR can be divided into three parts: linguistic, statistical or hybrid. Due to the fact that the Polish language has rich inflections, the statistical approach could produce incomplete results. On the other hand linguistic analysis is resource demanding and time consuming. We decided to implement statistical and hybrid methods in order to compare the effectiveness and computing cost of each. For the hybrid approach we use our own grammatical filter, and for counting the frequency we use C-value method. While for the statistical approach we will use an algorithm based on the one proposed by Cohen (Cohen 1995). Our goal is to create an efficient implementation that could be used by other linguistic systems, for which Automatic Term Recognition will be crucial.

1 Introduction

In today's world we are observing an exponential growth in readily available information. The sheer amount of data available , for example through the Internet, makes it very difficult to search through and localize the information that one is interested in. Many multinational companies such as Amazon, Ebay and Google (to name a few) are taking computational linguistics and applying them to their vast databases with great success. Improved user customization and advanced searches are helping everyone filter all unwanted data and get the information that one wants.

"... in computational linguistic we have recently witnessed a growth in the interest in automatic treatment of terms, or linguistic units which characterized specialized domain, especially when NLP systems are passing from the development stage to the application stage." [4]

The possibilities are far greater than what is currently commercially offered and with the growth of interest in applications such as Name-Entity Recognition, Information Retrieval or Information Summarization, the need for computing not only common words, but Terms and Named Entities is crucial because these parts of each sentence in natural languages are the parts that carry the most relevant information.

The discovery of knowledge relies heavily on the identification of relevant concept, which are linguistically represented by domain specific terms. [5]

As you can see Automatic Term Recognition is rather a necessary process in linguistic computations. Of course term dictionaries exist, however they have their drawbacks and we choose not to use them in our research.

- Automatic term recognition, ATR in short, aims at extracting domain specific terms from a corpus which consists of papers or documents of certain academic or technical domain. [8]
- ATR may also be described as it is the extraction of technical terms from a special language corpora with the use of computers. [7]

While Terms represent the most important notions in the relevant domain and characterized documents semantically and thus should be used as a basis for sophisticated knowledge acquisition. [5]

Here are some Term definitions:

- Terms are the linguistic representation of the concept in a particular subject fields, and are "characterized by special reference" as opposed to words that "function in general reference over variety of codes". [7]
- Term is "the intersection between a conceptual realm (a defined semantic content) and linguistic realm." [7]
- Terms differ from general language words primary by their nature of reference. Nevertheless, it is not always an easy task to divide terms from word.[9]

ATR is very important task in Natural Language Processing (NLP). Many studies on automatic term recognition are concerned with interesting aspects of terms and their uses, but most of them are not well founded and described. [4]

There's a lack of Term research applications for the polish language altogether.

Automatic Term Recognition in particular, is needed because simple but coherently built terminology is the starting point of many linguistic applications and methods such as: human or machine translation, indexing, thesaurus construction, knowledge organization, etc. Because a successful term recognition method has to be based on proper insights into the nature of terms, studies of automatic term recognition not only contribute to the application of computational linguistics but also to theoretical foundations of terminology. [4]

In this article we will describe two methods for Automatic Term Recognition. First we will describe the statistical, followed by the linguistic approaches. Then, we will present details of our hybrid algorithm. Finally we test each algorithm using a single test data set. As a conclusion we will presents our findings and show possible extension for the application.

2 Methods for automatic term recognition

An ATR procedure consists of two steps. Firstly the term candidates are extracted. Secondly the score is assigned to each term candidate that describes

how likely the candidate is a term or not. These steps can be obtained by a statistical, linguistic or hybrid approach.

Statistical algorithms are based on frequency of occurrence or on probability.

Usually some statistical studies for recognizing complex units take a straightforward standpoint concerning the unithood of complex units. [4]

- For Salton, Yang & Yu indicates the method of extracting and weighting complex index terms by simply extracting two adjacent words(...), and give weights on the basic of their constituent elements. [4]
- Damerau (1993) based his weight computation on independent probability.
- There is also the Cohen (1995) algorithm, which will be described in details in the next section.

Statistical methods don't always return the correct terms, but at the same time they are much easier to implement than any linguistic filter. These linguistic filters can also be used as a methods for Automatic Term Recognition.

Some early linguistic methods which attempt indexing units consisting two or more words are reported in Earl(1970,1972) and Klingbiel (1973). [4]

- Earl's PHASE system uses a parser to obtain grammatical information about words, and to select noun phrases as index term candidates. Final index terms are selected from candidates according to the frequency of their constituent noun elements. [4]
- Kliebel also uses grammatical information. The 'recognition dictionary' attaches parts-of-speech to each word in the text, from which certain grammatical sequences are selected as index term candidate. [4]
- Similar work has continued to appear to date, with some refinements in various aspects, e.g. Portability, computational effectiveness, or range of coverage of index term forms. [4]

The linguistic approach is time consuming, and requires additional resources, such as corpora, or morphological analizators. On the other hand term candidates are more accurate, since they are preselected by grammatical filters.

So even knowing the following:

- Terms differ from general language words primary by their nature of reference. Nevertheless, it is not always an easy task to divide terms from word. [9]
- And even thought we know that terms are related to the restricted part of communications such as medicine, law, IT. [7]
- Most of the difficulties encountered in ATR come from the fact that distinguishing term from words in not an easy task. Though there exists term formation rules, these are not strong enough to distinguish terms from non-terms. [7]

It is pretty obvious that a mix, or hybrid, solution is possible. ATR can also be implemented using both statistical and linguistic methods. For example grammar filters can be used to extract the term candidates, while statistical methods extract real terms from term candidates.

3 Our solution

At first we wanted to implement statistical methods and linguistic ones to compare their results. However during the research we decided to pursue the implementation of statistical and hybrid methods.

The hybrid method uses linguistic analysis and statistical computing.

1. First we extract all n-grams (where n=1,2,3) from the input text. Simply method of extracting all n-grams causes that we receive also nested terms.
2. Terms are usually a noun phrases so we apply a grammatical filter in order to get only grammatically correct sequences. Grammatical rules proposed by us are:
   - Noun Noun
   - Adj Noun
   - Noun Adj
   - Adj Noun Adj
3. All n-grams that passed through the filter are our candidate terms. At this stage we use a statistical method called C-value (described in article "Automatic Term Recognition using Contextual Cues"). The method is used to measure probability that a term candidate is a real term. The C-value method compares numerical values for an examined term 'a', namely:
   - $|a|$ - number of words in 'a'
   - f(a) - frequency of 'a' in the corpus
   - b_i - other terms that contain 'a'
   - T_a - number of terms that contain 'a'

$$c(a) = \begin{cases} log_2|a| \cdot f(a) \\ log_2|a| \cdot (f(a) - \frac{1}{|T_a|} \sum_{b \epsilon T_a} f(b)) \end{cases} \quad (1)$$

4. All candidates with a value higher than the threshold are treated as real terms.

In the Polish language terms appear in texts in many different forms due to their morphological and derivational variations, so there is a necessity of normalization which is done, at the preprocessing stage at the begging of a hybrid algorithm. We use a morphological analyzer Morfeusz and Polish corpora [1].

Our second approach to term recognition is the statistical one, it is implemented in two versions:

1. one that uses a Cohen's algorithm [4]
2. one with C-value.

The C-value method is the same as we used in hybrid approach, so we can compare its efficiency with and without linguistic pre-processing.

The Cohen's algorithm implemented by us has been modified a little with relation to its original version and it proceeds as follows.

1. In the first step we extract n-grams just like in the hybrid approach.
2. Following computing is done separately for all n-grams level.
3. For each n-gram its weight is calculated. Weight is assigned to the ceiling of (n/2)-th wors in n-gram.
4. N-grams are regarded as terms if weight of its ceiling(n/2)-th word is greater than the threshold or its first and last words' weight both exceed the threshold.
5. Finally for extracted terms scores are calculated as average value of its weighted words. Those scores are used to create terms ranking.

The weight is defined as follows:

$$\Psi(i) = \begin{cases} c_i log(\frac{c_i}{s}) + b_i log(\frac{b_i}{r}) - (c_i + b_i) log[\frac{(c_i+b_i)}{(s+r)}] \text{ , when } \frac{c_i}{s} >= \frac{b_i}{r} \\ \\ otherwise\ 0 \end{cases} \quad (2)$$

c_i and b_i are the occurrence numbers of the i^{th} n-gram in the document and corpus; s and r are the occurrence numbers of all n-grams in the document and corpus respectively,

4 Tests and results

Estimation of term recognition was done using general domain corpus of the Polish language, in which texts are taken from newspapers, parliamentary proceedings, prose, etc. For evaluation purpose we used a number of articles from internet portal as a source documents to be terms recognized within. The division of articles was as follow:

- 35% - general
- 20% - business
- 20% - information technology
- 25% - medicine domain

As a measure of term recognition performance we chose recall and precision ratios.

- Recall, that is effectiveness of recognition, was measured as a ratio of correctly recognized terms to all terms in an article.
- Precision, that is efficiency of recognition, was measured as a ratio of correctly recognized terms to all extracted terms.

Results produced by the application were compared against human assessment.

Unfortunately during test period failures of the corpus server (which was accessed by the Web) were occurring. A reason of failure wasn't solved by the time of sending article deadline, so the results we present are only partial and are not representative. (But they are shown in the table nr. 2)

We would also like to discuss results and test procedure based on following sentences:

1. Zawarte w preparacie antocyjanozydy borówki czernicy wywierają stabilizujący wpływ na naczynia kapilarne, zwiększają napięcie żył.
2. Związki farmakologicznie czynne zawarte w głogu zwiększają przepływ krwi w naczyniach wieńcowych i w mięśniu sercowym.

There are 6 terms in above sentences , including nested terms (see table 1). [3] introduces two types of counting extracted terms. A perfect hit occurs when the boundaries assigned by ATR system coincide with those of term maximal form (ie. naczynia kapilarne). An imperfect hit occurs when the boundaries assigned by ATR system do not coincide with term maximal form (ie. na naczynia kapilarne). We counted term extracted by the application as correct only if its form was exactly the same like in table below (a perfect hit), because other forms are useless for a user.

term / recognized by	C-value w/ grammar	C-value w/o grammar	Cohen
antocyjanozydy borówki czernicy			
borówki czernicy	*	*	*
naczynia kapilarne	*	*	*
naczyniach wieńcowych	*	*	
mięśniu sercowym	*	*	
Związki farmakologicznie czynne			

Table 1.

method / ratio	C-value w/ grammar	C-value w/o grammar	Cohen
recall	0,85	0,22	0,16
precision	0,71	0,71	0,28

Table 2.

As a conclusion we may say that hybrid method produces the best results. Although precision of the C-value is the same with and without grammar, there is much difference in recall ratio. The reason of linguistic (hybrid) methods has obtained a higher score, can be easy explained. Grammatical filter eliminate construction, which are not corrected for terms. Such as i w (and in), zawarte w (consists in), while statistical method is based only on frequency. So if frequency of such construction appears in corpora a lot, such construction can be treated as terms. There is no doubt that results produced by the Cohen's algorithm are the worst. But because of the small number of test articles we can't say if this method is really worse than C-value.

The log-likehood ratio adopted by cohen is considered to be statisticaly valid for binominal distributions with low-level occuring events, and in thus in a

stronger theoretical position. Hoever in the case of Cohen this is adopted at n-grams characters, thus it cannot be said to be statistically valid with respect to the terms. [4]

Another valuable remarks is that result of methods, which are using the corpora, depends much on size and domain of given corpora. Terms which are not in the corpora at all, won't be recognized in texts. (for example antocyjanozydy borówki czernicy)

We also notice that c-value method does not work for single-terms. Since we calculate the product of two attributes, and we want to obtain number greater than 0, any of the attribute can be equal to 0. Unfortunately the attribute log(a), where a is equals to 1 is 0.

5 Summary

Although the work on application is not fully finished, the proposed solutions are sufficient. Both statistical and hybrid solution during test stage obtained reasonably good result. We see, that ATR task is an important one, for other NPL application. Therefore probably we would like to extend this application more, so it could be a starting point to some other systems.

References

1. M. Wolinski, http://nlp.ipipan.waw.pl/wolinski/morfeusz/morfeusz.html
2. IPI PAN Corpus, http://korpus.pl
3. Lauriston, A. (1995). "Criteria for Measuring Term Recognition", Seventh Conference of the European Chapter of the Association for Computational Linguistics, Belfield (Dublin), Ireland, 27-31 March 1995, 17–22 http://ucrel.lancs.ac.uk/acl/E/E95/E95-1003.pdf
4. Methods of Automatic Term Recognition. Kyo Kageura, Bin Ummino. http://citeseer.ist.psu.edu/cache/papers/cs/1991/http:zSzzSzwebserv.rd.nacsis.ac.jpzSz~kyozSzpaperszSzterm-recognition.pdf/kageura96methods.pdf
5. Morpho-syntactic Clues for Terminological Processing in Serbian. Goran Nenadić, Irena Spasić, Sophia Ananiadou. http://personalpages.manchester.ac.uk/staff/G.Nenadic/papers/mpsl03_nenadic.pdf
6. Evaluation of Automatic Term Recognition of Nuclear Receptors from MEDLINE. Sophia Ananiadou, Sylvie Albert, Dietrich Schuhmann. Genome Informatics 11: 450–451 (2000). http://www.jsbi.org/journal/GIW00/GIW00P106.pdf
7. Automatic Term Recognition using Contextual Cues. Katerina T. Frantzi, Sophia Ananiadou. http://www.ercim.org/publication/ws-proceedings/DELOS3/Frantzi.pdf
8. Automatic Term Recognition based on Statistics of Compound Nouns and their Components. Hiroshi Nakagawa. http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/academic-res/term9-2.pdf
9. A Methodology For Automatic Term Recognition. Sophia Ananladou. http://citeseer.ist.psu.edu/cache/papers/cs/27244/http:zSzzSzacl.ldc.upenn.eduzSzCzSzC94zSzC94-2167.pdf/a-methodology-for-automatic.pdf

Parallel French-Slovak Corpus

Dorota Vasilišinová and Radovan Garabík

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences,
Bratislava, Slovakia
{dorota,garabik}@korpus.juls.savba.sk

Abstract. Presented French-Slovak parallel corpus «FRASK» is a sizeable corpus consisting of European Union legislative texts and fiction in both French and Slovak languages. Texts are sentence-aligned, lemmatized and contain morphological information. The searching mechanism includes the possibility to query single words, phrases, lemmas and morphology tag, using regular expressions. The corpus is publicly available on the internet.

1 Introduction and choice of texts

The intended scope of the corpus is twofold: first, to create an aligned corpus of French and Slovak text for general purposes, and second, to support cross-language terminology research, especially with emphasis on legal and economic texts of the European Union legislature. The corpus therefore consists of two kinds of texts, the first part consisting of fiction and the second consisting of a collection of texts of European Union law. At the moment, the fiction part of the corpus contains three French novels and their translation into Slovak. Texts of European Union law include The Official Journal of the European Union, treaties, legislation, case law, preparatory acts and parliamentary questions. These texts were obtained from the JRC-ACQUIS Multilingual Parallel Corpus, Version 3.0 [jrc07], where the texts were already downloaded from the European Union information portal and conveniently converted into the XML format – but without any additional linguistic annotation, nor language-aligned.

The size of the corpus is 334 021 French and 226 990 Slovak words for the fiction part and 65 797 270 French and 59 076 782 Slovak words for the EU law part, totaling 66 131 291 French and 59 303 772 Slovak words (punctuation included).

2 Text format and processing

Texts in the corpus are processed in several phases using a modular system where each conversion step is applied to the previous level of conversion. There are several levels of conversion:

1. Conversion from the original file format (HTML, MS Word, etc.) into a simple text format (UTF-8 encoding, paragraphs separated by a blank line).

2. Manual editing of the document, where applicable (not in the case of the EU subcorpus). Stray texts at the beginning and end of the documents were compared and brought into agreement – there are often differences across the translations in the format of the document title, author, editorial prologues or epilogues.
3. Conversion into TEI XML format, with paragraphs marked by a corresponding XML tag.
4. Lemmatization and part-of-speech (or full morphological) tagging, converting the document into TEI XML format with sentence delimiters and grammar information for each word.
5. Conversion into simple text format suitable for the hunalign aligning program (using only lemmas, to help the aligning process), with a special sign '¶' as a paragraph separator.
6. Adding the alignment back to the TEI XML format as an attribute for the sentence XML tag, linking to the corresponding sentence(s) in the opposite language document.
7. Converting the data into a vertical file format, suitable for the Manatee corpus manager indexing.

Before lemmatization, the texts were typographically normalized – different quotation marks (Slovak „ “ ” and French “ ” « ») were all internally translated into simple straight quotes `" (U+0022 QUOTATION MARK)` and various kinds of dashes were translated into `U+002D HYPHEN-MINUS` for the benefit of TreeTagger, which works internally in the Windows-1252 codepage and cannot properly deal with rich typographical characters.

2.1 French lemmatization and POS tagging

French texts have been lemmatized and morphologically annotated with TreeTagger, a tool for annotating text with part-of-speech and lemma information. The part-of-speech tag system used is described in [Ste03]. POS tags for the French language include 33 tags which describe major word classes and some of their inflectional variants (e.g. verbs in conditional, future tense, imperative etc.), tags for special word forms (abbreviations, acronyms), miscellaneous symbols and certain punctuation marks.

The French letter (ligature) *e dans l'o (Œ, œ)* has been retained in the corpus. Although the majority of the texts used the simple *oe* character sequence (probably due to inadequate historical use of the ISO/IEC 8859-1 character encoding), we decided to keep the *œ* character, if present in the source texts. This means that both the variants (e.g. *coeur* and *cœur*) are considered to be two different words and special care has to be taken when querying the corpus (e.g. by using the appropriate regular expression `"c(oe|œ)ur"`). Lemmatization contains the orthographically correct *œ* form regardless of the original variant, so when querying the lemma attribute only the canonical form needs to be used: `[lemma="cœur"]` (compare with `[lemma="moelleux"]`).

2.2 Slovak lemmatization and POS tagging

Slovak texts contain complete morphological information. Each word is assigned a lemma and a morphological tag, containing all the relevant grammar information (such as gender, case, number, tense, aspect). The tagset used is described in [Gar06], and for homonymy disambiguation, we are using the Hunpos tagger [HKO07] trained on a manually annotated corpus of about 511 thousand tokens.

3 Alignment accuracy

Texts were aligned using the hunalign [VNH+05] software, which works on a sentence level, using a combination of length and dictionary based similarities to align the parallel texts. Although hunalign is able to work without a supplied dictionary, using one can improve the alignment dramatically. Since no French-Slovak dictionary was available, we bootstrapped a dictionary from automatically generated aligned word pairs, manually correcting the entries, obtaining an initial dictionary of 1 505 entries, and then running the alignment again, generating a new automatic dictionary and correcting it again manually. At the end, we obtained a dictionary of 6 858 manually verified word pairs. Alignment accuracy was estimated by choosing several (Slovak) words and randomly choosing several hundred concordances semi-uniformly dispersed throughout the corpus and manually counting the number of matching bisentences. We considered only 'perfect' matches, i.e. only those, where one source language sentence was translated by one target language sentence and correctly aligned[1]. In the following tables, we see the accuracy compared using the initial small dictionary, using the final dictionary and for the fiction corpus only, for the whole corpus, and for the whole corpus with filtered bisentences only (taking into account only those bisentences where alignment score as given by hunalign exceeds 0.5 and the lengths of original and translated sentence differ by less than 30 %).

word	dictionary	
	smaller	bigger
malý	58.5	63.2
počuť	77.8	84.4
voda	62.5	60.5
alebo	62.6	66.7
total	63.9	66.9

Table 1. Improving alignment accuracy by increasing dictionary size, the whole corpus.

[1] Obviously, using this method we can never reach 100 % accuracy, because often there is not a 1:1 correspondence between original and translated sentences, and even if correctly aligned, we do not count such translations as accurate.

word	dictionary	
	smaller	bigger
malý	76.7	91.5
počuť	69.2	83.7
voda	69.0	84.5
alebo	69.3	79.7
total	71.5	85.0

Table 2. Improving alignment accuracy by increasing dictionary size, fiction only.

word	corpus		
	fiction	whole	filtered
malý	91.5	63.2	94.5
počuť	83.7	84.4	87.1
voda	84.5	60.5	83.0
alebo	79.7	66.7	90.3
total	85.0	66.9	88.8

Table 3. Comparing alignment accuracy, bigger dictionary.

4 Query interface

Corpus backend is provided by the Manatee server [Ryc00], where each half (Slovak and French) of the corpus is indexed separately. Links between the halves are provided in form of a `link` attribute to the sentence XML tag (i.e. `<s link="5+6" id="4">...</s>` means that the 4^{th} sentence in one language corresponds to the 5^{th} and 6^{th} sentences in the other language). On top of the Manatee libraries, a custom WWW-based search interface has been built, using the Karrigell web application framework [kar07] in the Python programming language. The query interface follows the CQP syntax and provides full regular expression queries for words, lemmas and POS tags (or morphosyntactic attributes), displaying the result in a KWIC-like format, with parallel text from the other language displayed alongside.

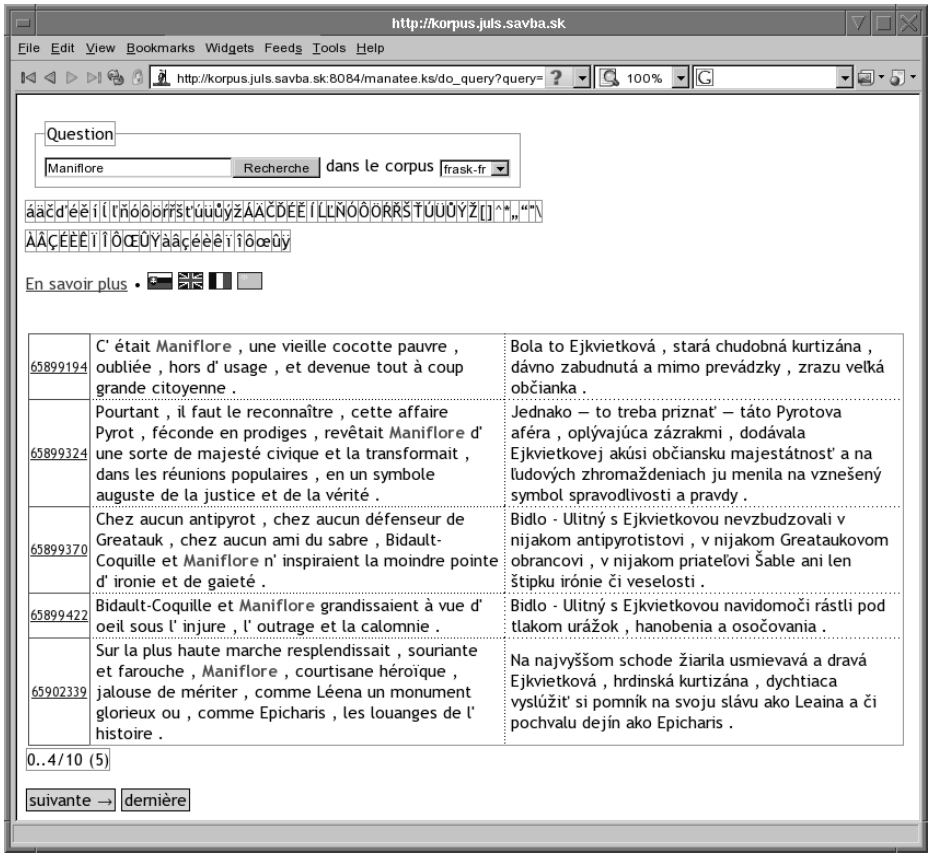


Fig. 1. Example of the query interface; searching for a proper noun.

5 Conclusion and further work

From the alignment accuracy comparisons we see that the alignment depends heavily on the size (and presumably quality) of the bilingual dictionary available. Our final dictionary of 6 858 words is obviously too small to cover much of the input texts, and does not contain many specialized words frequently present in legal texts. Our first necessary task will be to increase the size of the dictionary and to add the most frequent terms present in the European Union texts.

Since the provenience of the EU translations is not very clear, it is possible that we are dealing with two parallel translations into French and Slovak, not with the original and translation (in fact, the majority of the texts are probably just translations from original English). This does not diminish the usefulness of the corpus as such, but compels us to interpret the results with care and to apply additional measures to improve the corpus accuracy. In particular, we have to implement filtering, removing misaligned sentences and eventually also sentences

containing too much nontextual information – in the EU texts, there are often various lists, enumerations, tables and other elements, as well as complete texts in third unrelated languages in both the French and Slovak parts. Filtering out this content would improve the usefulness of the corpus texts and improve the aligning tools accuracy.

In the future we plan to provide the French part of the corpus with complete morphosyntactic annotation, using the FLEMM analyzer [Nam00]. In addition, an increase in the amount of texts in the corpus is a high priority, in order to augment the (rather small) fiction part to a more representative volume.

References

[Gar06] Radovan Garabík. Slovak morphology analyzer based on Levenshtein edit operations. In *Proceedings of the WIKT'06 conference*, pages 2–5, 2006.

[HKO07] Péter Halácsy, András Kornai, and Csaba Oravecz. HunPos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion*, pages 209–212. Association for Computational Linguistics, 2007.

[jrc07] `http://langtech.jrc.it/JRC-Acquis.html`, 2007.

[kar07] `http://karrigell.sf.net/`, 2007.

[Nam00] Fiammetta Namer. Flemm: Un analyseur Flexionnel de Français à base de règles. In Christian Jacquemin, editor, *Traitement automatique des Langues pour la recherche d'information*, pages 523–547, Paris, 2000. Hermes.

[Ryc00] Pavel Rychlý. *Korpusové manažery a jejich efektivní implementace*. PhD thesis, Faculty of Informatics, Masaryk University, Brno, 2000.

[Ste03] Achim Stein. French TreeTagger Part-of-Speech Tags, 2003. `http://www.ims.uni-stuttgart.de/ schmid/french-tagset.html`.

[VNH+05] Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*, pages 590–596, 2005.

Tools for the Input of Morphological Data – L 21 Solution Proposal[1]

Milada Voborská

Institute of the Czech Language of the ASCR
voborska@ujc.cas.cz

Abstract. The morphological tools developed within PRALED conform to our endeavour to describe the morphological characteristics of a word (taking into consideration also the vagueness between the word classes and their subcategories). By these means, it will be possible to capture and describe i.a. fluctuation in gender, the fuzziness of number in nouns, or to find a word with similar characteristics but classified under another wordclass.

1 Introduction

When describing the morphological characteristics of a headword in a computer database, we attempted to make use of the possibilities offered by the computer program and the advantage of the abundance of space for the input of preferably complete and synoptic information. The computer processing does not limit the compilers to listing only relevant forms of the word on the basis of which the user of a dictionary would complete the whole paradigm relying on his/her own knowledge by matching the word in question to the respective paradigm or look for the word forms in complete overviews of the paradigms of the word classes, as has usually been the case up to now. The abundance of space makes it possible not only to list a complete paradigm for every word but also to supply the necessary morphological data in the special blocks of the exemplification section. The exemplification of a word in a relevant form in a wider context allows the user to understand more easily the language situations in which the given form appears and provides an explicit demonstration of how marked the usage of a certain word form may become in some texts. The tools were designed in such a way as to make it possible to cover the fluctuation between word classes and their categories, overlaps and transitions between them and to provide space for such phenomena which cannot be unequivocally classified into individual grammatical categories. This attempt was motivated by both various possible linguistic interpretations and with regard to contemporary technological possibilities (i.e. when searching for particular phenomena).

1 This paper was created within the research plan of the ICL of the ASCR *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (AV0Z90610521).

Words which function as different word classes when used differently are processed as separate entries. The selection of the word class determines the tool menu for the description of the morphological categories which is activated by clicking on the respective button on the bottom toolbar of the entry being edited. The inputted data will then be displayed in Preview on the line below the heading of the entry (with the unfilled fields remaining hidden). It will also be possible to use the same tools for the individual meanings of polysemes. Another place for providing grammatical information should be a tool Note related to the paradigm, making it possible to comment on the marked usage of some word forms, variant endings, etc., to which an exemplification block showing their usage in specific contexts should correspond.

2 Specification of tools for individual word classes

2.1 Nouns

The declension of a head word can be described thanks to the selection of prepared items, which will provide the inflection of the described entry word. A situation when a word is both inflected and remains uninflected (*image*) may be marked separately or the tool duplicated if the word needs to be classified under two declension types (*slaneček*). The selected tools will make it possible to describe cases when the genders of a word differ in singular and plural (*kníže, dítě*). Possible differences in the usage of a noun in a certain number will be entered either in a grammar note or at the level of individual meanings if a new meaning is created by using the word in plural (*paměť* vs. *paměti*). Special items are prepared for the description of words which are in a certain meaning used mostly or exclusively in singular or plural. The fluctuation in gender with the meaning of the word and the form in the nominative singular being the same will be treated by adding a whole line with grammatical information and giving both (more) paradigms (*smeč*, *kápo*). Each paradigm will show all the standard alternative endings; should the usage of one of them be marked, a gloss will follow in a grammar note. Nouns of the type *choť, bačkora, naivka*, which have the same form of the nominative singular for both male and female genders but with a different subsequent declension will be treated as separate homonymous entries. The exemplification will then illustrate the notes given in the grammatical information and provide the contexts for all variants; possible markedness will be indicated in the relevant item next to the field for contexts.

'Subtype' is the working name for the item intended for providing information on the type of the noun (plurale tantum, singulare tantum, collective noun, abstract noun, mass noun, proper noun). The item 'Feature/Function' provides two characteristics, namely for nouns with a quantifying feature, for which the abbreviation 'kvant.' (quant.) is reserved (it will also be used for

other word classes) so that all words distinguished by this feature (*polovina, stovka, moře*) could be found upon entering the abbreviation. The abbreviation 'predik.'(predic.) will be used for the meaning of a polysemic lexeme in which a noun will be in the predicative position (*zima, hanba, radost, škoda*).

Deverbal nouns, nouns designating qualities (*-ost*), feminine nouns derived from masculine animate forms and diminutives (in the case of last two depending on the degree of lexicalisation and frequency in the corpus) will be given separate entries.

2.2 Adjectives

The first information will serve for the specification of the manner of inflecting adjectives. Should there be semantic differences in the usage of suffixed adjectival forms (longer) versus nominal (shorter) forms, it will be shown in the relevant grammar information on the card of the given meaning which of the forms refer to the meaning in question (eg. *hodný, hoden* – deserving something vs. only *hodný* – good, kind; a fair (way)). Another menu concerning adjectives is used for recording a more detailed information on the type of adjective, or on its formation (possessive adjectives, adjectives formed from transgressives and participles, eg. s*ynův, vařící, pečený*). Like in the case of nouns, also here the abbreviation 'kvant.' will be used to indicate the quantifying feature (*mnohý, pětimilionový*).

Comparison of adjectives will be shown at the end of the section for the grammatical information.

The inclusion of deverbal adjectives in the list of entries will be determined by the degree of their lexicalisation and the frequency of the lexeme. They will be explained or will have a link inserted, which will take the user to the verb in question.

3 Proposed solutions for other word classes, which will need to be further developed

3.1 Pronouns

We proceed from the fact that pronouns, along with other expressions of a similar nature, form a group of deictics. Apart from pronouns, this group is mainly formed by pronominal adverbs and pronominal numerals. All should be marked in the database with 'deikt.' (deict.) = a deictic word. The expressions bearing the semantic feature of quantification will, in addition to the abbreviation 'deikt.', also have the label 'kvant'. (quant.). Our further intent is to indicate the syntactic function of the pronoun in the sentence.

3.2 Numerals

It is necessary to solve both the classification of numerals and the method for their description to ensure that all essential features of the lexical unit being described are included. In *Mluvnice češtiny* [1], numerals are conceived very broadly, including also such designations considered by dictionaries to be nouns or adjectives, eg. *čtvrt, polovina, dvojice, stovka, nulový, tisícový, mnoha-miliónový*. We suggest that these groups of expressions be considered as nouns and adjectives while additionally labelling them with the abbreviation for the quantification feature at a relevant place. We would simultaneously like to indicate that some numerals have morphologically and syntactically a character of nouns (*milion*), adjectives (*pátý*) or adverbs (*dvakrát*).

3.3 Verbs

Our intent is to present a full paradigm of its forms for each verb, which will make it possible to include the information on the grammatical categories of verbal person, number, mood, tense and voice in this overview and describe possible particularities in the note and document them in the exemplification block. Items have been proposed for this word class which will contain a menu for the selection of adequate data rendering the headword being described. One of them will be intended for aspect, followed by a text field for the input of the aspect counterpart and also the syntactic and semantic characterisation of the verb, such as eg. valency, valency field, grammatical sentence pattern, semantic sentence pattern and others.

3.4 Adverbs

The traditional classification of semantic groups of adverbs into adverbs of place, time, manner, cause will be included in the explanation of meaning. Neither in the case of adverbs, however, will the classification be unequivocal, the individual usage of the words may differ (*neobyčejně krásný* vs. *neobyčejně oblečený*).

Tools have been prepared in the database which will make it possible to indicate stative and modal predicative adverbs, adverbial deictic expressions (*jak, kdy, tam, všude*, etc.), which will be indicated as 'deikt.' (in common for all deictic expressions), and 'kvant.' for adverbial expressions denoting quantum. In the case of pronominal adverbial expressions, it is possible to indicate if they are demonstrative, interrogative, relative, indefinite or negative.

Comparison of some adverbs will be shown, like in the case of adjectives, at the end of the grammatical information.

References

1. Mluvnice češtiny 2, 3. Academia, Praha (1986)

Comparing Natural Language Identification Methods based on Markov Processes*

Peter Vojtek and Mária Bieliková

Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology
{pvojtek,bielik}@fiit.stuba.sk

Abstract. We discover and experiment with categorization-based methods to natural language identification. Two approaches to language identification based on Markov processes are compared, both methods treat the incoming text on the character level. We performed series of experiments with the aim to make certain of high precision in language identification task of selected methods and also with the objective to compare them against themselves. Experimental evaluation was based on large-scaled Multilingual Reuters Corpus with various European and Slavic languages. Our research results showed that both methods are comparable in the task of natural language identification achieving recall as high as 99,75%.

1 Introduction

Natural language identification is the process of automated labeling textual documents by their language (e.g. this paper should be labeled as written in English). Although exact definition of the term *natural language* is not formed, the term covers languages used by humans for common communication (like Slovak or English), as a opposite of artificial languages (e.g. C++, Java).

Exploration of automated language identification is usually motivated by simplifying document preprocessing and organization of information, this is also the case of our research, which is involved in a project affiliating methods and tools for acquisition, organization and maintenance of information and knowledge in an environment of heterogeneous information resources[1].

As many language identification approaches exists (see survey by Cole et al. [1]), we point out our main demands with the aim to determine the proper identification method:

- *efficient* – capable to process large number of documents in real-time
- *language independent* – process text quantitatively in contrast to methods based on language specific features

* This work was partially supported by the State programme of research and development "Establishing of Information Society" under the contract No. 1025/04.

[1] Project NAZOU – http://nazou.fiit.stuba.sk/

– *document format independent* – identify language directly from text of document and not rely on meta-information bound with this document (which can be missing or incorrect)

Enlisted demands can be fulfilled by language identification method based on statistical modeling of text. A text modeling technique used by selected identification method should not make use of whole words or even sentences, rather putting stress on the lower level of granularity, hence chains of characters of text should be regarded. According to the mentioned requirements, two techniques satisfy our demands: Markov processes and the N–gram analysis. While we realized experiments with N–gram methods in our previous work [2], in this paper we explore, compare and improve two language identification methods based on Markov processes designed by Dunning [3] (Statistical identification of language) and Teahan [4] (Text classification and segmentation using minimum cross entropy). In the rest of the paper we will refer to these methods by their author's name.

The major contributions of this work are (1) theoretical and experimental comparison of two concurrent Markov processes based language identification methods using large-scaled Reuters Corpora, and (2) enhancement of the process of evaluating the best matching language in the identification phase by normalization by document length, which extends the scenarios of use of both methods.

The rest of the paper is structured as follows. Overview of related work is proposed in Section 2. Identification methods based on Markov processes are explained in Section 3. Proposal of additional castigation of categorization methods using normalization is in Section 4. After that, we report out experimental results aimed at comparison of the language identification methods in Section 5. Finally, Section 6 concludes the paper and points out some issues requiring further work.

2 Related work

One of the simplest approaches to language identification is based on common words and unique combinations of characters [5]. This approach works quite well for large documents, but fails when the incoming textual information in getting smaller (e.g. document containing only one sentence).

Another way of language identification is to use N–grams. On of the most cited method is designed by Cavnar and Trenkle [6], based on list of the most frequently observed N–grams (i.e. sequences of characters), variable N–gram length is used. Suzuki et al. [7] discovers a methods based on N–grams capable to identify language and character encoding together. We experimented with this method using Slavic languages and character encodings in [2].

Many other language identification methods are derived from universal categorization methods, e.g. Naive Bayes, Support Vector Machines [8] or k-Nearest Neigbour [9]. Survey by Aas and Eikvil contains overview of these categorization methods, tools and linguistic corpora [10]. The drawback of these methods is

that the text is usually represented as a bag-of-words and language specific preprocessing as stop-word removal or stemming is necessary, another disadvantage is that the feature space is usually large and must be reduced, although feature space reduction methods based on Information Gain, Principal Component Analysis [11] or Collaborative Filtering [12] are already well explored.

3 Language identification methods based on Markov processes

Both language identification methods use the well known supervised learning schema [13]. Statistical model is created for each language in the learning phase. Each language model is constructed from pre-selected training text. Then identification phase can be proceeded, documents to be identified are passed and language tags are assigned to them. The best-fitting language model for each document is determined by an evaluation function.

While the Markov processes theory serves as the basis for both language identification methods, we shortly describe this theory first. Further reading on probabilistic modeling of text can be found in [14].

3.1 Markov processes as text modeling tool

Stochastic process is called the first order Markov process if its state c_k in time k depends only on previous state c_{k-1} in time $k-1$ (Formula 1).

$$P(c_k|c_0, c_1, \ldots, c_{k-1}) = P(c_k|c_{k-1}). \tag{1}$$

In general, n-th order Markov process is described in Formula 2.

$$P(c_k|c_0, c_1, \ldots, c_{k-1}) = P(c_k|c_{k-n}, \ldots, c_{k-1}). \tag{2}$$

The character sequence $c_{k-n}, \ldots, c_{k-1}$ is named as Markov process $prefix$ (also the term *context* is used), c_k is usually named $suffix$.

3.2 Dunning's language identification method

Learning Phase – Creating Models of Language Categories A training text document (representative of particular language) is processed as a stream of characters. This stream is divided into Markov processes with length k characters (k is the order of Markov process). Each unique Markov process is stored together with information about its number of occurrences. After processing the whole document, all Markov processes counts are converted into probabilities using Formula 3 (k-th order Markov processes).

$$p(w_1 \ldots w_{k+1}) = \frac{T(w_1 \ldots w_{k+1}) + 1}{T(w_1 \ldots w_k) + |A|} \tag{3}$$

where $|A|$ is the size of an alphabet, $T(w_1 \dots w_k)$ is number of occurrences of Markov process prefix, $T(w_1 \dots w_{k+1})$ is number of occurrences of the whole Markov process and $p(w_1 \dots w_{k+1})$ is the computed probability.

As an example, processing the text "abracadabra" into Markov processes of order $k = 1$ is in Table 3.2, c is the number a particular Markov process occurred and p is the probability computed using Formula 3.

Order k = 1		
Predictions	c	p
a → b	2	$\frac{2}{5}$
→ c	1	$\frac{1}{5}$
→ d	1	$\frac{1}{5}$
b → r	2	$\frac{2}{2}$
c → a	1	$\frac{1}{1}$
d → a	1	$\frac{1}{1}$
r → a	2	$\frac{2}{2}$

Table 1. Processing the text "abracadabra" into 1st order Markov processes.

Identification Phase Language category models (i.e. persistently stored Markov processes bound with their occurrence probabilities) are created for each language in the learning phase. In the identification step, evaluation function is applied to the input text for each language model (Formula 4) and the best matching language model is determined.

$$log\ p = \sum_{w_1 \dots w_{k+1} \in S} T(w_1 \dots w_{k+1}) log\ p(w_{k+1}|w_1 \dots w_k) \tag{4}$$

where $T(w_1 \dots w_{k+1})$ are the number of occurrences of all Markov processes present in the text and $p(w_{k+1}|w_1 \dots w_k)$ is the probability stored in a particular model for each Markov process. While the model can handle only already observed Markov processes, yet unobserved processes on the input are skipped in this phase. Logarithm scaling is used due to avoiding problems of numeric underflow.

Overall probabilities computed for each language category by evaluation function are compared between themselves and the model with a result closest to zero is the best fitting.

3.3 Teahan's language identification method

Although the supervised learning schema and Markov processes are also used in this language identification methods, the process of creating the language models and evaluating the best fitting model differs. While this identification method is

more sophisticated and complex, detailed description of the method is beyond the scope of this paper. Deeper explanation and discussion can be found in [4] and [15].

Learning phase Dunning's identification method stores Markov processes only of particular length k, language models adopted in this method use various length of Markov processes together. Theoretically, this approach brings smoother modeling of a text.

At first, all Markov processes and their counts are extracted from training text. Dunning's language identification method extracts only Markov processes of exact order k, Teahan's approach takes into account also all lower orders Markov processes $k-1$, $k-2$, ..., 0 and -1 (Note that Markov process of order 0 is the distribution of separate characters in a text and Markov process of order -1 is the estimated distribution of all characters that did not appeared in the training text).

Next, Markov processes counts are converted into probabilities (Formula 5). While different orders of Markov processes are used, "escape" probability mechanism is involved, providing switching from higher orders of Markov processes to lower. Escape probabilities are important when Markov process of length k occurs on input and this process cannot be found in the highest order model table (length k). In this case, order of model is decreased to $k-1$, actual Markov process on input is also shortened and this overstepping between different process orders is count in with relevant escape probability.

$$e = \frac{t}{n+t} \text{ and } p(\phi) = \frac{c(\phi)}{n+t} \tag{5}$$

$c(\phi)$ is the number of times a particular prefix of Markov process was followed by the character ϕ, n is the number of all tokens that have followed and t is number of unique characters that have followed. e is the escape probability and $p(\phi)$ is probability for particular character.

Processing of the text "abracadabra" using method based on Teahan's method is displayed in Table 3.3, involving Markov processes of orders 1, 0 and -1.

Identification phase Models of selected languages are already created in the learning phase. Document written in yet unknown language is processed as a stream of characters and Markov processes of all lengths from k to 0 are extracted. For each model of language and set of all Markov processes present in input text, a cross entropy is computed (6). The language model which has the value $H(M)$ closer to zero is chosen as the best fitting and input document is labeled as written in this language.

$$H(M) = -\sum p_M(w_1, ..., w_m)\ log\ p_M(w_1, ..., w_m) \tag{6}$$

The probability for character model of length k is determined using Formula 7.

Order k = 1			Order k = 0			Order k = -1		
Predictions	c	p	Predictions	c	p	Predictions	c	p
a → b	2	$\frac{2}{7}$	→ a	5	$\frac{5}{16}$	→ A	1	$\frac{1}{\lvert A\rvert}$
→ c	1	$\frac{1}{7}$	→ b	2	$\frac{2}{16}$			
→ d	1	$\frac{1}{7}$	→ c	1	$\frac{1}{16}$			
→ Esc	3	$\frac{3}{7}$	→ d	1	$\frac{1}{16}$			
			→ r	2	$\frac{2}{16}$			
b → r	2	$\frac{2}{3}$	→ Esc	5	$\frac{5}{16}$			
→ Esc	1	$\frac{1}{3}$						
c → a	1	$\frac{1}{2}$						
→ Esc	1	$\frac{1}{2}$						
d → a	1	$\frac{1}{2}$						
→ Esc	1	$\frac{1}{2}$						
r → a	2	$\frac{2}{3}$						
→ Esc	1	$\frac{1}{3}$						

Table 2. Processing text "abracadabra" using Teahan's method.

$$p_M(w_1, w_2, ..., w_m) = \prod_{i=1}^{m} p'(w_i | w_{i-k} \dots w_{i-1}) \tag{7}$$

p' gives the probability returned by model of order k.

Although the escape mechanism (described in the learning phase of this method) helps deal better with already observed Markov processes (or their sub-processes), when yet unobserved Markov processes is present on input and its first character does not matches first character of any highest level Markov process in a language model, the escape mechanism cannot be applied and the actual Markov process must be skipped.

4 Normalization of the evaluation function by document size

In some cases, we are not aimed at identification of many languages, but only of one exact language – e.g. we have a set of documents written in many languages and we want to filter out only those written in Slovak (note that we even may not exactly know which languages are present in the document set, thus we cannot create models for all languages). In the current state, both language identification methods are not designed to deal with this problem, while they always assign a language label to the input document in the identification phase – when only Slovak language model will be created in the learning phase, all documents from the document set will be labeled as Slovak.

We can deal with this problem by involving normalization of evaluation function by document text length, which enables us to divide the document-space explicitly into two sub-spaces: a sub-space containing documents written in Slovak language and a sub-space where are non-Slovak document. Evaluation functions of both language identification methods are normalized using Formula 8.

$$F(language\ model,\ input\ text)_{norm} = \frac{F(language\ model, input\ text)}{\#_{chars}(input\ text)} \quad (8)$$

$\#_{chars}(input\ text)$ is the number of characters in input text. Note that different approaches to normalization are known (Overview of alternative approaches to normalization is in work of Singhal et al. [16]). In Formula 8, normalization by the number of characters in a text is used while character encoding independence is achieved.

5 Experimental evaluation

The main goal of our experiments is to determine, if more precision modeling of a text involved by the identification method proposed by Teahan (described in Section 3.3) brings better results in language identification when compared with results of the method proposed by Dunning (Section 3.2). Second experiment investigates, if the normalization of an evaluation function (in both language identification methods) allows to separate the state space between models of languages.

5.1 Language identification

Comparison of the language identification methods was performed using eight European languages from the Multilingual Reuters Corpus[2] – Danish, German, Spanish, French, Italian, Norwegian, Portuguese and Swedish. Many models were created for each language with the aim to compare the identification methods in various conditions. Different granularity of modeling of the text was achieved by using 1st, 2nd, 3rd and 4th Markov process orders were (larger orders of Markov processes were not used due to memory limitations), the amount of learning text varied from 25 kB to 200 kB.

After the learning phase was accomplished, 2 000 testing documents for each language were passed and the ability of the language identification methods to correctly label the testing document was measured. Average size of the testing documents in the corpus is 1,2 kB, Figure 1 displays the histogram of the testing documents.

Results of the language identification in Table 5.1 and 5.1 contains averaged values for all languages.

[2] Reuters Corpora – `http://trec.nist.gov/data/reuters/reuters.html`

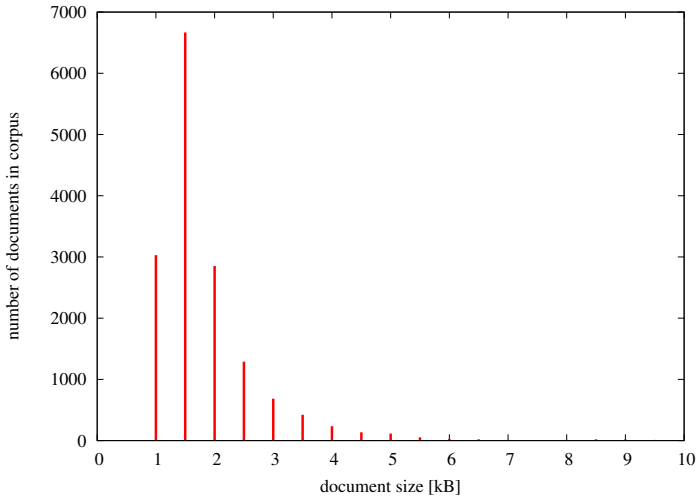


Fig. 1. Document size distribution in the Multilingual Reuters Corpus.

Markov process order	training text length [kB] / *Recall* [%]							
	25	*50*	*75*	*100*	*125*	*150*	*175*	*200*
1st	97.09	97.95	98.66	98.83	98.88	98.96	99.04	99.20
2nd	97.86	99.14	99.43	99.50	99.52	99.56	99.65	99.62
3rd	98.05	99.13	99.48	99.55	99.56	99.59	99.65	99.67
4th	97.67	8.83	99.08	99.28	99.36	99.61	99.63	99.70

Table 3. Language identification method proposed by Dunning, *Recall* values.

Markov process order	training text length [kB] / *Recall* [%]							
	25	*50*	*75*	*100*	*125*	*150*	*175*	*200*
1st	97.22	98.20	98.74	98.88	99.00	99.15	99.37	99.38
2nd	97.92	99.07	99.30	99.45	99.56	99.64	99.71	99.75
3rd	89.16	97.52	98.85	99.12	99.43	99.50	99.64	99.67
4th	64.79	62.60	62.94	62.01	70.37	77.47	81.03	84.20

Table 4. Language identification method proposed by Teahan, *Recall* values.

Comparison of the results in Tables 5.1 and 5.1 shows that both methods can deal very well when language models consisting of 1st, 2nd and 3rd Markov process order are used. Although the highest value of recall (99.75%) was achieved using the more sophisticated identification method proposed by Teahan, results shows that this method performs significantly worse when 4th of Markov processes are used. This degradation is present when the language model must treat yet unobserved Markov processes, which harms the Teahan's method more significantly. Unfortunately, adopting some mechanism to avoid this degradation will make already very complicated identification method even more complicated.

5.2 Normalization of evaluation function

The aim of this experiment is to determine, if the involvement of the normalization can clearly distinguish between languages, even when very similar languages are taken into account. If this hypothesis turns to be true, it will be possible to decide explicitly, where the boundary between languages lies, enabling us to avoid of incorrect labeling of documents written in not learned languages (as described in Section 4). The procedure is the same for both methods – only one language model is created (Slovak language) in the learning phase. Novels in Slovak, Czech and Polish language are evaluated in the identification phase.

Fig. 2 shows results for Dunning's identification method, Teahan's methods is evaluated in Fig. 3. Averaged values of the evaluation function are enhanced by standard mean value, where $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$. 95,4% of all documents of particular size should fall into the interval (assuming the Gaussian distribution).

The Y axis displays the normalized value of evaluation function applied in the identification phase. While only model of the Slovak language was involved, testing documents written in Slovak language naturally score best.

The results are similar for both methods – when documents smaller than 1 000 bytes are processed, documents written in Czech language are in many cases incorrectly labeled as Slovak. This is caused by the fact that Slovak and Czech languages are very similar, such a problem does not occurs when documents in Polish language are passed (e.g. value of $normalized\ evaluation\ function = 4.5$ safely divides Slovak and Polish documents in Fig. 3). The conclusion of this experiment is that when documents written in very similar languages are expected on input, the use of explicit division of state space should be carefully considered.

6 Conclusion and further work

We explored and compared two language identification methods based on Markov processes in this paper. Although method proposed by Teahan [4] is more complex than Dunning's identification method [3], our experiments based on Reuters Corpora and novels in Slavic languages showed that both methods treat the language identification task in similar way, achieving recall as high as 99,75 %. We

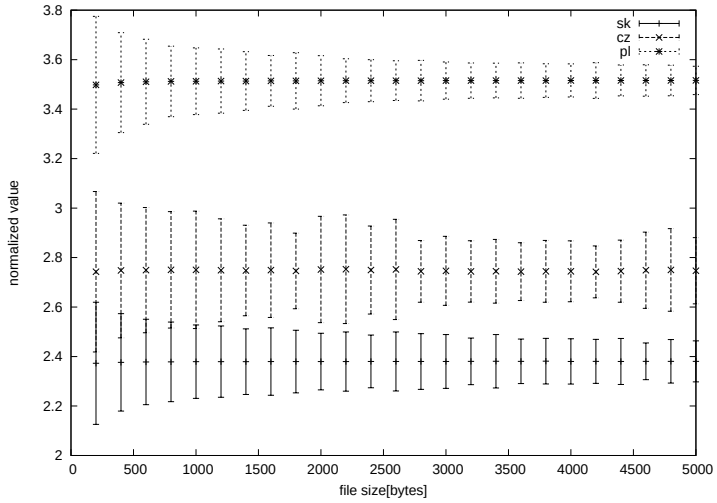


Fig. 2. Normalized evaluation function, Dunning's identification method.

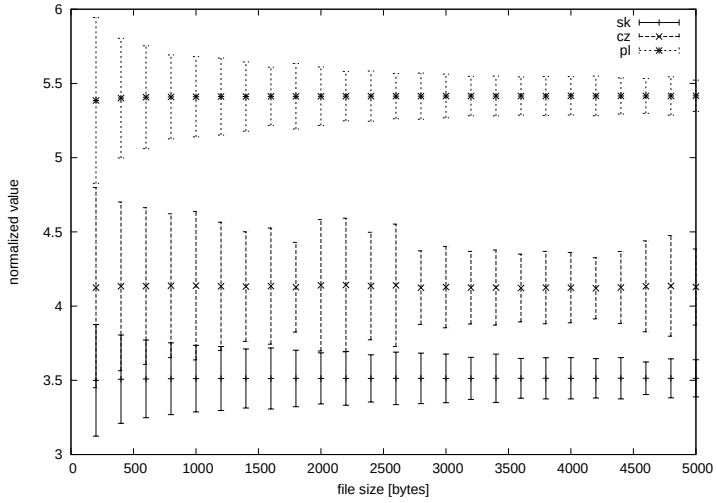


Fig. 3. Normalized evaluation function, Teahan's identification method.

improved the identification methods by involving normalization of evaluation function in the identification phase, enhancing the area of application of both methods.

Thanks to satisfactory results, Markov processes based identification methods served as the basis for a software tool incorporated into larger project affiliating tools for acquisition, organization and maintenance of information and knowledge in an environment of heterogeneous information resources [17]. This research project is experimentally evaluated in the domain of job-offers, our language identification tool serves in following ways – language identification of job-offers, document categorization [18] and semantic annotation [19].

Further work should focus on exploring the impact of character level text modeling (e.g. Markov processes, N–grams) in the task of general categorization, as a opposite to traditional bag-of-words representation. Already accomplished experiments include: subject classification [6], authorship categorization[4], genetic sequences classification [3] and we executed some preliminary research in categorization of job-offers [20] in Slovak language.

References

1. Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V.: Survey of the state of the art in human language technology (1995)
2. Vojtek, P.: Natural Language Identification in the World Wide Web. In Bieliková, M., ed.: IIT.SRC 2006: Student Research Conference, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava (2006) 153–159
3. Dunning, T.: Statistical identification of language. Techical Report MCCS-94-273, Computing Research Lab (CRL), New Mexico State University (1994)
4. Teahan, W.J.: Text classification and segmentation using minimum cross entropy. In: Proceeding of RIAO-00, 6th International Conference "Recherche d'Information Assistee par Ordinateur", Paris, FR (2000)
5. Grefenstette, G.: Comparing two language identification schemes. In: Proceedings of JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data. (1995)
6. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US (1994) 161–175
7. Suzuki, I., Mikami, Y., Ohsato, A., Chubachi, Y.: A language and character set determination method based on n-gram statistics. ACM Transactions on Asian Language Information Processing (TALIP) **1**(3) (2002) 269–278
8. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3) (1995) 273–297
9. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: 22nd Annual International SIGIR, Berkley (1999) 42–49
10. Aas, K., Eikvil, L.: Text categorisation: A survey (1999)
11. Zhang, R., Shepherd, M., Duffy, J., Watters, C.: Automatic web page categorization using principal component analysis. hicss **0** (2007) 73a
12. Song, Y., Zhou, D., Huang, J., Councill, I.G., Zha, H., Giles, C.L.: Boosting the feature space: Text classification for unstructured data on the web. In: ICDM '06:

Proceedings of the Sixth International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society (2006) 1064–1069
13. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer (2006)
14. Baldi, P., Frasconi, P., Smyth, P.: Modeling the Internet and the Web: Probabilistic Methods and Algorithms. Wiley (2003)
15. Teahan, W.J., Harper, D.J.: Combining ppm models using a text mining approach. In: DCC '01: Proceedings of the Data Compression Conference (DCC '01), Washington, DC, USA, IEEE Computer Society (2001) 153
16. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Research and Development in Information Retrieval. (1996) 21–29
17. Návrat, P., Bieliková, M., Rozinajová, V.: Acquiring, organising and presenting information and knowledge from the web. In: Proc. of Int. Conf. on Computer Systems and Technologies - CompSysTechŠ06, Varna, Bulgaria (2006)
18. Gatial, E., Balogh, Z., Laclavík, M., Ciglan, M., Hluchý, L.: Focused web crawling mechanism based on page relevance. In Vojtáš, P., ed.: Proc. of ITAT 05, Workshop on Theory and Practice of IT, Račková dolina, Slovakia (2005)
19. Laclavík, M., Šeleng, M., Gatial, E., Balogh, Z., Hluchý, L.: Ontology based text annotation ontea. In et.al., Y.K., ed.: Proc. of 16th European-Japanese Conf. on Information Modelling and Knowledge Bases, EJC 06, Paris, FR (2006) 280–284
20. Vojtek, P.: Improving Text Categorization Based on Markov Processes. In Bieliková, M., ed.: IIT.SRC 2007: Student Research Conference, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava (2007) 217–224

Spoken Corpus ORAL2006, Information It Provides and General Characteristics of Spoken Text

Martina Waclawičová

Institute of the Czech National Corpus
Charles University, Prague

The Czech National Corpus has several different parts such as various synchronic corpora, diachronic corpus and spoken corpora as well. Currently there are three corpora of spoken Czech in the Czech National Corpus available, namely Prague Spoken Corpus (PSC) built up in 1988–1996, Brno Spoken Corpus (BSC) from years 1994–1999 and ORAL2006 from years 2002–2006. In the following text I will concern on the last mentioned.

Spoken corpora offer (with respect to character of material they are built up from) information quite different from that in other corpora. ORAL2006 contains transcripts of recordings of informal speech transcribed in accordance with special transcription rules, whose aim is to take into account some special sound characteristics of speech. Regrettably the transcripts are not yet connected to the sound recordings.

Apart from the transcripts we observe two more fields of information about each recording, i.e. technical specifications and sociolinguistic data. *Technical specifications* include length of the recording, month and year of recording. *Sociolinguistic data* include place, region, type of situation and speakers' characteristics. For the specification of region we use Bělič's dialectal division of regions – Central Bohemia, Northeast Bohemia, Southwest Bohemia, Bohemian-Moravian (transient) region, Central Moravia, East Moravia and Silesia and the border areas (whereas ORAL2006 contains transcriptions of records only from Bohemia and bohemian borderland). As far as situation is concerned, we indicate its type (formal or informal – ORAL2006 covers only informal situations), topic, physical presence of the speakers (in ORAL2006 only speakers present in person) and preparedness of the speaker (in ORAL2006 only unprepared speakers). Further we note whether it is a dialog between two or more speakers who are present, the relationship between the speakers (whether they don't know each other, they only do know each other or they are friends) and the environment (private or public – in ORAL2006 only private). The most common situations of our recordings are a conversation at home, especially during a meal, visit, celebration or dinner in a restaurant.

Further on, anonymous information about the *speakers* is recorded. Analogous to PSC and BSC, it is their gender (male or female), age (two groups – younger, from 18 years of age to 35, and older than 35 years), education (elementary school, high school or university), place and region of birth, region of residence during childhood. Regions are specified according to Bělič as well.

Below I will discus three main groups of aspects that affect work with spoken corpora and character of information they allow to study. All the examples are chosen from ORAL2006.

The first aspect to mention is *construction of corpus* and its *representativeness*, i.e. which features were monitored and balanced – in the case of ORAL2006 it is sociolinguistic aiming. Texts for the ORAL2006 were chosen in certain amount according to the following features: gender and age of speakers and partly dialectal area, where speakers grew up. Restricted amount of all texts allowed us only partial balancing. The best-balanced feature is the age of speakers – there are speeches of 431 people aged 18–35 (57 %) and 323 people aged 36 or more (43 %).

Age	**Nr. of speakers**	**Nr. of tokens**
I	431	755 474
V	323	556 808

Table 1.

The second partly balanced parameter is the gender of speakers – ORAL2006 contains discourses of 302 men and 452 women, i.e. 40 % of all speakers are men and 60 % are women.

Gender	**Nr. of speakers**	**Nr. of tokens**
female	452	910 536
male	302	401 746

Table 2.

Next parameter is education. Speakers are grouped into two sets – those whose education is not higher than elementary or high school and those who at least started to study at a university.

Education	**Nr. of speakers**	**Nr. of tokens**
A	496	781 089
B	258	531 193

Table 3.

Feature, which is balanced least, is the region where speakers grew up: 452 (60 %) of them originate from Central Bohemia (this large proportion corresponds to the strongest nivelisation of speech in Central Bohemia and biggest medial influence coming from here), 139 (18 %) from Northeast Bohemia, 73 (10 %) from Southwest Bohemia and 87 (11,5 %) from bohemian borderland. A very small amount of speakers come from Bohemian-Moravian transient region: 3 (0,5 %).

Region	**Nr. of speakers**	**Nr. of tokens**
Central Bohemia	452	573 802
Northeast Bohemia	139	447 500
Southwest Bohemia	73	143 239
Bohemian-Moravian area	3	12 031
border areas	87	135 710

Table 4.

Thanks to this information we can among others investigate relations between linguistic features of spoken language and sociolinguistic categories. E.g. ORAL2006 shows with respect to *gender* of speakers that women use much more ensuring and contact words – *vid* ("am I right" or "doesn't it?") is used in 80 % of all cases by women. Investigating influences of the age of speakers we can in lexical area prove using of "trendy" words by young generation. Favorite word *hustej* "cool" is used in 80 % of cases by younger speakers and by older speakers only in 20 % cases. It concerns also using of some words serving usually to fill gaps in speech by younger generation. They used *normálně* "commonly" in 66 % of all cases or *prostě* "simply" in 77 %. Focusing on *dialectal* and *regional* language features we can examine them in the relation to the area where speakers grew up. E.g. we can assure that suffix of dative of masculine animate substantives *-oj* is typical for the Northeast part of Bohemia, because 90 % of its occurrence is in recordings from Northeast Bohemia, 10 % from Central Bohemia and only marginal occurrence is registered elsewhere. Dealing with age and region together we can see that some dialectal features, mainly the most regionally restricted or too noticeably different from literary or Bohemia-wide common Czech, are characteristic only for speech of older generations and vanish from younger people discourses. E.g. word *dolu* "downwards" has a dialectal variant with vocal length *dólu*. This variant is in 80 % used by older people and in 80 % by people grown up in Northeast and Central Bohemia.

The second aspect that affects work with spoken corpora and character of data concerns *sound* (partly *regional* or *dialectal*) *features*, which are available due to special *transcription*. The transcription used for ORAL2006 is not

phonetic, but is based on traditional script and differs from it in several ways. Words that are pronounced the same way in literary form of language as in common spoken form are obviously written according to the traditional script. Cases, where the literary spoken forms of language as well as the common spoken form differ from the script (*i – y*, *dě*, *tě*, *ně*, *bě*, *pě*, *mě*, *vě*, voiced and unvoiced sounds in certain positions), we use traditional script as well. But in other cases, where common speech regularly differs from literary pronunciation, the traditional script is not used and the differences are transcribed exactly as they occur in the common pronunciation. Equally, we try to record regional and dialectal distinctions from literary language as well. Consequently doublets appear in the transcripts. Thus we write *jsem* or *sem* ("I am"), *půjdu* or *pudu* ("I will go"), *dole* or *dóle* ("down"), *zrovna* or *zrouna* ("just"). Words at the beginning of a sentence are always written in lowercase. Capital letters are only used for proper names and some abbreviations. Interrupted or not finished sentences (both cases are very frequent in common speech), are marked in a special way, with the help of graphical marks (see below).

The segmentation of continuous speech to graphic units depends largely on transcribers' interpretation. Their decisions are based on intonation and meaning. It means that speech is not segmented according to pauses, but in this regard it comes closer to the character of written texts.

By the means of the transcription we can obtain such information, as percentage proportion of varying forms of words – e.g. nominative of variants of demonstrative pronoun masculine varying in the whole territory. (I leave out the most common variant *ten* that is used much more often than the other – 5611 tokens in ORAL2006.)

The third group of aspects influencing character of spoken corpora involves series of *sociolinguistic features*, that are recorded in addition to all the transcripts and which can be observed in corpus manager Bonito. Although they are not balanced, their helpfulness is unquestionable. We can make a subcorpus of texts with particular characteristic or show all the characteristics beside collocations or display frequency information including these characteristics etc. We can use these searching possibilities on examining lexical, morphological or phonetic features that embody some kind of variability in relation to recorded sociolinguistic characteristic.

One of them is specified *education*. We can display for every individual speaker, which of three types of school (elementary school, high-school, university) he or she studied. The next is *exact age* of speakers. Our basic grouping into two groups with the middle bound of 35 may be too raw and that's why there is also a possibility to display exact age and make individual grouping according to particular demands. Other data are *number of people* participating in the recording, *identification numbe*r of each speaker and *type of situation* – although it is always informal in ORAL2006.

In the following paragraphs I will focus on *general characteristics of spoken texts*. I will observe them in ORAL2006 through the use of tools provided by

the query engine Bonito. These features represent wide range of different language phenomena and I will choose only the most significant.

Spoken text are produced and perceived simultaneously during single process and in a particular situation, where speaker and listener as well as objects of communication are present. It causes such consequences such as unpreparedness, loss of sentential or textual perspective and strong situational boundedness. As the speaker produces his or her speech in real time, he or she lines up words gradually, adds words or phrases to once finished sentences, leaves one construction and continues in another etc. The speaker rectifies himself, repeats words and phrases or uses words that bring no information only to fill pauses in speech. It doesn't mean any obstruction for listener, but it makes perceiving and understanding easier.

Binding spoken text to situation causes much more frequent occurrence of *words with deictic function*. They refer outside the text to the reality, to common knowledge of speakers or inside the text to something that was or will be said. The most frequent word (4,1 % of all tokens) in ORAL2006 is pronoun *to* "it" and it has deictic and connecting function in general. Second most frequent deictic word is *tam* "there" – 1,1 % of all tokens. Deictic words often cumulate (*ale vypadal, víš jak? jako, jakoby takovej ten typ takových těch John Lennon prostě, jak sou takový ty, takhle dlouhý vlasy, takový ty řídký a brejle* "he seemed, you know how? like, as it were such that type of such those John Lennon just, how they are such those, that long hair, such that thin and glasses").

Pragmatic aspects are also manifested in special tools for expressing modality of a sentence. Presence of listener causes using *contact instruments* – not only lexical, but also lexico-grammatical or purely grammatical. The most frequent lexical instrument is *že jo* (resp. *žejo*, " doesn't it?"), 3596 tokens. The most frequent lexico-grammatical instrument is *vid'*, 1653 tokens (discussed above, typical for female speech). Grammatical tools are difficult to search because the corpus is not morphologically annotated. We can find some examples such as vocative *hele ty, ty Tondo, ty házíš dobře* („hey you, you Tonda, you can throw well") or transposition of function of verbal forms of 2nd person (*ale prostě když deš po tý tmě, tak prostě sem se bála* „but if you go through darkness, I was simply afraid").

One of the characteristics of spoken texts is that all speakers are aware of using *filling words*. As mentioned above, for the speaker they represent time to think up continuation or rectification, for the listener time to take a rest during reception and to interpret sense of what was heard. Speakers notice most often using such filling words as *prostě* ("just", 5922 tokens), *vlastně* ("actually", 1076 tokens), but there are much more frequent words, however not so noticeable (*jako* "like", "as", 18012 tokens).

Some words serve to building up text and dialog – these are among others words with preparative *and connecting functions*. They are signals for the listener that an utterance will follow, that the speaker wants to introduce a new theme and sometimes that the speaker wants to link to preceding text. Among ten most frequent words in ORAL2006 serving this purpose are *a* "and", *no* "indeed", *tak* "so", sometimes also *to* "it" or *já* "I" are used in this sense. They often cumulate, for example word *no* (28840 tokens, 11119 in the beginning of a sentence). Its most frequent right collocations are (ordered according to frequency): *a* ("and", 1920), *tak* ("so", 1694), *to* ("it", 1068), *jo* ("yeah", 975), *ale* ("but", 632) and *já* ("I", 467)), the most frequent left collocations are *no* ("indeed", 181), *jo* ("yeah", 83), *ale* ("but", 60) and *a* ("and", 41). One of characteristic text building features, *repeating* of words or whole phrases can serve several purposes – filling pause during formulating difficulties, making text more coherent or on the other side as a semantic tool of intensification or expressivity. There are 8736 repeated words in whole ORAL2006.

Another text-building characteristics are *rectifications*. During production of his or her discourse, the speaker often loses his perspective, makes mistakes in pronunciation, morphology, lexicon and syntax and consequently rectifies himself. Many of these rectifications are bound together with *unfinished sentences and words*. Unfinished sentences are completed with graphical sign ...: in their end (*voni právě ...: takže tam u nich se to dá vobjednat* „they just ...: so you can order it by them"), unfinished words with * sign (*pěk** nic*). There are 9012 unfinished sentences and 3378 unfinished words in ORAL2006. There are two possibilities how to continue the text after unfinished word – to start with completely another word or phrase (*voni spolu dost tah*, jako kamarádili* „they carr* were friends together"), i.e. lexical or syntactical rectification, or to pronounce the same word once more, but completed (*neu* neukazovala sem ti ji na fotce tam?* „di* didn't I show you her on that photo there?"), this can be pronunciation rectification or faltering.

The aim of this text was to show possibilities that spoken corpora, namely ORAL2006, offer to their users. Additional information is provided by the type of transcription that records some sound features, and sociolinguistic information that was used for balancing the corpus. Thanks to these data we can investigate language characteristic of spoken text either in general or in relation to various sociolinguistic parameters.

Acknowledgements

This research has been supported by the MSM0021620823 grant.

References

1. Bělič, J.: Nástin české dialektologie. Praha, 1972.
2. Čermák, F.: Mluvené korpusy. In: Korpusová lingvistika: Stav a modelové přístupy. Studie z korpusové lingvistiky, sv. 1. Eds. F. Čermák, R. Blatná. NLN a ÚČNK, Praha, 2006. P. 53–67.
3. Čermák, F.: Pražský mluvený korpus. 2001. http://ucnk.ff.cuni.cz
4. Čermák, F.: Today's Corpus Linguistics. Some Open Questions. In: International Journal of Corpus Linguistics. 2003. Vol. 7. № 2. P. 265–282.
5. Čermák, F. – Sgall, P.: Výzkum mluvené češtiny: jeho situace a problémy. In: SaS. 1997. № 58. P. 15–25.
6. Hladká, Z.: Tvorba a využití korpusů češtiny na FF MU v Brně. In: Hladká, Z. – Karlík, P. (eds.): Čeština – univerzália a specifika 4. Praha, 2002. P. 307–310.
7. Hladká, Z.: Brněnský mluvený korpus. 2001. WWW: http://ucnk.ff.cuni.cz
8. Kopřivová, M. – Waclawičová, M.: Construction of Spoken Corpus Based on the Material from the Language Area of Bohemia. In: Computer Treatment of Slavic and East European Languages, ed. R. Garabík. Veda, Bratislava, 2005. P. 137–140.
9. Kopřivová, M. – Waclawičová, M.: Representativeness of Spoken Corpora on the Example of the New Spoken Corpora of the Czech Language. In: Труды международной конференции "Корпусная лингвистика – 2006". Санкт-Петербург 2006. P. 174–181.
10. Kopřivová, M.: Struktura korpusu ORAL2006. 2007. http://ucnk.ff.cuni.cz
11. Rychlý, P.: Korpusové manažery a jejich efektivní implementace. FI MU, Brno 2000.
12. Waclawičová, M.: Mluvené korpusy v ČNK: několik poznámek k mluveným projevům a polyfunkčním výrazům. In: Korpusová lingvistika: Stav a modelové přístupy. Studie z korpusové lingvistiky, sv. 1. Eds. F. Čermák, R. Blatná. NLN a ÚČNK, Praha, 2006. P. 347–358.

Citation Card Files, Corpora of the Past

Victor Zakharov [1,2)]

[1] Department of Mathematical Linguistics
Philological Faculty, St. Petersburg State University
Universitetskaja emb., 11, 199034 St. Petersburg, Russia
[2] Institute for Linguistic Studies, The Russian Academy of Sciences
Tuchkov st., 9, 199053 St. Petersburg, Russia
vz1311@yandex.ru

Abstract. The paper explores the role of card files and corpora within the framework of modern lexicography. The Large Card File (LCF) of the Institute for Linguistic Studies in Saint-Petersburg comprises about 8 million cards. Its lexical stock surpasses all published dictionaries of Russian in its amount and diversity of artwork. The project to computerize the LCF was launched in 2006. The LCF database includes its word list and list of sources. The electronic word list allows for getting statistical data of card file structure as well as comparing concordances and word indices. The present work doesn't aim at creating the text database of citations. We propose that to cope with lexicographical tasks a system of semantic and other filters should be created which can help to search, choose and store data from a corpus for dictionaries. This selection will form a base of a modern card file.

1 Introduction

It is generally recognized that a dictionary should be based on a collection of lexicographic citation cards which give necessary data for its compiling and other lexicographic tasks. While compiling dictionaries of ethnic languages, one can deal only with the representative fundamental card file, which numbers in the millions of cards, and only thus can be considered reliable. Specially selected and processed texts are a source of such a card file, which reflect language in its variety.

Card collections of citations (card file, card index) are lexicographical corpora used for compiling dictionaries. They are prototypes for modern linguistic corpora. This topic was discussed in the "Language Corpora B.C." [1] by Nelson W. Francis, who developed the first corpus of English (Brown Corpus), where B.C. stands for 'before computers'. It deals with history and principles of creating card files for English dictionaries (XVIII – XX cent.): Johnson's Dictionary of English Language (1755), Murray's Oxford English Dictionary (1879-1928), Webster's New International Dictionary (1934-1961).

The Large Card File (LCF) of the Institute for Linguistic Studies of Russian Academy of Sciences, containing about 8 million of systematized cards with

citations, allows for various types of lexicographical and philological research [2]. Its stock was used by lexicographers while compiling a great number of dictionaries and grammars of Russian, including such basic and outstanding works as "Dictionary of Contemporary Russian" in 17 volumes, "Academy Russian Grammar" 1952–1954, "Russian Orthographic (Spelling) Dictionary" etc. [3]. Many researchers both from this country and from abroad use the Large Card File at their investigations on various topics. Nowadays the projects of "Big Academy Russian Dictionary" in 25 volumes and of "New Russian Phraseology Dictionary" are being done on its base.

The Large Card File was established in the 19th century under the guidance of the academicians J.K. Grot and A.A. Shakhmatov. At present the Large Card File consists of two parts. The one comprises about 5.5 million cards (collected from 1886 till 1968 г.), while the other contains more than 2.5 million cards (collected from 1968 till 1994 г.).

Two conferences dedicated to the Large Card File were held in 1986 and in 2001. All the speakers emphasized the importance of the Large Card File as Russia's cultural and national heritage.

However apart from its "memorial" function it also keeps on playing an outstanding scientific role. Suffice it to say that the collected lexical data outnumbers all published dictionaries of the Russian language. While even in big dictionaries one can find only a few citations, i.e. examples of word usage, the Large Card File might have several thousands. But its significance spans much further than in just the richness and variety of such a material.

Let's compare the volume of the LCF with that of main explanatory dictionaries of Russian [4-8].

Dictionary	**Volume**
Ozhegov-Shvedova [4]	72,500 entries
BAS-17 [5]	120,800 entries
MAS [6]	85,000 entries
BTS [7]	130,000 entries
BAS-25 [8]	150,000 entries

Table 1.

The cumulative dictionary composed by R. P. Rogozhnikova [9] on the base of 14 different dictionaries with word count of 170,000, while the word list of the LCF amounts to 400,000 entries. That means that the lexicon of Russian fixed by Russian dictionaries covers less than 50% of that of the LCF.

2 Automation of the large card file

Now it became critical to create a computer database of the Large Card File and to expand it on a regular basis with modern information technology. In the first place the database is meant to make the users' life easier giving them new resources.

The automation of the LCF boils down to computer aided maintenance, expansion and usage of the database. This includes

- compilation of the card index and its expansion;
- retrieval of the card you need;
- retrieval and sorting of cards for the purpose of creating new dictionary entries.

The expansion process, in its turn, comprises two operations:

- the further optimization and expansion of the LCF in terms of its representative content;
- compilation of brand new card files (or expansion of the LCF) as applied to the tasks of compiling new dictionaries.

The main stock of LCF are citation cards. There was worked a document which specifies principles of citation selection [10]. Citation cards contain also information that can be dubbed both bibliographical and technological. The first one refers you to the source where the citation has been taken from, while the second presents information, valuable for a lexicographer, such as the name the card researcher, disclosure of pronouns, explanation of word meanings, locales etc. (see Fig. 1).

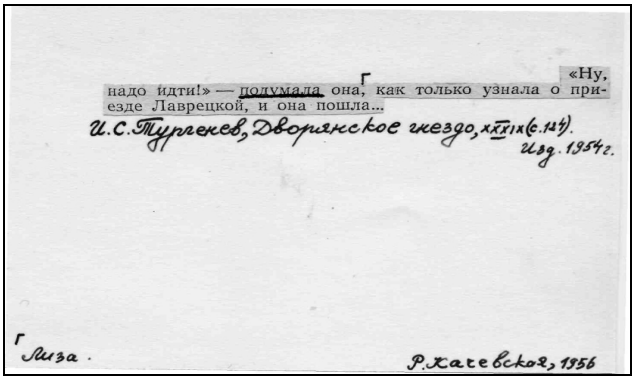


Fig. 1. Citation card.

The second dataset of the LCF, an array of entries (head words), could be called a lexicon of the card file. Inside of the lexicon there are links and references to other words which have to be indexed as well.

The third array is the list of sources which were used to make a so-called "sampling".

The ideal informational model of LCF could be illustrated with following chart:

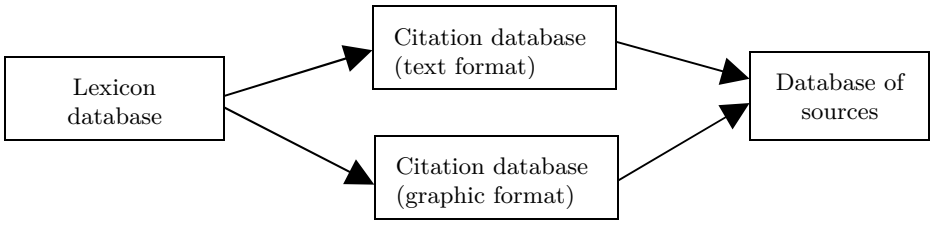


Fig. 2. The information model of the Large Card File database.

The first and foremost that both must and could be implemented in the LCF is compilation of the lexicon database (word list). At present there isn't even a complete paper lexicon in existence. The electronic version allows to obtain various statistical data: a number of different words, a number of cards in the LCF, a number of composite words, etc. Also the possibility exists of checking different word indices against it to determine which words are presented in the Large Card File and which are missing.

The word list (lexicon) database format has the following structure:

A word	**Number of cards**	**Reference**	**Comments**	**Accent**

For a composite hyphenated word (e.g. 'coach-grass', 'dust-coat') there are as many entries as there are components, e.g.:

багряно-красный (coach-grass)	6			
……	……	……	……	……
красный<= (grass<=)		*багряно-красный (coach-grass)*		

An ACCESS database is currently under construction, and by the end of 2006 about 70 thousand entries had been entered.

A word list should be cross with the citation card database. Should this database exist at all it would have allowed for using it as corpus. But to digitize hand-written citation cards by simple scanning is hardly doable as yet. As an alternative, a simple storage of cards' graphical images in the database can

be used. But that would be very labor consuming and require special equipment, too.

It would also allow working with it as with a corpus. But the task of digitization of citation cards can hardly be solved today while citation cards are mainly written by hand and in most cases their input isn't possible by simple scanning and recognition in order to create a text data set. An alternative way of dealing with the problem can be found in storing cards' graphical images in the database. But it also requires man-hour and special equipment.

The next database to be created is a bibliographical one of all the books and periodicals underlying the Large Card File. It should also be connected to the word list that would give information about, for example, the source of citation for a given word. Bibliometric analysis of the database is allow for dating the sources, identifying their genres and authors cited, as well as estimating the extent, their works were represented in the card file etc. This information will be of use while searching for new sources and citations.

3 Creation of an electronic card file

At the same time the expansion of the Large Card File proves to be one of the key problems. At its present form the card file is not enough representative enough. This can be accounted for by both by its inherent defects (as during the Soviet time a number of authors and works could not be included due to ideological reasons), and by lack of finance – as a consequence for the last 15 years very small amount of new entries have been added to it.

It is obvious that only cutting edge information technologies, i.e. electronic libraries, text corpora, programs for lexicographical tasks, can take care of current lexicography needs. Thus, further development and expansion of the LCF should be done electronically. So, the next step is a digital database of citations as a LCF supplement.

The LCF expansion and its digital "extension" implies the following steps:

- manual selection of citations from designated paper sources and adding them to the database;
- selection of citation from chosen digital sources;
- selection of citations from corpora;
- selection of citations from any digital sources.

This calls for finding the right manner of coordination between the paper and digital card files as well as for their concurrent use.

4 Card files vs. text corpora

The next issue to be discussed is the principals of "coexistence" for card file and text corpora.

The notion of 'corpus' is a next generation tool born within the tradition of card files that have been used by linguists for a long time. In fact, card file compilers starting from XVIIIth century have been discussing – and successfully resolving – contemporary corpora linguistics problems. Among the discussed issues were the sufficiency and representativeness problems: the Oxford English Dictionary comprised 4 mln cards, the Webster (Merriam-Webster) – 4.5 mln cards. Johnson, the compiler of the first English language dictionary in 1755 wrote: "I extracted from philosophers principles of science; from historians remarkable facts; from chymists complete processes; from divines striking exhortations; from poets beautiful descriptions" [1]. To follow natural and impartial selection principles of selection while working on the Webster Dictionary, W. Freeman Twaddell, for one, introduced a technology of independent (occasional) selection of citations and auxiliary words.

However card files neither reveal wider context, nor give statistics about word frequency or usage in a language or a sublanguage. A peculiarity of modern corpora consists in their annotation, namely, metadata. This metadata determines the opportunities that the corpus can give to a researcher.

The search potential of corpus managers outstrips that of the card files by a large margin, too. For example, there is a wide variety of search, namely, search for a single word, a lemma or a phrase, restricted search by a part of speech, by the text's type, search in parts of the corpus only, and much more [11]. The use of corpora can be helpful not only in studying lexical units in their contexts but also in getting data about their occurrences, frequency, grammatical categories, collocability etc.

Nevertheless national corpora and specifically the Russian National Corpus (`http://www.ruscorpora.ru`) [12] are generally only part-of-speech tagged and therefore can hardly be used for lexicographical tasks. Thus, for lexicographic research solving the problem metadata of a special kind must be assigned to the corpus, with special attention paid to the lexical and semantic variants of a word where the object of annotation is the meaning rather than the word per se. Some means of stylistic and subject-matter indexing of the text could be also of help for proper attributing lexical units.

At the same time, card files often contain comments or special lexicographic marking, normally missing in corpora, such as specification of homonyms, identification of the part of speech, of accent etc. (see Fig. 3 and 4).

A word	Number of cards	Reference	Comments	Accent
баба	180		женщина	
баба	1		кулич	
баба	16		орудие	
баба	2		растение	
закаменело	1		наречие	
закапывать	4		капать	
закапывать	16		копать	
забегать	38			забЕгать
забегать	49			забегАть
гусь	1		верхняя одежда	

Fig. 3. Examples of special lexicographic marks in the LCF.

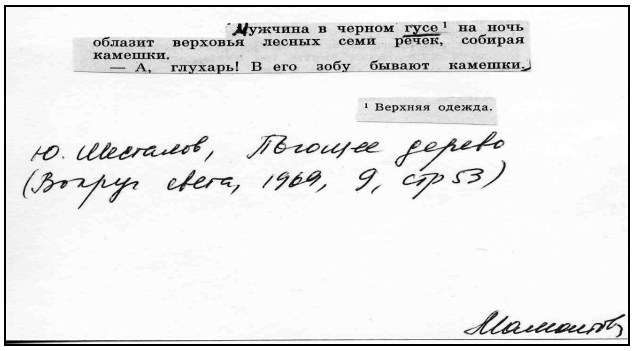


Fig. 4. An example of a citation card with specification of word meaning.

5 Conclusion and outlook

Hence there is a need for a special corpus as a supplement to the existing card files and to the Russian National Corpus that would meet the requirements of lexicography. The content marking should be performed automatically against a pre-set system of coordinates. At this point it is only a reputable fundamental dictionary that can provide for a ready system of linguistic coordinates, something like the "Dictionary of Contemporary Russian" [5] or a normative "Explanatory Dictionary of Russian" [4]. An automatic conversion of existing explanatory dictionaries into a structured XML-based form can be one of possible solutions. Such tagged dictionaries can be used for corpus annotation. Further corpus managers can deal with such corpus data, the latter being a source of contexts to illustrate the word usage in the Russian language.

"Coexistence" with corpora can be viewed as a next stage in the development of the Large Card File based on information technologies. A triad "Internet – corpus – card file" represents one of the possible solutions to the task [13, 14]. But in any case there should be a specialized intermediary system helping interaction between a corpus and a lexicographer and aimed at lexicographical tasks. Corpora often give thousand and thousands of contexts to a word, and it is absolutely impractical to watch them all on the monitor in terms of time and quality of research. "We need a system of distinct semantic filters, namely, metadata, that would help to find and arrange the data for an academic dictionary and would turn 'the huge slub' brought from the excavation called the Russian National Corpus into a material relevant both for 'artists and sculptors' working in the area of academic lexicography. The field that has proved to be an art of penetrating into semantic depths of a word" [15]. A system of word sketches being developed by English and Czech researchers that deals with lexical and grammatical collocability of words, can be seen as an example of that filter [16, 17, 18].

Finally (outside the task of the automation of the Large Card File) there is a need for an integrated system that will include as components different dictionaries and statistical instruments for data processing as its constituents not to mention card databases and an electronic citation card file. Such a system is aimed at supplying lexicographers with necessary and sufficient lexical array and tools that allow for unbiassed information about a word, its relations to other words, classified contexts etc. To make a long story short, a user needs a programmatically oriented linguistic system that permits operating both with corpus data and the data extracted from card files and dictionaries.

Acknowledgements

This research has been partly supported by grants of the Saint-Petersburg Scientific Centre of the Russian Academy of Sciences in the frame of the scientific program for years 2006-2007. I am grateful to Polina Ivanova and Yanina Krupina (Saint-Petersburg State University of Culture and Arts) and to students of the Mathematical Linguistics Department of the Saint-Petersburg State University for their manual work of entering data to a database. I'm grateful to Elena Gekkina and Vasilii Kruglov for inspiring discussions concerning issues of card file automation. My cordial thank goes also to Boris Neyman, Maria Khokhlova and Tamara Nevleva for their support in this work.

References

1. Nelson W. Francis. Language Corpora B. C. In: Svartvik J. (ed.): Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. Stockholm 4. – 6. August 1991. Mouton, Berlin (1991), 17–32.
2. Rogozhnikova R. P. Sokrovishchnitsa russkogo slova. Istoriia bol'shoi slovarnoi kartoteki Instituta lingvisticheskikh issledovanii RAN. Saint-Petersburg, Russia (2003).
3. Priemysheva M. N. Kartoteka kak istochnik leksikograficheskoy i nauchnoy raboty. In: Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskikh issledovanii. Saint-Petersburg, Russia (2003), 23–28.
4. Ozhegov S. I., Shchvedova N. Yu. Tolkovyi slovar' russkogo iazyka. Moscow, Russia (1992, 2997) (Ozhegov-Shvedova).
5. Slovar' sovremennogo russkogo literaturnogo iazyka: vol, 1-17. Moscow, Russia (1948–1965) (BAS-17).
6. Slovar' russkogo iazyka: vol. 1-4. Moscow, Russia (1957–1961) (MAS).
7. Bol'shoi tolkovyi slovar' russkogo iazyka. Saint-Petersburg, Russia (2000) (BTS).
8. Bol'shoi akademicheskii slovar' russkogo iazyka: vol. 1–6. Saint-Petersburg, Russia (2004–2007) (to be continued) (BAS-25).
9. Svodnyi slovar' sovremennoi russkoi leksiki: vol. 1–2. / Editor Rogozhnikova R. P. Moscow, Russia (1991).
10. Razrabotka leksiki i frazeologii sovremennogo russkogo literaturnogo iazyka: Posobie po vyborkam. Moscow, Russia (1972).
11. Zakharov V. Russian Corpus of the 19th Century. In: Text, Speech and Dialogue. Proceedings of the 6th International Conference TSD 2003, České Budějovice, Czech Republic, September 2003 / Václav Matoušek, Pavel Mautner (Eds.). – Springer-Verlag, Berlin, Heidelberg, 2003. – P. 146–151. (Lecture Notes in Artificial Intelligence, 2807) (2003).
12. Natsional'nyi korpus russkogo iazyka: 2003-2005. Rezul'taty i perspektivy. Moscow, Russia (2005).
13. Volkov S. Sv., Zakharov V. P. Informatsionnaia sreda sovremennoi leksikografii: korpus tekstov i/ili elektronnaia kartoteka? In: Sbornik trudov VII Vserossiyskoi ob"edinennoi konferentsii "Tekhnologii informatsionnogo obshchestva".
   (IST/IMS-2004). Saint-Petersburg, Russia (2004), 52–54.
14. Zakharov V. P. Web-prostranstvo kak iazykovoi korpus. In: Komp'iuternaia lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoi konferentsii *Dialog-2005* (Zvenigorod, 1-6 iiunia 2005). Moscow, Russia (2005), 166-171.
15. Gerd A. S. RNK i akademicheskaia leksikografiia. In: Trudy mezhdunarodnoi konferentsii *Korpusnaia lingvistika–2006*. Saint-Petersburg, Russia (2006), 88–91.
16. Kilgariff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine. In: Proceedings of EUROLEX-2004.
17. Rychly P., Smrz P. Manatee, Bonito and Word Sketches for Czech. Trudy mezhdunarodnoi konferentsii *Korpusnaia lingvistika–2004*: Sbornik dokladov. Saint-Petersburg, Russia (2004), 324–334.
18. Pala K. Word Sketches and Semantic Roles. In: Trudy mezhdunarodnoi konferentsii *Korpusnaia lingvistika–2006*. Saint-Petersburg, Russia (2006), 307–317.

Povaha a úzus interjekcí: případ češtiny

František Čermák

Ústav Českého národního korpusu
Filozofická fakulta Univerzity Karlovy Praha
frantisek.cermak@ff.cuni.cz

1 Úvod

1.1 Pohledy a aspekty

Jako všude tak i u interjekcí se lingvisté z různých důvodů co do jejich podstaty zcela neshodnou. Pracovním východiskem, zde v dalším užívaným a prověřovaným, tu bude poměrně rigorózní přístup daný jejich primárně jasnou funkcí, který se musel na rozdíl od jiných vyrovnat s kontextovou funkcí všech kandidátů na interjekce v Pražském mluveném korpusu (PMK) individuálně a manuálním způsobem. Hlavními problémy při tom byly pro své některé styčné plochy partikule a adverbia. Že jsou interjekce primárně v úzké vazbě na jazyk mluvený, každý intuitivně tuší, že však tu je jejich index opakování vyšší než u verb je překvapující. Interjekce jsou zaznamenány ve všech jazycích, jsou tedy univerzální, jako třída se však můžou v některých sekundárních rysech a jednotlivostech v nich lišit.

K základním rysům českých interjekcí v tomto pohledu patří (**1**) funkční, formální a syntaktická samostatnost, (**2**) sémioticky ikoničnost a semimotivovanost vnější realitou (jen u jádra jednoho typu), (**3**) fonologicko-morfologická specifičnost a jedinečnost (u interjekcí vlastních, tj. jádra), (**4**) plynulý přechod do víceslovných ustálených kombinací, které lze vnímat jako frazémy a idiomy. Takto rys (1) například implikuje, že za interjekce se pro svou plnou samostatnost nutně považují i tak častá slova jako *jo, ano, ne* apod.; zdůraznit přitom je však třeba i častou přítomnost pragmatické evaluativní funkce. Rys (2), na němž se dlouho a často demonstruje, že např. pes štěká v různých jazycích podobně, se dál snad připomínat nemusí. Rys (3) je ve fonologickém aspektu poměrně povědomý (cizí a neobvyklé fonémy či jejich kombinace, *haf, fuj*), málo se však zdůrazňuje a studuje přitom rys druhý, že pouze u interjekcí nemusí být hranice morfému jasná, především v důsledku emfatického prodlužování (*vrrr, ááále, chachacha, fííí,* kde skutečná délka závisí pouze na mluvčím a v textu je téměř vždy deformována, často přičinlivým panem redaktorem, a dostává matoucí určitou podobu a délku). Konečně v posledním rysu, kde se mj. projevuje dobře i nesmyslnost kodifikace, je dobře vidět, že např. mezi *kčertu* vs *k čertu*, anebo *ahoj* vs *dobrý den* žádný funkční rozdíl není, jde o zástupce dvou homogenních a sevřených funkčních tříd, a to kleteb a pozdravů, lišených pouze ne/diskrétností formy jejich zápisu.

Toto vymezení také ukazuje na nevhodnost staršího pohledu na citoslovce kladoucího do centra jejich vyjadřování citovost (interjekce voluntativní, imitativní ani kontaktové citové však nebývají) a tak i nevhodnost samotného českého ekvivalentu. I když lze zaujmout kritický postoj i k latinskému názvu (<*inter-iacere* a tedy „to co bylo vhozené, vhozek"), je nepochybně funkční víc (vhozením se etymologicky míní vložení interjekce do proudu mluvy, textu). Z dominantního hlediska funkčního vycházejícího z toho, že interjekce (s výjimkami) jsou nejen samostatnými větami, ale především promluvami a že z hlediska teorie slovních druhů se tu neutralizuje hranice mezi lexémem a větou, resp. promluvou, pak vyplývá, že se této interjekční funkce časem a vývojem můžou ujímat i slova další, která sem historicky nepatří, srov. (a) *Prosím!* (interjekce, např. v odpovědi na dík či při podávání) vs (b) *Von to* ***prosím*** *zapomněl doma!* (partikule) vs (c) ***Prosím*** *o rychlou odpověď* (původní verbum). Zvláště druhý případ překrývání a obdobných funkčních vlastností interjekcí a partikulí je jak významný, tak prakticky nestudovaný. Menší a okrajové rysy, které do jádra konstitutivních rysů jevu nepatří, tu uváděny nejsou, jakkoliv můžou být zajímavé (srov. např. extenzi flektivní morfologie do jinak morfologicky neměnných interjekcí v dichotomii sg-pl u *na* vs *nate!* či *ahoj* vs *ahojte!* aj.) anebo řadu aspektů fonologických a fonetických, jako je vyjadřování rytmu, jejich intonace apod.

1.2 Pojetí a klasifikace

Pracovním rámcem zde bude klasifikace interjekcí jednoslovných i víceslovných do pěti funkčních tříd (podtřídy jsou předběžné), užitých a testovaných na PMK:

1. **Faktuální** (objektivní relace k FAKTu), např. s podtypy *možnost, nutnost, ne/určitost* (snad, určitě), *ne/jistota*, *ne/pravdivost* (bůhví, čertví, opravdu, rozhodně, vážně), *ne/pravděpodobnost* (asi, nejspíš, patrně, pravděpodobně, sotva, že by?, cože?, prosím, určitě), *vyplývání, vyvození* (zřejmě, jó tak/ten!), *změna platnosti, všeobecnost* (takhle!), nebo s relací *suspendované* (tj. odmlka, výplňková slova, eh, hm, no) aj.
2. **Voluntativní** (relace z VŮLE (k tobě/faktu)), např. s podtypy *rozkaz, zákaz, žádost, vybídnutí, povel* (alou, hajdy, hop, pozor, račte, psst, syp, ticho, vpřed, kuš, marš, pohov), *zájem, touha, přání* (do toho!, skol), *uspokojení, uznání* (bohudíky, chválabohu, konečně, výborně, no proto!, tak vida!), *politování, omluva* (bohužel, pardon), *povzbuzení, útěcha, nabídka* (no ták!, do toho), *varování, upozornění, vyhrůžka, napomenutí* (no no!, ty ty ty!), *rada*: (když už tak už), *ne/spokojenost, rozhořčení, podiv* (jak to?, no tohle, tak ty tak! krucifix), *ne/souhlas, ne/přijetí*, *odmítnutí, vzdor* (ále, ano, jo, no, ne, houby, hovno, jasně, kdepak, ovšem, pochopitelně), *registrace, dovození* (aha, hm, ehm),

popř. s relací *suspendovanou* (nuda, lhostejnost, zdrženlivost, tak ať, ať si, eh, nic) aj.

3. **Emocionální** (subjektivní v relaci k FAKTU), např. s podtypy *ne/chuť, odpor, ne/libost* (á, aah), *obdiv, opovržení, posměch, ironie* (pff, to zrovna), *bolest, radost* (au), *strach, zděšení* (brr), *starost, úleva* (bohudík, konečně, zaplaťpánbu), *překvapení, údiv, úžas, zklamání, povzdech* (no ne!, no teda!, no tohle!), *lítost, soucit, účast* (ajaj), *dojetí, nadšení* (no teda), *pobavení, smích, smutek* (chacha), *zlost, vztek* (hergot) aj.
4. **Kontaktové (fázové)** (objektivní relace k TOBĚ), např. s podtypy *pozdrav* (ahoj, čao, hej, nazdar, papá, sbohem, servus, těbůh, zdařbůh, zdravím, zdravíčko), *přivolání, udržení* (tak co?, jářku, lidi, pane, panečku, člověče), *předávání* (na-te, tumáš/te) aj.
5. **Onomatopoické** (**imitativní**, relace Agenta k člověku/věci/zvířeti n. kontakt s ním), např. podtypy *lidská ústa* (ccc, mňam mňam), *člověk jinak* (fíí), *zvíře* (mňau mňau), *volání na zvíře* (čiči), *pohyb věci* (cink), *proces-exploze* (prásk, bum), *jiné* aj.

V tomo příspěvku se pokusíme se podívat, do jaké míry je hlavní rys interjekcí, jejich samostatnost (1), o kterém se mnoho pochybností nevyslovuje, nesporný a skutečně realizovaný, a to na stomiliónovém reprezentativním korpusu SYN2005, popř. PMK. Tento přístup bude součástí širšího pohledu na zjištění a ověření syntagmatických aspektů interjekcí obecně, o kterém se nikdy nemluví.

Dodejme, že počet výskytů interjekcí v korpusu SYN2005 není v důsledku nepřesné lemmatizace známý, v PMK se jich vyskytlo 8134 a v korpusu Karla Čapka jich najdeme 4743.

Pro představu si ještě uveďme 50 nejčastějších interjekcí ve Frekvenčním slovníku češtiny (Čermák, Křen, ed., 2004):

1- *ne, ále, ano, no, ahoj, jo/jó, ach, ó, hm*, **10**- *proboha, hele/heleď, aha, sakra, sbohem, hej, běda, haló, vida, fuj,* **20**- *panebože, nazdar, jé, bravo,* hurá, hop, cha/cha, *och, hop, bum,* **30-** *hergot, ehm, kruci, ježíšmarjá, ejhle, haha, panečku, eh, basta, ksakru,* **40-** *šup, bodejť, prásk, zaplaťpánbůh, probůh, bác, au, haf, che/che, pst,* **50**- *uf .*

2.1 Syntagmatika vnitřní

Uvažujeme-li o syntagmatice interjekcí, jakkoliv je to neobvyklé, pokoušíme se uvažovat eo ipso primárně o jejich kombinacích a interjekce izolované zůstávají prozatím stranou. Není to však tak úplně na místě, protože je třeba napřed a především lišit (I) kombinatoriku interjekcí s jinými slovnědruhovými lexémy

v rámci věty, tj. pokud bude zjištěna, od (II) kombinatoriky interjekcí mezi sebou navzájem. Říkejme pro jednoduchost prvnímu typu syntagmatika vnější (viz dál 2.2) a druhému syntagmatika vnitřní.

2.2 Ustálenost a neustálenost

Podívejme se napřed na druhý případ (II), na syntagmatiku vnitřní. Z poměrně hojných dat vidíme, že nesporně existuje. Nejprve je tu však třeba mezi kombinacemi interjekcí navzájem odlišit ty, které jsou (1) **ustálené** a tedy součástí systému, a ty, které jsou (2) **textové** a mají jen ad hoc povahu. Potíže s určením těch prvních, ustálených, nejsou ani tak dány nezvykem o víceslovných interjekcích vůbec uvažovat, popř. neodstraněnou nejasností v jejich formě (psát zvlášť či dohromady?, srov. *fuj tajbl* vs *fujtajbl*, resp. */tajfl*, anebo *abraka dabra* vs *abrakadabra*), jako potížemi s rozpoznáním ustálenosti některých kombinací, které se zvykově chápou jako neustálené, srov. *ach jo, ach ouvej*. Staré víceslovné interjekce typu výpůčkového *Á propos* (z franc.) nezná ani Mluvnice češtiny, i když ta možnost existence interjekcí víceslovných opatrně připouští, ale jen ve třech příkladech. Situaci v jejich vnímání komplikuje i chybně zapsaná forma a definice některých z nich, např. SSJČ zná pouze heslo *buch* s definicí pravící, že „označuje temný zvuk při úderu, pádu, výstřelu, výbuchu ap.“ a uvádí až pod ním jako příklad *buch buch* v jasném smyslu zabouchání na dveře. Je pak zřejmé, že uvedená definice tu neplatí a autoři se dostávají do rozporu s vlastním tvrzením. Přesto tvůrci slovníku nepřipouštějí, že tu jde o něco jiného, o jiný, a to víceslovný lexém, jehož úzus je silně vázaný na bouchání na dveře ve smyslu dožadování se vstupu, což vůbec neplatí o jednoslovném *buch*. Podobně problematicky je zachycena v SSJČ pouze izolovaná interjekce *ha* se svou definicí (překvapení, podiv apod.) a až jen dál, v exemplifikaci, se uvádí *hahaha* s poznámkou, že jde o smích. Tady je přirozená důležitost věcí, podporovaná i intuitivně frekvencí, postavená na hlavu: vyjádření smíchu je nesporně mnohem běžnější než vyjádření překvapení. Interjekci *bum bác* tento slovník vůbec nezná, uvádí jen zvlášť *bum* a zvlášť *bác*, bez sebemenší zmínky o této nesporně ustálené víceslovné interjekci.

2.3 Multiplikace a kombinace

Napříč oběma typy, kombinacemi ustálenými a neustálenými, jde častý případ reduplikace, lépe **multiplikace** některých interjekcí, obvykle z důvodu důrazu, popř. větší zřetelnosti nebo i k naznačení rytmu apod. Je-li na jedné straně tudíž reduplikaci *ach ach* nebo *hm hm* možné považovat spíše za neustálenou a náhodnou, zřejmá většina reduplikací se v korpusu už kvůli vysoké frekvenci, resp. opakování zdá být spíše ustálená, srov. *ha ha, haf haf, hip hip, hou hou, cha cha, mňam mňam/ňam ňam, pa pa, pi pi, puť puť, šup šup, ťuk ťuk* aj.

Vraťme se ale ke skutečným víceslovným interjekcím, jejichž horní hranice zřejmě není uzavřená a většina jejich kombinací patří do frazeologie (musíme

mezi ně počítat např. i pozdravy a kletby). Srov. pár příkladů na takové frazémy s interjekční povahou jen z části písmena **A** ze *Slovníku české frazeologie a idiomatiky. Výrazy větné* (SČFI4, v tisku): *A bác ho!, A hele!, A hrome!, A jéje!, A jó!, A kruci!, A sakra!, Á propos!, Aby ne!, Aby do toho už!, Aby ne!, Ach ano!, Achich ouvej!, Ale ale!, Ale ano!, Ale co!, Ale kdepak!, Ale no tak!…* Jednou z prvních věcí, která upoutá, je to, že původní interjekce tu jsou v menšině a komponenty těchto frazémů se rekrutují odleckud. Tyto příklady, ve zmíněném slovníku poměrně bohaté, lze doplnit i dalšími, excerpovanými z korpusu (SYN2005): *čáry máry, cupity dup, fuj tajbl, hej hou!!, hej hola, hej rup, hergot sakra, hopsa hejsa, houpity hou, ratata bum* aj. (z nichž většina do SČFI4 je zařazena také).

Interjekce se ale navzájem kombinují i bohatěji, nejen binárně. V SYN2005 nacházíme takto kombinace, resp. kumulace a multiplikace interjekcí v rozmezí od 2-7, srov. absolutní počty výskytů: binární 1094, ternární 242, kvatrální 64, kvintální 13, sextální 6, septimální 2. Aniž tu lze zacházet do detailů, už na první pohled je zřejmé několik věcí. Extrémní kombinatorické případy (zvl. šest a sedm) sestávají prakticky výlučně z opakování, iterace téže formy, o heterogenní kombinace tu nejde, srov. 7x po sobě jdoucí, opakované *aj*, které je téměř stejné u šesticové interjekční grupy s jedinou výjimkou *hej hou hej hou hola hou*, zatímco kvintální iterace je vedle zmíněných tří obohacena ještě o multiplikované *uf* atd; určitá monotónnost je tu tedy zřejmá. Na druhou stranu se zdá, že některé kombinace kvantitativně preferují větší počet komponentů, opět při iteraci, než dva, srov. kombinaci ternární *jé jé jé* (12 výskytů) oproti binární *jé jé* (1 výskyt). Otázka optimálního počtu členů při multiplikaci s ohledem na určitý typ interjekcí stejně tak jako otázka, které interjekce připouštějí multiplikaci a mají tuto kombinatorickou schopnost relativně snadno a která ji blokují a proč, je třeba teprve studovat. Souvislost mj. s typem textu je tu přitom nasnadě.

2.4 Tendence a korelace

Odlišíme-li kombinace náhodné (což není bez problémů), může nás však zajímat i pozitivní pohled, otázka možností a tendencí vzájemných kombinací. Je zřejmé, že tu lze rozpoznat aspoň dvojí kolokační tendenci, (1) **lexikálně-sémantickou** a (2) **gramatickou**. Kombinace, resp. kolokace prvního typu charakterizuje souvýskyt interjekcí s obdobnou sémantikou (pak jde obv. o důraz), srov. *ach bože/panebože* a povzdech, *eh co* a bezstarostnost, *hej hola* a zavolání a kontakt, *aha* a výraz pochopení, dovtípení (formulace SSJČ), popř. poněkud méně průhledné *ne, proboha!* a odmítání apod. Srov. ***Hej hola**, lidičky, kam se to valíte?,* ***Ach panebože**, byl to ten nejhorší tanečník, s jakým jsem v životě tancovala.,* ***Aha,** tak proto sis na dnešek napsal službu!*

Komplexní a smíšenou povahu mají kombinace s korelací gramatickou (2), z nichž nejvýznamnější a v textu nápadný je souvýskyt interjekce s pádem nebo modem, a to (A) a vokativu (jména), nebo (B) imperativu (verba), srov.

(A) *Ach Jirko, ach kamaráde, haló pane kolego, hej člověče, hej ty!, hej páni zedníci!*

(B) *ach slyšte!, hej počkej na moment!, pozor, nenarazte do mé hlavy!, pst, spi už!, prokrista, dejte mi pokoj!*

Základní funkcí těchto afinit a souvýskytů je vyjádření kontaktu (popř. oslovení apod., v A) a v druhém případě pragmaticky, resp. evaluativně modifikovaný rozkaz, výzva apod. Při bližším pohledu je tu kompatibilita a sémantická a funkční blízkost těchto dvou gramatických kategorií a některých interjekcí zřejmá, jakkoliv nestudovaná a stávající gramatické systémy ji neumějí uchopit. V kontrastu k imperativu (typ B) stojí někdy i jiné interjekce, pokud se takto kombinují, a ty pak preferují naopak indikativ (viz dál).

2.2 Syntagmatika vnější

Pozice a typy distribuce. Podívejme se na ni (typ I) z obou hledisek, formálního i sémantického. Z hlediska své formální **distribuce** ve větě interjekce vykazují většinou jasné tendence, někdy kombinace i výlučné distribuční preference. Obojí pochopitelně souvisí významně s jejich sémantikou a funkcí. Rozdělme si, bez další specifikace, formální postavení interjekce s ohledem na větu na čtyři případy, a to na pozici (1) na začátku věty či promluvy, resp. před ní, (2) uvnitř ní, (3) na jejím konci a (4) kdekoliv z (1-3) bez jasné tendence. Poslední (5) případ jsou interjekce skutečně izolované, bez jakéhokoliv vztahu k větě a kontextu. Ukazuje se, že tohoto posledního případu, jediného, který se až dosud zmiňoval, je výrazná menšina, a to obv. tehdy, když je promluva ukončena, je formulovaná vágně apod. Naopak se zdá, že většina případů interjekčního úzu spadá do prvních čtyř případů, které bývají o to zřetelnější, o co je kontext explicitnější a jednoznačnější. Případy (1) a (3) bývají často odděleny čárkou, někdy dokonce v uvozovkách, psaný úzus je tu ale spíš rozpačitý a bez zásad, jakkoliv o těsnosti spojení či případné pauze jasně vypovídá akustická stránka. Srov. příklady:

Iniciálová pozice (1), většina velmi pevná a výlučná, interjekce vyskytující se zde v jiné pozici nestávají: *Á, Aha, Ach, Au, Bode(j)ť, Haha, Haló!, Hej!, Inu, Vida… Proč by to nešlo? ptám se ho.* ***Inu,*** *povídá děda, protože…, štve je proti vládnoucí počasí… říkaje* ***Bodejť*** *by nebyla chřipka!*

Intrapropoziční pozice (2) bez bližšího určení, vždy však jako vsuvka: *Kdo vám,* ***jářku,*** *obloudil ducha a cit? Test,* ***hm****, komplexní kádrový test. Možná vám,* ***ehm****, zavolám zítra.* Už starší interjekce *jářku* je v této pozici nejčastější, stává však také v pozici iniciálové, i když se ale zdá, že nestává v pozici terminální. Zatímco interjekce *hm* preferuje opět vnitřek nebo začátek věty, není

jasné, do jaké míry se užívá na konci. Podobná interjekce *ehm*, která se typicky užívá zde, se naproti tomu zřejmě spíše na konci věty neužívá.

Terminálová pozice (3): *amen, a basta, haha…*: poslední příklad se tu uvádí spíš ilustrativně, sama interjekce *haha* patří do nespecifického typu (4): *hergot, kruci, sakra…*

Naproti tomu se zdá, že víceslovné interjekce preferují většinou pozici první. Tuto otázku a další je však třeba teprve v detailu zkoumat.

2.3 Valence a kolokabilita

K aspektům formální syntagmatiky lze počítat i další nezkoumanou otázku, a to **valenci** interjekcí. Z uvedených příkladů ***Běda vám****/jinověrcům* (D)*,* ***Kuk na*** *strejdu !* (A) *Opatrně se rozhlédl a* ***šup do*** *komory* (G), *Právě teď kolem vezou citróny a tak* ***tradá za*** *autem.* (I) ***Prásk do*** *koní!, Prásk bičem.* (G, I) ***Vida ho****, pána v botách* (A), *Vida je!,* ***Hybaj ke*** *koním, hybaj odtud* (ADV). je zřejmé, že jev valence tu není nijak výjimečný i to, že tu nejde o prosté kopírování, resp. extenzi valence verba (*kuk, šup, hybaj* k sobě jasná verba nemají).

Za zmínku ještě stojí i některé další případy. První je na úrovni kolokační poměrně známý a zřejmě svou omezenou kolokabilitou je vhodné ho řadit do frazeologie, srov. výzvu obvykle vůči malému dítěti *Udělej pa!* aj. Druhý pevně buduje na předchozím kontextu, jehož jednu část nahrazuje interjekcí, srov. *Visí na věšáčku ! –* ***Prdlajs*** *visí !* Zároveň je tento typ ilustrací na ještě jiný a obvykle nezmiňovaný případ skutečné a pevné integrace interjekce do věty.

Při bližším pohledu je však na distribuci těchto případů zjevné, že se řídí primárně **funkcí a sémantikou** interjekcí. V 2.1 výše se pod názvem lexikálně-sémantické kombinace stručně pár možností tohoto primárního sémantického půdorysu už naznačilo. Přijmeme-li podmínku dostatečně explicitního, nekráceného a nevágního kontextu (viz výše zde), jeví se sémantický základ a rámec užití většiny interjekcí jako přirozený a běžný, jakkoliv i zde bližší výzkum chybí. V úzu uváděný kontext naznačující daný význam a funkci může sloužit více cílům: zexplicitňuje, zdůrazňuje (což je jakási formálně-sémantická reduplikace), specifikuje jeden význam z polysémního, vysvětluje, resp. komentuje apod.; může se tak často jevit jako druh synonymního vyjádření, srov. příklady: ***Aha*** *vyhrkl s úlevou,* ***Hle*** *chyba/procesí pro maso.* ***Prásk!*** *děsné švihnutí bičem,* ***Prásk,*** *výpadek zvuku.* ***Proboha****, mějte s ní slitování.* Zajímavou otázku tu mj. představuje příklad s *hle*, který lze chápat jako synonymní s imperativum *podívejte se/pohleďte,* je však ekonomičtější a pohodlně vágnější. Poslední příklad navíc ukazuje, že mluvčí, vědomý si sémantické šíře a tedy vágnosti interjekce *proboha,* ji chce zjednoznačnit a zároveň signalizovat, proč ji užívá. Ještě lépe tuto sémantickou specifikaci příliš široké, resp. polysémní interjekce *fuj* ukazují čtyři následující příklady ***Fuj****, to je*

odporné!, ***Fuj!*** *Byl to strašný smrad.,* ***Fuj,*** *to jsem se lekla !,* ***Fuj****, styď se!*, které jsou v daném úzu nejen výhodné, ale zřejmě přímo nutné. Sémanticko-formální kolokabilita interjekcí a její potřeba je tu tedy dobře vidět. Poslední příklad navíc zaznamenává případ věcné a pragmatické implikace (vyplývání), která je zde řečena jasně a explicitně, aby posluchači byla zcela jasná. Takové vyplývání může však být spojeno ještě se zapojením indexálně-ikonického aspektu zvuku, srov. ***Chramst*** *a je po uzence.* Srovnejme, už bez komentáře, ještě další příklady (které lze snadno rozmnožovat), ***Pšt,*** *do voleb prosím nerušit,* ***Prásk*** *– ozvaly se dveře.* ***Proboha****, to ne!* ***Pst****, zašeptala po chvilce.* ***Šup*** *s ní do klece. A teď* ***šup*** *do postele.* ***Tradá*** *do lesíčka.* ***Uf,*** *člověk si musí vydechnout.* ***Vida*** *jak to tu poklidila.* ***Vida,*** *je hezká.* ***Haló*** *je tu někdo?* ***Haló*** *jste to vy doktore?* ***Haló*** *pomoc!* ***Haló*** *paní, slyšíte mne?****Haló****, otevřete!* ***Haló!*** *Je tam někdo? Kde se* ***hergot*** *couráš? Neštvete mě,* ***hergot.*** *Ukradli mu peněženku s další tisícovkou –* ***Inu*** *, auto není trezor.* ***Kuš!*** *křikl na hlasité hosty.* ***Kuš,*** *potvoro! zadupal…****Pa,*** *měj se pokud možno dobře.*

3 Aspekty sémiotické, zvláště pragmatické

Zastavme se tu už jen stručně aspoň u třech aspektů sémiotických. První typ (**1**) **sémantického vyplývání** úzu interjekce (sémanticky však ne nutně v tomto sledu), resp. jejího **navazování** na „kontext“ byl už naznačen výše. Patří sem případy jako *Ponebože, to je vůl., Fuj, to smrdí.* Jeho sémiotická povaha, jejíž formální stránka nebývá plně zřetelná a jednoznačná, je indexální, kdy užitím jednoho výrazu se ukazuje na užití druhého (zpravidla interjekce).

Druhý typ (**2**) má povahu **ikonickou**, (imitativní, zvukomalebnou) a je nejstarší avšak obvykle taky jedinou sémiotickou stránkou interjekcí, na kterou se poukazuje (srov. interjekce *haf* a štěkání psa, *cink* a zvuk vznikající při lehkém dotyku kovů či skla aj.); zároveň se právě v této oblasti jako jediné demonstruje, že jazyk a jeho prostředky mají přímý vztah (ne ale zcela exaktně) k vnější realitě a tyto znaky, resp. onomatopoia jsou jí tedy motivovány. Všimněme si však ještě aspoň třetího typu aspektu, který je pragmatický, a to specificky (**3**) **evaluativní**, hodnotící.

Srovnáním kontaktových interjekcí *hybaj, marš* aj., např.

(A) ***Hybaj*** *do postele*

či Čapkova příkladu s velmi dobrým kontextem, který jasně ukazuje sociální souvztažnost osob,

(B) *Potvora jedna zlořečená. Tak si to představ. Předevčírem mně dá šéf generálního štábu jeden spis a povídá. Hample, zpracuj to doma; čím míň lidí o tom ví, tím líp, v kanceláři ani muk; tak* ***marš****, máš dovolenou …*

je dobře vidět, že mj. i užitím příslušného interjekce se oslovený zařazuje situačně, sociálně n. jinak níže či jako takto už zařazený se považuje za snadný objekt rozkazu, manipulace apod., kterého si výše postavený či autoritativní mluvčí nemusí nutně ani vážit či ho respektovat aj. Srov. ještě dál obdobný

vztah (depreciativní), obvykle jen momentální, který se signalizuje užitím např. víceslovné frazeologické interjekce *To zrovna*! aj. Zcela jiný vztah přátelskosti a důvěrnosti vyjadřuje, vedle vlastní sémantiky uznání a respektu, naopak interjekce *panečku* aj.

4 Závěr. Otevřené otázky

Ukazuje se, že takovéto na korpusu založené studium do značné míry vyvrací nejstarší tradovou představu o izolovanosti interjekcí, která nebrala v úvahu to, že text je složitý celek, předivo především mnoha typů sémantických vztahů, kde izolovaný prvek zcela bez vztahu by neobstál. Čistě formálně založené přístupy tu, jak je vidět, selhávají. Detaily, podoby a míru naznačených syntagmatických souvztažností je však třeba teprve studovat.

Stejně tak zůstává otevřená ještě celá řada dalších otázek, z nichž některé zde ani nemohly být zmíněny. Necháme-li stranou možnosti paradigmatických aspektů jako celku, připomeňme aspoň heslovitě ještě:

1. míra ustálenosti některých kombinací (*ach co*!),
2. možnosti a míra extenze a tedy znejistění hranic morfému (*jééééééje*, *fííííííí* apod., viz též výš),
3. polyfunkčnost a překrývání, zvl. interjekcí a partikulí apod. (viz 1 výše a dál: *Sakra!* x *Von je* ***sakra*** *dobrej*),
4. funkční přechody, resp. transformace jinam (*Haló! x Haló noviny*, *hip hop* vs tanec a muzika *hip hop*),
5. variabilita (*hergot/hernajs/heršvec, proboha/propána/Prokristapána*),
6. rytmus a fonologie (*abrakadabra, bim bam, cupity dup*).

Bibliografie

1. Blatná R., 1996, Zvukomalba a pragmatika, In *Jazyk a jeho užívání.* Sborník k životnímu jubileu prof. O. Uličného,
2. FFUK eds. I. Nebeská, A. Macurová, 93–102.
3. Čermák F., M. Křen, ed. 2004, *Frekvenční slovník češtiny.* Praha NLN.
4. *Encyklopedický slovník češtiny,* ed. P. Karlík, M. Nekula, J. Pleskalová, NLN Praha 2002.
5. Havránek B., A. Jedlička, 1960, *Česká mluvnice,* Praha SPN.
6. *Mluvnice češtiny* I-III, 1986, 1987, Academia Praha.
7. Šmilauer V., 1972, *Nauka o českém jazyce* Praha SPN.
8. Vondráček M., Citoslovce a částice – hranice slovního druhu, *NŘ* 81, 29–37.

Appendix

Attribution-ShareAlike 3.0 Unported

License

1. Definitions

a. **"Adaptation"** means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

b. **"Collection"** means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in

unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.

c. **"Creative Commons Compatible License"** means a license that is listed at http://creativecommons.org/compatiblelicenses that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.

d. **"Distribute"** means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

e. **"License Elements"** means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

f. **"Licensor"** means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

g. **"Original Author"** means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

h. **"Work"** means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

i. **"You"** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

j. **"Publicly Perform"** means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

k. **"Reproduce"** means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

b. to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";

c. to Distribute and Publicly Perform the Work including as incorporated in Collections; and,

d. to Distribute and Publicly Perform Adaptations.

e. For the avoidance of doubt:

    i. **Non-waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

ii. **Waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

iii. **Voluntary License Schemes**. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.

b. You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US)); (iv) a Creative Commons Compatible License. If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the "Applicable License"), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of

each Adaptation You Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

c. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Ssection 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

d. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in

which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

b. Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

c. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

d. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

e. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

f. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Computer Treatment of Slavic and East European Languages

Editors
Jana Levická
Radovan Garabík

Návrh obálky: Vladimír Benko
Technický redaktor: Peter Luciak
Prvé vydanie. Vydalo vydavateľstvo Tribun,
v Brne roku 2007. 318 strán.