

# INSIGHT INTO THE SLOVAK AND CZECH CORPUS LINGUISTICS

Editor  
MÁRIA ŠIMKOVÁ



Bratislava 2006



VEDA

Publishing House of Slovak Academy of Sciences

Ľudovít Štúr Institute of Linguistics of the SAS



© Silvie Cinková, Radovan Garabík, Lucia Gianitsová-Ološtiaková, Markus Giger, Jan Hajič, Eva Hajičová, Veronika Kolářová-Řezníčková, Michal Křen, Markéta Lopatková, Karel Oliva, Jarmila Panevová, Vladimír Petkevič, Věra Schmiedtová, Miloslava Sokolová, Mária Šimková (ed.), Zdeňka Urešová

Zborník Insight into Slovak and Czech Corpus Linguistics bol pôvodne koncipovaný ako súbor prednášok, ktoré odzneli od 4. 11. 2002 do 28. 6. 2004 na pôde oddelenia Slovenského národného korpusu v Jazykovednom ústave Ľ. Štúra Slovenskej akadémie vied v Bratislave (<http://korpus.juls.savba.sk/activities/>). Nie všetci prednášajúci však dodali príspevok na publikovanie, iní svoj príspevok poskytli v doplnenej a prepracovanej verzii. Viaceré prednášky technického charakteru a témy workshopov ani neboli plánované na vydanie v textovej podobe, niektoré časti z tých, ktoré sa v zborníku nachádzajú, sú už v čase vydania prekonané. Naopak, do zborníka pribudli štúdie o budovaní Slovenského národného korpusu a jeho lingvistickej (najmä morfolologickej) anotácii. V tomto zložení predstavuje publikácia Insight into Slovak and Czech Corpus Linguistics vybraný súbor štúdií o stave korpusov a možnostiach i výsledkoch korpusovej lingvistiky v Slovenskej republike a v Českej republike v prvej polovici prvého desaťročia 21. storočia.

Proceedings entitled Insight into Slovak and Czech Corpus Linguistics were to have been a collection of lectures that had been delivered since November 4<sup>th</sup> 2002 until June 28<sup>th</sup> 2004 at the department of the Slovak National Corpus of the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences in Bratislava (<http://korpus.juls.savba.sk/activities/>). However, not every author had handed in his or her paper for publishing, others had sent them enlarged and rewritten. Several lectures being of more technical character as well as subject matter of workshops had not been even intended for publishing in print. Some parts of those, which are included into proceedings, are recently outdated. On the contrary, proceedings were enriched with papers on building of the Slovak National Corpus and on its linguistic (especially morphological) annotation. The content of the publication Insight into Slovak and Czech Corpus Linguistics represents a collection of selected papers reflecting the state and possibilities of corpora as well as the outcomes of Corpus Linguistics in the Slovak and Czech Republic in the first half of the 21<sup>st</sup> century.

© Ľudovít Štúr Institute of Linguistics of the SAS, Slovak National Corpus, 2006

ISBN 80-224-0880-8

## TABLE OF CONTENTS

**Věra Schmiedtová**

Corpora and Dictionaries ■ 7

**Michal Křen**

Frequency Dictionary of Czech: A Detailed Processing Description ■ 16

**Vladimír Petkevič**

Reliable Morphological Disambiguation of Czech: a Rule-Based Approach Is Necessary ■ 26

**Karel Oliva**

Discovering and Employing Ungrammaticality ■ 45

**Jan Hajič**

Complex Corpus Annotation: The Prague Dependency Treebank ■ 54

**Eva Hajičová**

Towards the Underlying Structure Annotation of a Large Corpus of Texts ■ 74

**Markéta Lopatková – Jarmila Panevová**

Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank ■ 83

**Zdeňka Urešová**

Verbal Valency in the Prague Dependency Treebank from the Annotator's Viewpoint ■ 93

**Silvie Cinková – Veronika Kolářová**

Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank ■ 113

**Markus Giger**

On the Delimitation of Analytic Verbal Forms ■ 140

**Mária Šimková**

Slovak National Corpus – History and Current Situation ■ 151

**Radovan Garabík**

Processing XML Text with Python and ElementTree – a Practical Experience ■ 160

**Lucia Gianitsová**

Morphological Analysis of the Slovak National Corpus ■ 166

**Miloslava Sokolová**

Options for the Generation of a Corpus-Based Slovak Morphology (as Part of Corpus Morphosyntax) ■ 179





# Corpora and dictionaries

VĚRA SCHMIEDTOVÁ

This text was developed and has been used as a basic study material for courses on corpus linguistics taught at the Charles University in Prague over the last couple of years. In addition, it was presented at a workshop in 2003 at the Institute for the Slovak National Corpus in Bratislava.

## INTRODUCTION

It is sometimes claimed that we are experiencing the age of dictionaries. In recent years, many new lexicographical organizations and institutes have organised specialized conferences where lexicologists from all over the world come together to evaluate and discuss theoretical as well as practical issues concerning their field of expertise.

(There are such conferences as Euralex, Afrilex, Asialex, and DSNA – lexicological organizations of North America). Also, many books and papers whose focus is lexicology have been published. For example, a very important piece of work is the *Encyclopaedia of Lexicography*, which was edited by Franz Josef Hausmann and to which 349 collaborators contributed their work. New specialized lexicographical journals have started to appear (Dictionaries, Lexicographica, International Journal of Lexicography), etc.

This boom has mainly affected the English-speaking countries, especially Great Britain. Naturally, other countries can learn from this turbulent development in documenting their languages, and can attempt to use the results already achieved.

My contribution pays attention to monolingual dictionaries, and addresses the current language situation in the Czech Republic compared with the situation in Great Britain.

After several years of improving conditions, modern lexicology is pausing for breath and all the data that have been collected and made ready by the Czech National Corpus are now ready to be processed and analysed. It is possible to say that, since the political changes in 1989, the Czech lexicology situation is still fragile; nevertheless, it is not as hopeless as previously was the case.

## THE TRADITIONAL WAY OF DEVELOPING A DICTIONARY

Prior to the use of computers, it was necessary to collect a large amount of language material in order to compile a new dictionary. The basic prerequisite method was excerption. That is, recording individual words and collocations in their respective contexts of occurrence onto excerptions slips. For more detail, see below.

## HISTORY OF MODERN CZECH ACADEMIC DICTIONARIES

We will not review all the dictionaries written in recent periods in the Czech language sphere, but instead focus on a few academic dictionaries, pertaining to Czech vocabulary from 1870 to 1978, whose origins goes back to the first years of the 20<sup>th</sup> century.

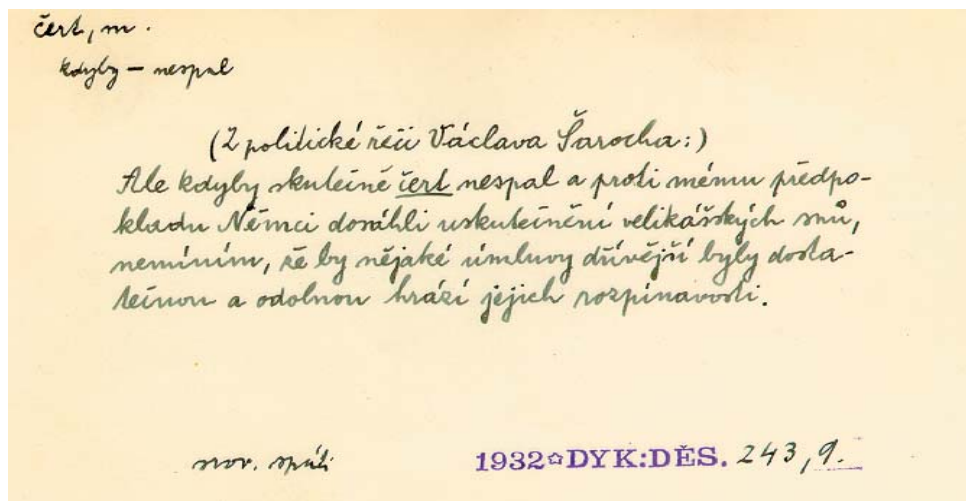
## 1 REFERENCE DICTIONARY OF THE CZECH LANGUAGE 1935–1957 (PS) (9 VOLUMES)

The lexicological archive where all the excerption slips are stored is located at the Institute for the Czech language in Prague. The archive has been operating since 1911 although preparations started in 1906. The main initiator of the entire project was Mr. František Pastrnek, who also became the president of the lexicological and dialectological committee of the Czech Academy of Science in 1911. In the same year, the Office of the Dictionary of the Czech language was established and, thereby, the first rules for scientific excerptions. The Office was situated in one room of the Hlávkův Palace in Jungmannova Street in Prague. The sole employee was an internal secretary by the name of František Trávníček. Soon afterwards, the Bureau expanded with new members from the ranks of high school teachers and university students. Interestingly, until the Bureau was changed into an academic institution, it was supported by the Ministry of Education by making high school teachers available for work on a new dictionary.

The future dictionary was based on excerptions from Czech vocabulary from 1770. Initially, so-called “total excerption” was applied, i.e. every word in a given text was documented in all its contextual occurrences, collocations, and idioms. A word in its dictionary form was the header of an excerption slip. That means that declined or conjugated words were to be found in nominative singular or in infinitive in the header; other possible collocations, further sufficient context where the form of a given word was underlined, and finally bibliographical information regarding the occurrence of the word were put in the commentary line.

In 1917, a second improved edition of the excerption rules appeared. The total excerption style was enhanced by a partial excerption and by specialized excerption that was dedicated to capturing linguistic peculiarities.

An example of an excerption slip:



Authors of books selected for excerption were supposed to be “the good authors”. Thus, every year, between 10 and 15 books were chosen. Subsequently, filters were set up. These were lists of words and their meanings that were already well documented. For these words, no additional excerption was required. In addition, for each period, a significant piece of work was singled out. Excerpts from such books formed the basis for a filter. For example, the novel *Babička* (Grandmother) by Božena Němcová was selected for the period 1854–1878.



The actual lexicographical work began in 1932. Originally, the idea was to publish a thesaurus of the Czech language, but in the end this plan was changed and it was decided to create the reference dictionary of the Czech language. The focus was on the current Czech language. For the description, excerpted data collected since 1870 were used. Older language material was used in order to check that the language link with the past was preserved. In 1935, the first issues of the Reference dictionary of the Czech language started to appear (15 issues in 1935, 19 issues in 1936). Besides the lexicographical research, linguists also continued with their excerption work.

The compilation of the dictionary was complicated and slowed down by both world wars. During the Second World War, after the Czech universities were closed down, the Office became a refuge for many university students and research assistants. In 1946, the Office was changed into the Institute for the Czech language. After the establishment of the Czech Academy of Science in 1953, the Institute for the Czech Language became a part of the Academy. Publication of the Reference Dictionary continued.

## **2 DICTIONARY OF THE LITERARY CZECH LANGUAGE 1951 – 1970 (SSJČ) (4 VOLUMES) (SECOND UNCHANGED EDITION FROM 1989) (8 VOLUMES)**

The last volumes of the Reference Dictionary of the Czech language appeared in 1957. From 1958 on, the first issues of the new dictionary began to appear – the Dictionary of the Standard Czech Language (in four volumes). In 1945, a second excerption data pool was established based on unprocessed language material from the Reference Dictionary. Between 1955 and 1969, a full excerption resource was completed for novels from authors such as Ivan Olbracht, Marie Pujmanová, or Marie Majerová. As can be seen from these names, the lexicographical work was not exempt from ideology. In 1969, a new set of excerption rules was established and the Czech lexicological team tried to create a new lexical standard. In contrast with the earlier decision to consider “good authors” only (see Ertl) excerptions were gathered from the texts of different authors coming from various functional styles (e.g. scientific written language, journalism). Additionally, a special status was assigned to spoken language and idiomatic expressions. The source for the documentation of the spoken language was modern Czech novels and the excerption criteria were rather vague. Nevertheless, these new rules and regulations did take into account the changes in the language situation at that time. It is necessary to stress here that this dictionary is still considered the largest synchronic description of the Czech Lexicon. Although its title alludes to a specific description of the standard (ie. written) Czech language, it also includes common (ie. spoken) Czech.

## **3 THE DICTIONARY OF LITERARY CZECH 1978 (SSČ) (1 VOLUME) (2<sup>nd</sup> REVISED EDITION 1994) (1 VOLUME)**

This dictionary is designated for use in schools and for the general public. It is a reduced version of the Dictionary of Literary Czech as described in 2 and it represents the Czech standard. It contains the core of the Czech vocabulary and grammatical information in a form that a standard Czech native user can refer to when in doubt.

## **ADVANTAGES OF COLLECTING LANGUAGE MATERIAL BY USING ELECTRONICALLY STORED TEXTS AND DISADVANTAGES OF COLLECTING LANGUAGE MATERIAL THE TRADITIONAL WAY**

Materials for new dictionaries assembled from electronic texts guarantee typicality, objectivity, and systematicity. Entire texts are stored and, depending on the query, where random use is

suppressed, a very specific search regarding all external annotations and internal grammatical tagging can be submitted. On the other hand, when a word was selected by manual excerption, the intuition or common sense of the excerptor played an important role. This factor makes the traditional excerption way less systematic and more subjective.

The old material, placed in an archive, is ordered alphabetically and hence the system is fixed. That means, for example, that it is impossible to extract the complete works of K.H. Mácha that are in full, are excerpted, and a part of this archive, from the alphabetically ordered body of the Lexical Archive of the Institute for Czech language. In addition, as pointed out above, obtaining language material by means of manual excerption is cumbersome and time-consuming. Collecting data and the conversion of texts for building electronic corpora are overwhelmingly complicated issues.

### THE LEXICAL ARCHIVE

According to calculations by the Institute for the Czech National Corpus, the Lexical Archive of the Institute for the Czech Language contains between 12 and 14 million excerption slips. Despite the confusion as to the precise number of these slips, we know that the excerption procedure upon which all existing dictionaries are based started in 1911. Until the 80's, more and more frequently, the standard excerption was enhanced by partial excerption or excerption of linguistic peculiarities. Then this type of excerption ended completely. Also in the 80's, the lexicographic activities around the Dictionary of the Literary Czech Language were terminated.

From the beginning, the aim of the Czech National Corpus was to collect language data for a new dictionary of a modern Czech language. Its last description was finished in 1970 when the last volume of the SSJČ appeared. The newer SSC dictionary, however, is based on the SSJC dictionary. In other words, there is a fifty-year gap in the continuity of the linguistic description of the current vocabulary of the Czech language.

### TYPES OF DICTIONARIES THAT CAN BE DEVELOPED ON THE BASIS OF A CORPUS

Every dictionary type requires a different type of corpus. For example:

- 1.1 Monolingual dictionary
  - Requirement: as extensive and diverse a corpus as possible
- 1.2 Translation dictionaries (bilingual)
  - Requirement: parallel corpora are now a prerequisite for a good bilingual dictionary
- 1.3 Frequency dictionary
  - Requirement: a representative large corpus
- 1.4 Author's dictionary
  - Requirement: a corpus based on all works written by a particular author
- 1.5 Terminology dictionary
  - Requirement: a corpus based on texts from a given field
- 1.6 Dictionary of neologisms
  - It is not possible in a corpus to establish neologisms automatically. However, a corpus aids verification of one's observations/intuitions.

The Czech National Corpus (CNC) is an extensive representative corpus of the synchronic written Czech language. It is eminently suitable for forming a frequency dictionary. The manuscript of such a dictionary is currently ready for publication. Another way of using the CNC is for grammatical description and monolingual dictionary (ie. 1.1 and 1.3).

## **HISTORY OF THE MOST RELEVANT CORPORA**

The real turning-point in the field of lexicography was the building of computer corpora. This was made possible by the invention of the computer.

### **1 SURVEY OF ENGLISH USAGE (1959 – 1989)**

This attempt comes from the pre-computer era and is associated with the name of Randolph Quirk from the University College in London. During the period from 1959 to 1989, he (and his colleagues) assembled a corpus of one million words. As a basis, he made use of 200 different samples, each containing 5000 words. One half consisted of written, the other half of spoken texts. In his corpus, Quirk tried to document English as used by British intellectuals. Each word was assigned a grammatical tag.

Later, Jan Svartvik, from the Lund University, transferred the spoken part of the corpus into a computer-driven version and added another 435,000 words of spoken English. In this way, the London-Lund Corpus of spoken English came into being. Svartvik completed this corpus in 1980. Both corpora were used for a ground-breaking and well-known book, the *Comprehensive Grammar of the English Language*.

### **2 BROWN CORPUS (1963–64)**

The first computer-based corpus of American English was created by Henry Kučera and W. Nelson Francis. This corpus contains one million word forms organized into 500 samples, each including 2000 words. Texts were selected from different genres and styles from the year 1961.

#### **2.1 LOB (LANCASTER – OSLO/BERGEN) 1978**

This corpus is also a corpus of British English using the same model as the Brown corpus. The initiator of this project is Geoffrey Leech from the University of Lancaster. The other two universities that took part in this project were the University in Oslo and the Norwegian Computing Center in Bergen. These two corpora made it possible to compare American and British English. On their basis, grammatical descriptions of both varieties were developed.

### **3 COLLINS – COBUILD CORPUS AND DICTIONARY**

The Cobuild project was founded as a joint project between the publisher, Collins, and the University of Birmingham in 1980. (The abbreviation *Cobuild* means “Collins Birmingham University International Language Database”.) The leading person in this project was John Sinclair. In 1982, the corpus consisted of 7.3 million words and it was called the Birmingham Corpus. In the 80's, it mainly included written texts. Later on, the corpus was extended by adding other British texts, some American texts, but also 26% of all texts were in the spoken language.

The Birmingham Corpus had 20 million words in 1987. In the same year, the Cobuild English Language Dictionary was published. This dictionary represents a new type of dictionary – grammatical information is organized in a new typographical way; new definition types are included (the descriptions are based on the speaker's point of view); occurrences found in the corpus are consequently used as examples. The Cobuild English Language Dictionary is intended to be used by students of English. This dictionary affected the way dictionaries were edited and published throughout Great Britain.

#### **3.1 THE BANK OF ENGLISH**

The Birmingham Corpus continues as the Bank of English. It is a so-called monitor corpus, i.e. an uncompleted corpus that monitors current changes in language. The name was changed

in 1991. This corpus contained 450 million words in January 2002. It includes complete texts of different types created mainly after 1990. The corpus is neither tagged nor lemmatized, but within its structure it is possible to create virtual corpora.

#### **4 THE BRITISH NATIONAL CORPUS (BNC)**

This corpus was established in 1991 and consists of 4000 samples. There are 90% written texts (75% informative texts; 25% artistic texts) and 10% spoken texts (spontaneous conversations). The British National Corpus has 100 million words and it claims to be a representative corpus of British English. It does not contain complete texts. The corpus was created in cooperation with several institutes and the Oxford University Press. The BNC is the basis for all dictionaries now published by Oxford University Press. A big supporter of this project was the British government which covered 50% of all costs.

#### **5 OTHER BRITISH CORPORA**

Over the years, individual dictionary publishers formed their own corpora that became the basis for writing and publishing new dictionaries for college and university students. These publishers are, for example, the Longman Corpus Network (in comparison with other corpora, the Longman corpus has a high percentage of novels) or the Cambridge International Corpus. In the latter corpus American texts are also included. The Cambridge International Corpus attained a fairly extensive size and in this respect was, for a time, comparable to the Bank of English.

### **DICTIONARIES BASED ON CORPUS DATA**

#### **A**

##### **STUDENT DICTIONARIES (ELT)**

The year 1995 was very productive for English lexicography. In this year all the leading British publishing houses published student dictionaries that were corpus-based with, in addition, a new graphical layout:

##### **1.1 COBUILD ENGLISH DICTIONARY**

##### **1.2 LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH**

##### **1.3 CAMBRIDGE INTERNATIONAL DICTIONARY OF ENGLISH**

All these dictionaries continue to be published in more recent editions. In this competitive environment, publishers keep bringing out new improvements that often fail to go beyond a few superficial features.

Just recently (in 2002), a really new student dictionary with an ELT focus appeared that learned from the previous attempts of other publishers. This dictionary is called the

##### **1.4 MACMILLAN ENGLISH DICTIONARY**

It is becoming more and more common that a CD-ROM is part of the standard equipment of a dictionary. Such a CD-ROM is suited to a quick and efficient search throughout the entire text.

#### **B**

The first corpus-based English dictionary for native speakers of English is

##### **1.5 THE NEW OXFORD DICTIONARY OF ENGLISH**

#### **C**

Different corpora present an ideal starting point for the compilation of a collocation dictionary. An example of this type of dictionary is the

## 1.6 OXFORD COLLOCATIONS DICTIONARY FOR STUDENTS OF ENGLISH

### THE CZECH NATIONAL CORPUS

It is the case that the British corpora and dictionaries based on these corpora were the models for the Czech language situation after the Velvet Revolution in 1989. Soon after this dramatic political change, an initiation group called The Czech Computer Fund was established. Its aim was to bring together all parties interested and/or already involved in computer-based language description and documentation and to seek financial means to start solving the difficult Czech situation in the field of lexicography and lexicology.

During several trips abroad (e.g. Great Britain or Italy), members of this group studied the advanced corpus situation in the other countries. After the political change, there was free access to international references and literature and hence it was finally possible to get a better impression of the new situation and the challenges connected with the project of writing a new corpus-based dictionary.

After several years of joint efforts, the Institute of the Czech National Corpus was established in 1996 at the Faculty of Arts at the Charles University of Prague. To establish such an Institute was a necessary prerequisite to starting a new Czech corpus-based dictionary system.

In 2000, the first 100 million representative corpus of current written Czech (SYN2000) was made accessible to the public (for more detail, see Čermák, Schmiedtová 2001). The representativeness of the Czech National Corpus is based on empirical sociolinguistic data (see table below). This research shows that, in contrast to the traditional excerpt method where the selection of texts was driven by the “good authors” principle, readers nowadays choose text types other than novels. For illustration, see the following overview:

#### ARTISTIC TEXTS 15 %

Novels	15 %
Poetry	0,81 %
Drama	0,21 %
Fiction	11,02 %
Other artistic texts	0,36 %
Transition between styles	2,60 %

#### INFORMATIVE TEXTS 85 %

Journalism	60 %
Specialized texts 2	5 %
Arts	3,48 %
Social science	3,67 %
Law and security	0,82 %
Natural science	3,37 %
Technology	4,61 %
Business and marketing	2,27 %
Religion	0,74 %
Life style	5,55 %
Administration	0,49 %

SYN2000 builds on complete written texts of the current Czech language. These texts are incorporated into the corpus according to the distribution outlined above. Additionally, finer grained criteria were created, which basically include works created after 1960 or published after 1990. The author, in whatever event, must have been born after 1880. For technical reasons, only journalistic texts from 1990 and later are part of the corpus. The selection criterion for specialized texts is that they were published after 1990. The SYN2000 has an internal tagging system and has been lemmatized.

#### **VIEWS ON THE WAY A COMPUTER-BASED CORPUS SHOULD BE BUILT**

Corpora and their creators have differing views on how a corpus should be built. It holds true that the model and distribution of different types of text should be suitable for the aims of the corpus in the first place.

When general corpora are created, people disagree on whether the corpus should fulfil some representativeness criteria or not. For example, the Bank of English view is that only a large spectrum of various text types can guarantee the real repetitiveness of a corpus. Not everybody has the same opinion regarding lemmatisation and tagging of a corpus. Another point of disagreement is the question of whether it makes sense to include complete texts or only samples. For example, the BNC includes samples only. The answer of the CNC to those questions is SYN2000.

#### **HOW BIG MUST A CORPUS BE IN ORDER TO SUPPORT THE WRITING OF A DICTIONARY?**

Many lexicographers have raised this question. The more corpus lexicography and linguistics grow, the stronger the opinion that corpora should be as big and as diverse as possible. Ramesh Krishnamurthy, one of the first members of the Cobuild team and today a collaborator of the University of Birmingham, said the following during an Internet discussion:

“Half of all word forms occur in the corpus only once. A dictionary entry cannot be established on the basis of one occurrence in the corpus database. For this purpose, we need at least ten different occurrences. However, many expressions that fulfil this criterion are not included in the corpus. These are, for example, numbers, proper names, etc. According to my calculations, a 100 million corpus would be enough for 45 thousand dictionary entries and this is the extent of a pocket dictionary. In lexicographical work, we need corpora that contain billions of words.”

Compare: The dictionaries mentioned above have the following scope:

- Longman 1995 – 80 000 entries;
- Cambridge 1995 – 100 000 entries;
- Cobuild 1995 – 75 000 entries.

That means that a corpus of 100 million word forms is not adequate for the compilation of such dictionaries.

#### **A NEW CZECH DICTIONARY**

Based on what we have learned from other parties – especially from the British experience – it is safe to assume that the new Czech dictionary will be a corpus-based dictionary. It will be a dictionary of a medium to large size and it will describe the vocabulary from a completely new point of view. The entries will supply information about the valency of words, word collocations, frequencies of meaning, occurrences of idioms, and it will avail itself of a more effective graphic structuring of the text. These are only some of the aspects that will be found in the new dictionary.

## BIBLIOGRAPHY

- BEJOIN, H. (2003): *Modern Lexicography: Past, Present and Future*. Proceedings ASIALEX 2003 Tokyo, Japonsko.
- Cambridge International Dictionary of English 1995.
- COLLINS Cobuild English Dictionary 1995.
- ČERMÁK, F., Schmiedtová, V. (2001): *The Czech National Corpus: Its Structure and Use*. PALC 2001: Practical Applications in Language Corpora, Lodž, Polsko, ed. Barbara Lewandowska-Tomaszczyk, PETER LANG Europaeischer Verlag der Wissenschaften.
- ČERVENÁ, V. (1981): *Kancelář Slovníku jazyka českého. Naše řeč 1981/1*, Praha.
- LANDAU, S. I. (2001): *Dictionaries, The Art and Craft of Lexicography*; 2<sup>nd</sup> edition; Cambridge University Press, Great Britain.
- Longman Dictionary of Contemporary English 1995.
- Macmillan English Dictionary 2002.
- Oxford Collocations, dictionary for students of English 2002.
- SINCLAIR, J. M. (1988): *Looking up*. Collins Cobuild, Great Britain.
- The New Oxford Dictionary of English 1998.
- 25 let Ústavu pro jazyk český (K 25. výročí založení Ústavu pro jazyk český a k 60. výročí vytvoření akademické Kanceláře Slovníku jazyka českého). Tiskem ÚJČ 1971.
- ERTL, V. (1929): *Dobrý autor*. In: *Časové úvahy o naší mateřštině*, Praha.

## ABSTRACT

Říká se, že naše doba zažívá rozkvět slovníků. Tento rozkvět zasáhl hlavně anglicky mluvící země, speciálně Velkou Británii. Ostatní země a jejich jazyky se ovšem mohou na tomto bouřlivém vývoji poučit a snažit se dosažených výsledků využít.

Náš příspěvek se věnuje jednojazyčným anglickým výkladovým slovníkům vzniklým s pomocí korpusu a pojedná o historii moderní české lexikografie a o české situaci výkladové lexikografie na pozadí situace anglické.

Co se týká češtiny, současná lexikografie po letech rozkvětu teprve nabírá dech a připravuje se zpracovat materiál, který jí připravil projekt Českého národního korpusu, který je stručně ve článku také popsán.

Dá se říct, že po politickém zlomu v roce 1989 byla a stále ještě je česká lexikografická situace neutěšená, není už ovšem beznadějná, tak jak byla.

Tento článek vznikl v roce 2004 jako písemná podoba přednášky ve studentském semináři Korpusová lingvistika na FF UK Praha, v rámci Výzkumného záměru č. MSM0021620823.

This paper was developed in 2004 as a written talk given in the course 'Corpus Linguistics' at the Charles University Prague, Research Grant Number MSM0021620823.





# Frequency Dictionary of Czech: A Detailed Processing Description

MICHAL KŘEN

## 1 INTRODUCTION

This paper describes in detail all the steps leading up to compiling the recently published Frequency Dictionary of Czech (FDC) [1,4]. As a base material for the dictionary, corpus SYN2000 was used. It is a 100-million corpus of contemporary written Czech, whose composition reflects the survey of written language reception. It was composed at the Institute of the Czech National Corpus and is comprised from 15 % fiction, 25 % specialized professional literature and 60 % newspapers and magazines [5]. All the texts in the corpus are from the 1990s, the only exception being fiction, which can be older. The whole corpus was morphologically tagged and lemmatized by tools developed at the Institute of Formal and Applied Linguistics, Charles University, Prague, under the supervision of Jan Hajič.

One of the aims of this paper is to show that, even with the powerful tools available, it would simply not be possible to print out the lemma list as a frequency dictionary. Due to the need for extensive manual corrections of the lemmatization described hereafter, compiling the FDC turned out to be more complicated than it was initially expected. However, after all the corrections, a new corpus FSC2000 was created, whose significantly improved lemmatization was finally projected into the quality of the whole dictionary.

## 2 DESCRIPTION OF THE DICTIONARY

The concept of the FDC comprises several characteristic features:

- it lists proper names and abbreviations separately in special dictionaries,
- it shows the distribution of occurrences across the three main genres (fiction, specialized professional literature, newspapers and magazines),
- it uses average reduced frequency (ARF) as a main measure of word commonness instead of usual absolute frequency (FRQ – total number of occurrences of all forms of a given word).

The value of ARF is always equal to or less than FRQ, reflecting the evenness in distribution of occurrences of a given word in the corpus: the more even the distribution, the closer the value of ARF approaches to FRQ and vice versa [6]. The ARF of evenly distributed words is typically around a half of their FRQ, but it gets considerably – ten times or more – smaller than FRQ for words that occur only in a few sources (technical terms, proper names etc.). For example, the words *antigen* (*antigen*) and *kopanec* (*kick*) both have the same FRQ in the corpus (221). While *kopanec* is a widely comprehensible word evenly distributed in the corpus, the occurrences of *antigen* are concentrated mainly in a few medical texts. This is reflected by ARF: its value for *kopanec* is 117, while for *antigen* it is only 26, i.e. more than four times smaller. ARF thus reduces the excessive influence of such words caused by the fact that any corpus is purely a sample.



FDC consists of five dictionaries:

- An alphabetically sorted dictionary of the most frequent common words (50,000 entries). This is the key dictionary; its entries were chosen according to ARF and for each entry it lists the values of both ARF and FRQ together with their corresponding ranks as well as the genre distribution.

- A dictionary of the most frequent common words (20,000 entries) sorted by FRQ. Its entries list only basic data with references to the wider numeric description in the key dictionary.

- A dictionary of the most frequent common words (20,000 entries) sorted by ARF. Similarly, its entries list only basic data with references to the key dictionary.

- A dictionary of the most frequent proper names (2,000 entries) sorted by ARF.

- A dictionary of the most frequent abbreviations (1,000 entries) sorted by ARF.

In addition to these dictionaries, the FDC also contains a list of the most frequent graphemes and punctuation marks, as well as a brief research into the lexical coverage of text.

Here is an example of the entry for *veletrh* (*trade fair*) in the key dictionary:

entry	Rank ARF	ARF	Rank FRQ	FRQ	fiction	prof.lit.	news
veletrh	3040	1482	1748	6807	1%	65%	34%

The word *veletrh* occurred 6,807 times in the corpus, but its ARF is only 1,482, i.e. more than four times smaller. It indicates that its distribution is not very even, and this finding is further confirmed by the genre distribution. It is reflected also in the ranks: *veletrh* is the 1,748th most frequent word according to FRQ, but only the 3,040th according to ARF, which shows that FRQ was reduced more than an average.

In its list of entries, FDC distinguishes homonyms only if they are different parts of speech (PoS); in such a case, the PoS abbreviation is given after a slash, e.g. *obrat/N* as a noun (*turnover*) vs. *obrat/V* as a verb (*to rob*). However, no distinction is made when different PoS only describes different roles of a given word in the sentence. Typical examples include particles or common relation between preposition and adverb (e.g. *okolo* (*around*) is only one entry in the dictionary).

The printed version of the FDC is accompanied by a CD containing a complete electronic form of all the dictionaries – common words, proper names and abbreviations. Its service program EFES enables users to view and re-sort the dictionaries included, to search in them according to several criteria, and to save the results to the clipboard for further processing. In addition to the CD, corpus FSC2000 has been made available on the internet – for more information see <http://ucnk.ff.cuni.cz>. The corpus is a complementary and reference entity to the FDC and its lemmatization exactly corresponds to that of the dictionary. Thus the users can easily find out which word forms prevail in the corpus or what context is typical for a given dictionary entry, or they can perform a statistical analysis of the collocates etc. The combined concept of the FDC – a printed dictionary with accompanying CD and the corpus – hence enlarges the number of possibilities of how to further exploit the information value of the dictionary.

### 3 PROCESSING PROCEDURE OVERVIEW

As a basis for the dictionary, corpus SYN2000 was used, the morphologically tagged and lemmatized 100-million representative corpus of contemporary written Czech. However, it

was not possible to compile the dictionary directly from the existing corpus without the following fundamental changes:

- Several duplicate texts were discarded from the corpus together with parts of other texts that contained mostly tables, numbers etc. Although a similar clean-up had been carried out previously, it proved to be insufficient. All the discarded texts made up about 5 % of the original corpus, so that the size of the new one – FSC2000 – was reduced to about 95 million tokens. Due to the reduction, the corpus texts have undoubtedly gained a better quality, which has increased the overall reliability of the dictionary data.

- New versions of morphological tagging and lemmatization tools were applied on the new corpus. These tools automatically assign a lemma and morphological tag to each token in the corpus [2,3]. The processing has two stages: during the first stage (morphological analysis), each token is assigned with all possible morphological interpretations – a set of pairs consisting of a morphological tag with corresponding lemma. This is done regardless of context, so that the same word forms always get the same set of pairs. In the second stage (disambiguation), stochastic methods are used for choosing the most probable tag-lemma pair depending on the context of each token. It should be pointed out that Czech is a language with a relatively free word order, rich inflection and a high degree of homonymy, so the problems concerning tagging are of a different nature than in English.

- The output of the lemmatization was still inappropriate for the dictionary and it required extensive and mostly manual corrections. There were generally two sources of imperfections: homonymy-caused errors in stochastic disambiguation, when a lemma was selected incorrectly from the set offered by morphological analysis, and the concept of morphological analysis itself, as it treated some language phenomena unsuitably. Both types were basically handled independently; the correction procedures are described in the next two chapters.

It should be pointed out that no attempt was made to correct also the morphological tags, all the effort was aimed at correcting the lemmas only. The reason for this decision is that if a given token has wrong lemma, it is unlikely that the morphological tag could be correct. However, if the lemma is correct, it is not improbable that there could be an error in the morphological tag (determination of case, number etc.) due to very common homonymy within the paradigms. Thus the number of incorrect morphological tags is several times greater than the number of incorrect lemmas, which makes the cost of correcting them too high. On top of that, it would not improve the dictionary as such, since it is the lemmatization that is essential for counting word frequencies, although correcting the tags would permit the inclusion of a set of useful tables concerning the most frequent parts of speech, noun cases, verb tenses etc. in the appendices.

#### 4 CORRECTIONS OF MORPHOLOGICAL ANALYSIS

Although the dictionary used by the morphological analysis was comprehensive, its concept included several questionable features, mainly the inappropriate or inconsistent treatment of some grammatical phenomena:

- pluralia tantum (e.g. *brambůrky* (crisps) lemmatized *brambůrek*, even if the singular is very rare);
- spelling variants (e.g. both *stadion/stadión* (stadium) lemmatized *stadión*, while for *citron/citrón* (lemon) both lemmas existed);
- negations (e.g. *nežádoucí* (undesirable) lemmatized *žádoucí* (desirable), although it is a negation with semantic shift) etc.

In addition, there was also an enormous number of proper names, abbreviations, words from foreign languages and various errors in the corpus that made the processing even more difficult. Unfortunately, the morphological analysis sometimes failed in their detection. It should be mentioned that, during the conversions of texts into the corpus, a module for detecting and discarding foreign texts at the paragraph level was involved. This means that the corpus would not contain any larger part of text written in a language other than Czech (abstracts, summaries, advertisements, Slovak newspaper articles etc.). Of course, this does not apply to short quotations that are considered to be an integral part of the text, and that have thus become the main source of foreign language words in the corpus. The following examples show that proper lemmatization is not at all easy and that unexpected meanings of individual word forms can occur – and even prevail – in the corpus.

- the form *an* can be either archaic Czech relative pronoun (ca 5 % of occurrences in the corpus), or also English indefinite article (ca 20 %), German preposition (ca 20 %), part of Chinese or other name (rather surprisingly 50 %) or a typing error (ca 5 %);
- one would expect that the form *sky* can only occur in an English context or as a part of an English loanword, but almost 90 % are journalistic initials or typing errors;
- the frequent Czech surname *Čermák* (1,430 occurrences) was not present in the dictionary used by the morphological analysis; thus, all its occurrences were lemmatized as the common noun *čermák* (*redstart*), which is very rare in the corpus;
- each of the forms *PES*, *PSA*, *PSU*, *PSE*, *PSI* can be interpreted as a part of paradigm of the word *pes* (*dog*) written in uppercase, but each of them can also be an abbreviation, whether common or rare and unexpected.

As a base for the corrections, the first draft version of the dictionary was simply printed out of the originally lemmatized corpus. Proper names and abbreviations were not yet listed separately, so the whole dictionary consisted of only one list sized at more than 83,000 entries. The size was determined in order to get a list that would contain 60,000 of the most frequent common words according to ARF, i.e. the size of the key dictionary (50,000) plus a reasonable reserve (10,000). This number of common words contained more than 23,000 of proper names and abbreviations, or to be more precise lemmas with the first letter in uppercase – the original lemmatization was not very reliable in this respect.

The first inquiries into the list showed that the extent of the corrections necessary would be enormous. A system of corrective operations was developed and tested by means of the inquiries, so that any kind of error in the list could be fixed by a manually-determined sequence of these operations. It is important to note two things: first, all the operations dealt only with the lemmatization (i.e. the markup), so individual tokens in the corpus remained untouched. Second, the meaning of “kind of error” is purely technical and has nothing to do with various kinds of language phenomena. Generally speaking, correspondence between the language phenomenon and the kind of corrective operation was very loose. The individual operations were suggested by linguists who went through the list thoroughly. Their previous experience with the lemmatization was invaluable for their anticipation of imperfections that were often not obvious at first glance. The whole list was checked three times, each time by a different person, in order to minimize the number of inconsistencies and omissions. This was the most laborious part of the corrections and it took several linguists almost half a year, devoted mostly to the dictionary.

Two kinds of corrective operations were applied:

- Corpus operations – operations performed directly on the lemmatization of the corpus. They included determining a new lemma for a given word form (all the occurrences regardless of context or actual lemma) and determining a new lemma for all forms of given lemma (i.e.

renaming the lemma). The latter was used also for joining two lemmas (if the new lemma had already existed) or e.g. for renaming singular lemmas to plural in cases of pluralia tantum. In addition to these operations, it was also possible to mark any word form as ambiguous, i.e. to add it to the list of ambiguous word forms (group 1 in the next chapter). However, in the vast majority of cases the individual word forms did not need to be re-lemmatized because of their ambiguity, so this option was used very rarely.

– List operations – operations to be performed later on the next version of the list of entries. The lemmatization of the corpus was not affected by them. They included deliberate omissions of a given entry from the dictionary (mainly various errors or words from foreign languages that occurred mostly in foreign language contexts, e.g. *see, und, an, pej, sky* etc.), completion of dictionary entries (with spelling variants, reflexive *se, si* etc. – they are given in the dictionary, but not in the corpus), addition of plus signs (they mark entries which occur mostly as parts of multi-word units, e.g. *zbla+, break+*) and *viz* references (e.g. *dbalý viz nedbalý*).

The total number of all the corrective operations performed was 19,149, the most frequent of them being the corpus operations (12,397). After all of them were ready in separate lists, the lemmatization of the corpus was processed and altered using the corpus operations. This permitted the re-computing of all necessary frequencies, including ARF and genre distribution. Then the second version of the dictionary was printed out of the corpus and modified using the list operations. As a side-effect of performing the corrective operations, it was possible to separate the lists of proper names and abbreviations, since the lemmatization had become much more reliable. A similar correction procedure was also used for the final revision of the dictionary (see chapter 6 for processing chronology overview).

Let us now refer back to the earlier examples:

- word forms *an* and *sky* remained lemmatized *an* and *sky* in the corpus, but they were omitted from the dictionary due to their context;
- lemma *čermák* was renamed to *Čermák*, thus correcting the lemmatization of all its forms as well;
- lemmatization of forms *PES, PSA, PSU, PSE, PSI* was not changed at all; their lemma remained *pes*.

The last example merits a brief analysis and explanation: the primary aim of the frequency dictionary is to give the correct frequencies of words (not e.g. word forms), and all the corrections were carried out bearing this in mind. Due to the time limitations, only very little effort could have been devoted to determining the correct lemma for word forms, whose frequency had minimal influence on the overall frequency of any dictionary entry. It means that correct lemmatization cannot be granted for every single token. This is the case of the abbreviations above: the most frequent of them is *PES*, which occurs 84 times in the corpus. All its occurrences are lemmatized *pes*, which is correct for 62 of them, the rest being various abbreviations. In view of the total frequency of the lemma *pes* (13,091), the error caused by incorrect lemmatization of form *PES* is insignificant (0.17%) and can be ignored, especially when the lemma *PES* does not exceed the frequency limit for inclusion into the abbreviations dictionary.

Apart from the manual corrections described so far, some features were corrected automatically. Although the automatic corrections were applied wherever possible, the scope of their use was limited to the following cases:

- recognition of multi-word prepositions (e.g. *v rámcí*),
- revision of lemmatization of personal and possessive pronouns (e.g. lemmatization of form *nám* (*to us*) was changed from *já* (*I*) to *my* (*we*)),

- detection of auxiliary forms of the verb *být* (to be) in order not to influence the total frequency of the verb (e.g. frequency of the form *budete* in expression *budete muset* (you will have to) did not increase the frequency of the lemma *být*).

## 5 CORRECTIONS OF DISAMBIGUATION

This chapter describes verification of the output of the original stochastic disambiguation. In the beginning, a list of all ambiguous word forms together with the lemmas they had been assigned with by the morphological analysis has been printed out of the corpus. These were precisely all the word forms which could have been incorrectly disambiguated. It should be mentioned that the corrections of disambiguation were treated independently of the corrections of morphological analysis, because they were of a different nature and their overlapping was rather rare. The only problematic point was that the lemmas in the list were generated by the morphological analysis, and thus were sometimes renamed during the corrections. However, it was not complicated to discover the new lemma. Therefore, it was possible to keep the two correction methods basically independent, with only a slight effort devoted to maintaining correct references between them.

Another important point is that the list of ambiguous word forms was almost exhaustive, due to the comprehensive dictionary of the morphological analysis (especially in the case of Czech common words), so additions to it were infrequent. The list contained approximately 76,000 different word forms, but it was reduced to about 16,000 whose frequency in the corpus was greater than 20. This cut-off point is questionable from the present point of view and a lower frequency limit would probably be selected nowadays, although the size of the reduced list would greatly increase. The reduced list was then manually divided into the following five groups, and this categorization was always verified on random samples from the corpus. The approximate size of each group is given in the parentheses:

1. the suggested ambiguity is real and the lemma must be determined with respect to the context (1,200)
2. the suggested ambiguity is theoretically possible, but very unlikely; it is thus safer to lemmatize all such forms directly regardless of the context, rather than to rely on the disambiguation (6,700)
3. the suggested ambiguity is either not real or the suggested lemmas will not be distinguished in the dictionary, so that there is unambiguously only one possible lemmatization of the given form (2,300)
4. the suggested ambiguities differ only in the case of the first letter; the lemma will be determined according to the position of the form in a sentence (3,000)
5. a subcategory of group 2 (unlikely ambiguity), introduced due to the similar nature of its elements (2,500)

Examples (rejected lemmas are in square brackets):

group	word form	suggested lemmas	
1	je	být (to be)	oni (they)
1	jedli	jíst (to eat)	jedle (fir)
2	pilo	pít (to drink)	[pila (saw)]
2	patře	patro (floor)	[patřit (to belong)]
3	k	k (to)	[kuo (?)]
3	folkloru	folklor (folklore)	[folklór (spelling variant)]

4	Procházku	Procházka ( <i>surname</i> )	procházka (walk)
4	Ostrov	Ostrov ( <i>place name</i> )	ostrov (island)
5	nařízení	nařízení (command)	[nařízený (commanded)]
5	povolení	povolení (permission)	[povolený (allowed)]

All the word forms in groups 2, 3 and 5 were lemmatized automatically, and the manually selected lemma was assigned to all of their occurrences regardless of context. Forms in group 4 were also lemmatized automatically, but according to the position of the form in the sentence. There remained only about 1,200 forms in group 1 that required further manual processing because of their real – rather than theoretical – ambiguity. Therefore, at the cost of a minor trade-off, the size of the original list was significantly reduced.

Thus far, it was possible to lemmatize a given word form automatically, mostly regardless of the context. However, this is not conceivable for the forms in group 1, where the context is essential for determining the correct lemma. Therefore, an extensive manual verification of the disambiguation of these ambiguous forms was carried out. For each of them, random samples from the corpus were used – sized 100 concordances – to determine what share of the given word form's occurrences should be lemmatized with the particular lemma. There were at least three samples used for each form, the final figure being an average of the partial results. This way “ideal” shares were obtained for each of the word forms, that were used for the verification of results of the stochastic disambiguation by comparing them with the real shares extracted from the lemmatized corpus. The difference between them constituted a partial correction (signed plus or minus), that was temporarily stored together with the corresponding lemma. Its meaning could be described as “the number that should be added to the lemma's frequency in order to correct the influence of the incorrect disambiguation of one of its forms”. If the lemma had several ambiguous forms, the partial corrections were added up to constitute the final correction of the given lemma. If it was greater than 5 % of the lemma's total frequency, the corresponding dictionary entry was marked with an asterisk to indicate that its frequency is not accurate. In this case, the numeric value of the final correction was also shown on the line below together with all the ambiguous forms of the lemma that had caused it.

Let us take the word form *bouří* as an example. It occurs 518 times in the corpus, 75 % (388 occurrences) being lemmatized *bouře* (*storm*) and 25 % (130 occurrences) *bouřit* (*to rage*). During the verification, different results were obtained: only 60 % should have been lemmatized *bouře* and 40 % *bouřit*. This means that, due to this ambiguity, the partial correction is -78 (15 % of 518) for lemma *bouře* and +78 for lemma *bouřit*. Because both the lemmas have no other ambiguous form present in group 1, the partial corrections are also the final ones. They are compared with the total frequencies of both lemmas that are 2,333 (*bouře*) and 420 (*bouřit*), only one of the corrections exceeding the 5 % limit. The result is illustrated by the two following dictionary entries: one of them is marked with an asterisk and the correction (in % of the total frequency) is given on the next line, while the other one is not:

entry	Rank ARF	ARF	Rank FRQ	FRQ	fiction	prof.lit.	news
bouře	3777	1107	4092	2333	52%	21%	27%
*bouřit (se)	11153	219	12995	420	50%	20%	29%

korekce +19%, tvar bouří



Perhaps it should be explained why such a laborious process has been carried out only to pinpoint the entries with inaccurate frequencies. It would certainly be possible to add the final correction to the frequency of each entry so as to get the correct numbers if only FRQ was used in the dictionary. However, in order to be able to compute ARF and genre distribution, it is necessary to know the exact positions of all the occurrences of a given lemma in the corpus. This becomes a real problem with highly frequent ambiguous word forms, as not only random samples would have to be verified, but in effect every token. Since the total frequency of all the ambiguous forms that caused their corresponding lemmas to be marked with asterisks is bigger than one million, the re-computing of ARF and genre distribution is not possible without an enormous amount of additional manual work. Therefore, it was decided to use the asterisk symbol as a simple and feasible correction method.

Finally, it should be pointed out that the corrective asterisks should be regarded positively as a feature of assurance that is only rarely present in similar dictionaries. Extensive verification of both morphological analysis and disambiguation was carried out, during which all the errors discovered were corrected, the only exception being the disambiguation corrections that are marked with asterisks in the dictionary (453 entries, i.e. less than 1 %). Although there is a strong possibility that some of the errors remained undiscovered, the absence of an asterisk in front of the individual entry shows that the given frequency data are presumably accurate, perhaps only with the insignificant error rate explained above. The open opportunity to access and query the dictionary's base corpus FSC2000 on the internet provides even more assurance to the users of a certain reliability of the dictionary data.

## 6 CONCLUSION

The following scheme summarizes the processing chronology:

- ❖ morphological analysis
- ❖ partial disambiguation (word forms from groups 2, 3, 4 and 5 only)
- ❖ full stochastic disambiguation
  - first draft version of the dictionary
- ❖ manual detection of disambiguation errors (group 1) – simultaneous with the two following steps
- ❖ manual corrections of morphological analysis
- ❖ automatic corrections (multi-word prepositions, personal and possessive pronouns, verb *být*)
  - second version of the dictionary in proper output format
- ❖ final revision
  - final version of the dictionary

The data in the dictionary are supported by the size and representativeness of the corpus, as well as by extensive professional corrections of its lemmatization. However, the authors do not claim that the dictionary is free from errors or omissions, because the manual processing of large and complex data cannot avoid them, but their number was undoubtedly minimized. Nevertheless, the dictionary offers an objective overall picture of contemporary written Czech, and its practical applications extend widely from lexicography to information technology.

## 7 ACKNOWLEDGEMENT

Many thanks to all the people of the FDC processing team, whose devotion and thoroughness were essential in compiling the dictionary.

## 8 REFERENCES

- [1] ČERMÁK, F. – KŘEN, M. (et al.) (2004): Frekvenční slovník češtiny. NLN, Praha.
- [2] HAJIČ, J. – VÍDOVÁ-HLADKÁ, B. (1998): Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: Proceedings of the Conference COLING – ACL '98, Montreal.
- [3] HAJIČ, J. (2004): Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Praha.
- [4] JELÍNEK, J. – BEČKA, J. V. – TĚŠITELOVÁ, M. (1961): Frekvence slov, slovních druhů a tvarů v českém jazyce. SPN, Praha.
- [5] KOCEK, J. – KOPŘIVOVÁ, M. – KUČERA, K. (eds.) (2000): Český národní korpus: Úvod a příručka uživatele. ÚČNK FF UK, Praha.
- [6] SAVICKÝ, P. – HLAVÁČOVÁ, J. (2002): Measures of Word Commonness. In: Journal of Quantitative Linguistics, pp. 215–231, no. 3, vol. 9.

## ABSTRACT

Příspěvek se zaměřuje především na podrobný popis kroků, které předcházely vytvoření Frekvenčního slovníku češtiny (FSC). Základem pro tvorbu FSC byl stomiliónový reprezentativní korpus současné psané češtiny SYN2000, který byl navíc lemmatizován a morfologicky označován. Tento korpus však přesto nebylo možné použít přímo, vytvoření slovníku předcházelo vyřazení některých technicky nekvalitních textů, a zejména rozsáhlé opravy lemmatizace, které byly převážně manuální. Výsledkem těchto oprav, které jsou popsány dále, je korpus FSC2000, který svojí lemmatizací plně odpovídá slovníku.

FSC se skládá celkem z pěti hlavních slovníků a několika dodatků, přičemž propria a zkratky uvádí odděleně od apelativ (slov obecných). FSC dále kromě běžných frekvenčních údajů udává pro všechna hesla také rozložení jejich výskytů po třech hlavních žánrech. Hlavním charakteristickým rysem slovníku však je, že namísto běžně užívané absolutní frekvence (počet výskytů všech tvarů daného slova v korpusu) používá pro posuzování běžnosti slov průměrnou redukovanou frekvenci (ARF), která kromě prostého počtu výskytů odráží také rovnoměrnost jejich rozložení v korpusu. Tištěná verze slovníku je dále doprovázena CD s elektronickou verzí hesláře, a také zpřístupněním korpusu FSC2000 na internetu.

Výše zmíněné opravy lemmatizace je možné rozdělit do dvou skupin, které byly zpracovávány v zásadě nezávisle na sobě. První skupinu tvoří opravy chyb ve stochastické desambiguaci, které byly způsobeny homonymií, druhou pak opravy dané koncepcí morfologické analýzy samotné, v jejímž jinak velice rozsáhlém slovníku byly nevhodným nebo nekonzistentním způsobem zpracovány některé jazykové jevy (např. negace, pomnožná substantiva, pravopisné varianty atd.). Opravy se však v obou případech týkaly pouze lemmatizace, nikoli morfologických značek. Z časových důvodů se dále nebylo možné zabývat lemmatizací každého výskytu každého slovního tvaru, rozsah oprav byl proto omezen pouze na tvary, jejichž frekvence v korpusu byla dostatečná k ovlivnění celkové frekvence některého ze slovníkových hesel.

Opravy morfologické analýzy vycházely z hesláře, založeného ještě na původní lemmatizaci. Pro tento účel byla vytvořena sada korekčních operací, které jednak opravovaly vlastní lemmatizaci korpusu – provedením těchto operací nakonec vznikl korpus FSC2000 –, a zároveň také doplňovaly novou verzí hesláře, který měl na základě tohoto korpusu teprve vzniknout (šlo např. o doplňování pravopisných, hláskových a jiných variant daného hesla, reflexivního *se, si* apod.). Použití jednotlivých korekčních operací v konkrétních případech určovali lingvisté při procházení celého původního hesláře, což byla nejnáročnější část celého zpracování. Kromě těchto manuálních korekcí proběhly ještě opravy automatické, jejichž rozsah byl však omezen pouze na nevlastní předložky, pomocné tvary slovesa *být* a osobní a přivlastňovací zájmena.

Na počátku oprav desambiguace byl seznam všech homonymních slovních tvarů v korpusu. Tento seznam byl manuálně rozdělen do několika skupin podle toho, zda šlo o homonymii reálnou (např.



tvary *je* nebo *jedli*) či spíše pouze teoretickou (např. tvary *pilo*, *patře* nebo *povolení*). Ve všech případech kromě první skupiny (reálné homonymie) bylo možné přiřadit danému slovnímu tvaru lemma bez ohledu na kontext s tím, že možná chyba bude zanedbatelná. V případě první skupiny to však možné nebylo, a proto byla manuálně ověřena desambiguace všech těchto reálně homonymních tvarů. Pokud byla zjištěná chyba způsobená stochastickou desambiguací příliš velká, byla hesla odpovídající těmto tvarům označena hvězdičkou a doplněna příslušnou korekcí frekvence.

Jedním z cílů tohoto příspěvku je ukázat, že i když byl k dispozici vhodný lemmatizovaný korpus, nemohla být tvorba frekvenčního slovníku otázkou pouhého vyjetí hesláře z korpusu, ale naopak byly nezbytné rozsáhlé opravy lemmatizace. Autoři jsou si vědomi toho, že slovník může i přes veškerou pečlivost při zpracování obsahovat chyby, protože těm se při manuálním zpracování velkého množství dat úplně vyhnout nelze. Přesto však věří, že FSC jako celek přináší objektivní obraz současné psané češtiny, a že tedy bude užitečným zdrojem informací pro široký okruh uživatelů.

This research has been supported by a MSM 0021620823 grant.



# Reliable Morphological Disambiguation of Czech: a Rule-Based Approach Is Necessary\*

VLADIMÍR PETKEVIČ

## 0 INTRODUCTION

Automatic morphological disambiguation of natural language texts (esp. those collected in language corpora) is one of the most difficult problems of contemporary natural language processing. This very difficult problem has not yet been solved in a satisfactory way for any natural language. It consists in:

- disambiguation of lemma(s), i.e. in the assignment of correct lemmas to a word-form occurrence (token) in a text being processed, the set of all possible lemmas pertaining to the given word-form being supplied by lemmatization as one of the modules of morphological analysis
- correct part-of-speech (POS) and morphological interpretation of a morphologically ambiguous word-form occurrence (token) in a text, i.e. in the assignment of proper POS and morphological tag(s) to the token, the set of all possible tags characterizing the given word-form (the tags are related to all the lemmas assigned to the word-form) having been supplied by morphological analysis.

As is well-known, this problem is a crucial bottleneck in corpora build-up and, accordingly, it should be paid due attention. Moreover, it is a very important step towards a successful syntactic analysis of any natural language in which the problem of POS and morphological ambiguity arises. There are three basic methodological approaches attempting to cope with the problem, namely:

- stochastic approaches
- rule-based approaches
- combined stochastic and rule-based approaches.

In this article I shall concentrate on the *rule-based disambiguation* of Czech as one of the most morphologically and syntactically intricate Slavic languages. I claim that if a method for the successful disambiguation of Czech exists, then also other less morphologically and syntactically complex languages (those belonging to the Slavic language family or other families) can be morphologically analyzed and disambiguated (I use the term *tagging* for POS and morphological analysis *and* disambiguation) with *incomparably greater success* than has been the case with the various stochastic methods applied so far. In the sequel, I shall try to provide evidence for this most unequivocally formulated statement; moreover, I even claim that no method other than the one presented can disambiguate texts in a sufficiently correct way.

---

\* The work described was funded by the Grant Agency of the Czech Republic (grant No. 405/03/0913).

## 1 WHY IS CORRECT POS AND MORPHOLOGICAL DISAMBIGUATION EXTREMELY DIFFICULT?

The answer to this question seems to be simple but it is not unhelpful to understand the principal reasons why disambiguation is so difficult and why it is such a real intellectual challenge that a computational linguist is confronted with:

- quantitative problem: natural languages comprise hundreds of thousands of word forms and many paradigms (understood here as groups of related word-forms and lemmas) which can be combined in a huge number of syntagmatic combinations
- a natural language system is a very complex system of rules and exceptions to these rules
- in POS and morphological disambiguation, not only sentence segmentation and tokenization, phonology and morphology are used but first of all syntax and also semantics. This means that the solution to a simpler problem (morphological disambiguation) requests more sophisticated means (e.g. semantic analysis). Thus, bootstrap methods<sup>1</sup> known from computer science cannot, unfortunately, be used without serious problems because:
  - (a) for solving a problem on a lower level (morphology) one needs information supplied from a higher level (syntax, semantics);
  - (b) in general, one cannot entirely rely on lower levels of description (this is related to point (a): e.g. morphological and POS disambiguation needs a perfect sentence segmentation and tokenization, and, unfortunately, vice versa)
- in natural language texts, unknown and foreign words (in view of the language under investigation) are often encountered and they must also be tagged
- various varieties and dialects of the given language appear in texts: for instance in Czech, standard Czech, colloquial Czech and dialects should be recognized and distinguished
- any living language undergoes constant changes, primarily in its vocabulary and collocations, less so in syntax and morphology.

I claim that only a very sophisticated and fine-grained linguistic analysis of the system of a particular language can cope with the given task. The indispensable necessity of such an analysis will be demonstrated below.

## 2 TAGGING CZECH

### 2.1 MAIN GENERAL FEATURES OF CZECH FROM THE PERSPECTIVE OF TAGGING

As stated above, only a deep linguistic analysis of the primarily syntactic structure of language (i.e. Czech in this study) resulting in a rule-based approach can perform tagging with satisfactory results (no less than at least 99% accuracy!). The tagging task is comprised of the two main subtasks:

- morphological and POS analysis which:
  - performs context-independent lemmatization, i.e. the assignment of all possible lemmas to the given word-form
  - assigns all context-independent morphological and POS interpretations (tags) to the given word-form according to a morphological tagset previously designed
- context-dependent POS, morphological and lemma disambiguation.

---

<sup>1</sup> Bootstrap is a term used in computer science: simpler building blocks on level  $x$  are construed as means to build up a more complex construction of a higher level  $x + 1$  which is subsequently used as a building block for a more complex construction on the level  $x + 2$  etc., i.e. a sequence of means is thus created under the condition that the means on each level are perfectly reliable.

The tagset reflecting the morphological richness of Czech is quite extensive: out of approx. 4400 theoretically extant tags, more than 2000 distinct tags are actually used. Therefore, morphological analysis is quite a laborious task but, unlike morphological and POS disambiguation, it is realizable in an almost error-free way (at least for known words). By contrast, the disambiguation which is the main topic of my study is very difficult. One has to take into account the well-known general properties of Czech that make the disambiguation task extraordinarily difficult (cf. also Oliva et al. 2000):

- free word order
- very rich inflection (esp. with nominal paradigms)
- high degree of case syncretism in nominal paradigms
- relative absence of syntactic fixed points in sentence structure
- high degree of accidental part-of-speech and morphological ambiguity of Czech word forms.

The following two examples reflecting the above-mentioned aspects show the richness of Czech morphology and the complexity of the disambiguation task:

#### EXAMPLE 1

(1) *Teprve tato řešení, jež mají vliv na výrobu, rozhodnou o opatření, které se bude muset přijmout.*

(E. lit.: Only these solutions which have an impact on the production will determine the measure which will have to be adopted.)

The following lemmas, POS and morphological interpretations are supplied by one of the existing morphological analyzers of Czech (cf. Hajič 2004), the correct disambiguation being underscored:

form: *Teprve*, lemma: teprve, tags: Adverb

form: *tato*, lemma: tento, tags: PronDem (nom. sg. fem., nom. pl. neut., acc. pl. neut.)

form: *řešení*, lemma: řešení, tags: Noun (nom. sg. neut., gen. sg. neut., dat. sg. neut., acc. sg. neut., voc. sg. neut., loc. sg. neut., nom. pl. neut., gen. pl. neut., acc. pl. neut., voc. pl. neut.); lemma: řešený, tags: Adj (nom. pl. mascanim., voc. pl. mascanim.)

form: „,“, lemma: „,“, tags: Punctuation

form: *jež*, lemma: jenž, tags: PronRel (nom. sg. fem., nom. pl. masculin., acc. pl. masculin., nom. pl. fem., acc. pl. fem., nom. sg. neut., acc. sg. neut., nom. pl. neut., acc. pl. neut.); lemma: *ježit*, tags: Verb (imper. sg.)

form: *mají*, lemma: mít, tags: Verb (pres. 3<sup>rd</sup> pers. pl.)

form: *vliv*, lemma: vliv, tags: Noun (nom. sg. masculin., acc. sg. masculin.)

form: *na*, lemma: na, tags: Prep (acc, loc)

form: *výrobu*, lemma: výroba, tags: Noun (acc. sg. fem.)

form: „,“, lemma: „,“, tags: Punctuation

form: *rozhodnou*, lemma: rozhodný, tags: Adj (acc. sg. fem., instr. sg. fem.); lemma: rozhodnout, tags: Verb (pres. 3<sup>rd</sup> pers. pl.)

form: *o*, lemma: o, tags: Prep (acc, loc)

form: *opatření*, lemma: opatření, tags: Noun (nom. sg. neut., gen. sg. neut., dat. sg. neut., acc. sg. neut., voc. sg. neut., loc. sg. neut., nom. pl. neut., gen. pl. neut., acc. pl. neut., voc. pl. neut.); lemma: opatřený, tags: Adj (nom. pl. mascanim., voc. pl. mascanim.)

form: „,“, lemma: „,“, tags: Punctuation

form: *které*, lemma: *který*, tags: *PronRel* (nom. pl. masculin., acc. pl. masculin., gen. sg. fem., dat. sg. fem., loc. sg. fem., nom. pl. fem., acc. pl. fem., nom. sg. neut., acc. sg. neut.); tags: *PronInterrog* (nom. pl. masculin., acc. pl. masculin., gen. sg. fem., dat. sg. fem., loc. sg. fem., nom. pl. fem., acc. pl. fem., nom. sg. neut., acc. sg. neut.);  
 form: *se*, lemma: *s*, tags: *Prep* (gen, instr); lemma: *se*, tags: *PronRefl*  
 form: *bude*, lemma: *být*, tags: *Verb* (fut. 3<sup>rd</sup> pers. sg.)  
 form: *mušet*, lemma: *mušet*, tags: *Verb* (inf.)  
 form: *přijmout*, lemma: *přijmout*, tags: *Verb* (inf.)

We immediately perceive the high degree of case syncretism with the nominal word forms *tato*, *řešení*, *jež*, *opatření*, *které*. As to syntactic fixed points, the prepositions *na* and *o* requiring a certain case (accusative/locative) and governing the prepositional phrase *na výrobu* and *o opatření*, respectively, can be considered fixed points as well as the pair (, and *které*) and (, and *jež*) introducing their respective relative clauses following the main one. Accidental ambiguity is represented by the word-forms *rozhodnou* and *se*.

#### EXAMPLE 2

(2) *Poté se*(Refl.Pron) *ředitel, který byl znám svou poddajností při vyjednávání se zahraničními partnery, opravdu snažil*(Verb-PastPart) *jednání rychle uzavřít.*

(E. lit.: Afterwards the director who was well-known for his submissiveness during negotiations with foreign partners really tried to close the negotiations very quickly.)

This sentence demonstrates the free word order in Czech: lexically and syntactically related elements can be separated by an arbitrary number of word-forms in Czech sentence as is shown by the reflexive only verb *snažit se* whose reflexive pronoun/particle *se* takes up the second syntactic position in the main clause, whereas *snažil* is separated from *se* by some elements of the main clause as well as by the entire embedded relative clause.

#### 2.2 BRIEF EVALUATION OF EXISTING STOCHASTIC DISAMBIGUATION OF CZECH

Czech texts, especially those contained in the *Czech National Corpus* (CNC, cf. Czech National Corpus 2000; Český národní korpus 2000), have so far been morphologically tagged almost exclusively by stochastic methods (cf. Hajič et al. 1997; Hajič et al. 1998, Hladká 2000, Hajič 2004). These methods yielded a success rate attaining a maximum of 94.5 % which is a relatively (with regard to the 97–98 % success rate achieved e.g. for English and French) poor result if we take into account the fact that unambiguous word-forms are also included in this success rate. In addition to the complexity of the syntactic structure of Czech mentioned above, the following specific factors are responsible for the success rate of stochastic methods being so low:

- inadequacy of stochastic methods applied on a free word order language because of sparse data
- very rich tagset (more than 4400 distinct tags used for Czech).

Our criticism of stochastic methods applied to tagging Czech texts can be summarized as follows (cf. also Oliva et al. 2000). Stochastic methods:

- use only positive information based on the training data rather than the negative information (Oliva 2001; Oliva et al. 2002)
- use only a limited context
- are totally dependent on very limited (sparse) training data and therefore they cannot adequately reflect the system of language as a whole, i.e. so-called smoothing is necessary

- are crucially dependent on the size of the tagset: the larger the tagset the more tag sequences exist as the result of morphological analysis and the sparseness of the training data is painfully felt
- make error identification impossible or at least very difficult (Oliva 2001; Oliva et al. 2002)
- commit naive errors primarily due to smoothing necessitated by the insufficiency of the training data
- they may “overdisambiguate”, i.e. they inadequately disambiguate morphologically inherently ambiguous sentences.

In order to avoid the shortcomings listed above, a group of computational linguists began to develop a system based on the language system of Czech, rather than on statistical chance (cf. Oliva et al. 2000).

## 2.3 PURELY RULE-BASED SYSTEM OF AUTOMATIC DISAMBIGUATION OF CZECH

### 2.3.1 MAIN CHARACTERISTICS

The low success-rate achieved by stochastic tagging for Czech led to a totally different approach to tagging, viz. the *purely rule-based approach* which should not have any of the above-mentioned negative properties of the stochastic methods. This approach is based on the manual development of negative and positive syntactic disambiguation rules (unlike e.g. Brill 1992) reflecting the syntactic system of Czech. The negative disambiguation rules remove all or some of the incorrect POS and morphological interpretation(s) (encoded in tags) of the given word-form, whereas the positive rules select the correct POS and morphological interpretation(s) of the given word-form in a sentence. The motivation for developing such rules is both theoretical and practical. From the theoretical point of view, the rules make it possible to obtain deep insights into the syntactic (and partly also semantic) structure of Czech and, in the final analysis, to develop a grammar of Czech based on corpus data. The practical objectives of the development of syntactic disambiguation rules can be summarized as follows:

- to perform much better morphological tagging of Czech language corpora
- to prepare a solid basis for a syntactic analysis of Czech, i.e. the rule-based system can be considered as a preprocessing stage (a kind of shallow analysis) for a full-fledged syntactic analysis. It is evident that, if the disambiguation system is almost error-free, it can considerably facilitate a subsequent syntactic analysis proper. In addition to the disambiguation itself, the rule-based disambiguation system can make it possible to identify especially:
  - the syntagms, i.e. the relation of the governor and its dependent node
  - nominal groups
  - prepositional groups
  - analytical verbo-nominal predicates
  - analytical verbal predicates
  - reflexive and other verbs/adjectives
  - agreement of various kinds
  - clause structure in compound sentences (at least in simpler cases)
  - valency relations
  - word order relations
  - collocations
- to prepare the ground for semantic analysis of the sentence, e.g. for the word sense disambiguation

- to make it possible to develop a grammar-checker for Czech.

All these objectives can be achieved – as our tests and experience show – because the rule-based disambiguation system, in contrast with the shortcomings of the stochastic approach, has the following properties:

- the system is based on the cooperation of disambiguation rules reflecting the system of language and a collocation component responsible for processing various deviations from the system and idiosyncrasies of language
- the rule system captures the system of the given language (de Saussure's langue), i.e. Czech, as reflected in parole; the method is linguistically based, i.e. it exploits specific features of the language system of Czech
- the rules (primarily (morpho)syntactic ones) are developed on the basis of linguistic intuition and analysis and verified on corpus data; there is no automatic (and, very often, erroneous) inferring of the grammar from a corpus
- the rules are based on unlimited context
- the rules use both negative and positive facts about language
- the disambiguation system uses a reduction method which consists in the following: the input to the system is the output of the morphological analysis where:

*recall* = 100 % (in a fault-free case, i.e. in a case where the set of lemmas and tags assigned to a given word-form contains the correct one)

*precision* = lowest possible (maximum number of *incorrect* lemmas and tags is assigned);

the method tries to retain the maximum *recall* (100 %) simultaneously maximizing *precision* by the following basic operations:

- the removal of incorrect morphological interpretations down to (in the optimum case) the only correct lemma(s) and tag(s) – this is a negative approach (primarily used by the rules)
- direct identification of the correct(s) tag(s) only – this is a positive approach (primarily used by the collocation component)
- the system needs no training data (but it needs relatively well-tagged corpora)
- the performance of the system does not deteriorate if the size of the tagset increases
- the system does not try to “overdisambiguate”, i.e. to disambiguate morphologically inherently ambiguous sentences. This means that each corpus position is assigned correct interpretations (possibly more than one, i.e. not necessarily the only one, cf. Oliva 2001b)
- the rules and the collocation component work together as follows: the collocation component comes first; then the rule-based system follows it and subsequently the collocation component is invoked again and the whole cycle may be repeated
- the rules are mutually independent and unordered, and they operate on continually more and more disambiguated data; each rule is applied until it cannot disambiguate any more, and after all the rules have thus been applied the whole bunch of rules (starting from the first one whichever that may be) is applied again till it is detected that in one cycle the data were not changed
- during rule application, no overt syntactic structures (such as trees) are built
- negative n-gram conception ( $n \geq 2$ ) is used by the rules, i.e. the system makes use of tuples of incorrect sequences of tags assigned to word-forms, i.e. these sequences violate the (mainly syntactic) system of Czech (examples are given below) – these negative n-grams can be automatically extracted from already tagged corpora and they can thus be used as an auxiliary means for rules' development



- the rules make it possible to immediately localize an error – it has nothing to do with the black box approach of stochastic methods
- the rules' validity can be measured in terms of decades at least because every language is very slow in changing its syntax
- for rules' development, all available sources of linguistic information should be used (phraseological dictionaries, valency dictionaries, dictionaries and classes of ambiguous forms, paradigmatic lists of lexemes sharing the same property/ies etc.)
- the rules are written in a special programming language LanGr (cf. Květoň 2003; Květoň in prep.) which is especially suited to the effective and clearly organized development of rules.

The whole rule-based strategy can be labeled as the *horror erroris* approach, i.e. the method primarily tries to keep recall close to 100 % (the method tries to avoid errors as much as possible) and to gradually increase precision (as already noted, initially, i.e. immediately after the morphological analysis before the disambiguation process starts, the precision is the lowest one). This strategy is adopted both by the rules and by the collocation component.

In the following section, the intrinsic structure of a rule will be described.

### 2.3.2 THE STRUCTURE OF A RULE

A disambiguation rule consists basically of four types of components:

- context
- disambiguation area
- report
- disambiguation action.

which are basically related as follows (cf. Květoň in prep.; Petkevič et al. 2002; Petkevič 2004):

$cont_1 \text{ disamb}_1 \text{ cont}_2 \text{ disamb}_2 \dots \text{cont}_n \text{ disamb}_n \text{ cont}_{n+1} \quad \text{report} \quad \text{action}$

where

$cont_i$  is the description of a *context*

$disamb_i$  is the description of a *disambiguation area* where the actions (see below) are performed (i.e. data are modified)

*report* is the report of a disambiguation action performed

*action* is a disambiguation action resulting in removing one or more incorrect tags.

The *context* is always unambiguously specified by means of the *IsSafe* quantifier – a word-form or a sequence of word-forms in question must have *only* the specified property, i.e. *all of its morphological interpretations* (tags) *must comply with the condition specified*. This means that *context* must always be unambiguously specified and it is *not changed* by the rule application.

The *disambiguation area* is subject to data change and it is specified by means of the *Possible* quantifier stating that *at least one of the morphological interpretations* (tags) *of the given word-form must comply with the condition specified*. The disambiguation area is typically ambiguous and there are in principle two basic actions (operations) that modify the data:

*DELETE some (not necessarily all) incorrect interpretation(s) from one or more corpus positions (i.e. word tokens equipped with lemmas and tags)*



*LEAVE ONLY* correct interpretation(s) in one or more corpus positions (i.e. word tokens equipped with lemmas and tags)

In addition to these basic functions, there also exist other functions in the system that, in fact, serve as macros for performing more *DELETE* and *LEAVE ONLY* operations at the same time. For instance, one of such key functions is:

*UNIFY [CONDITIONALLY] x y IN [gender,number,case]*

which has two operands, *x y* (the operands being two corpus positions, each with its own repertory of tags), and *leaves only* those respective values of the gender, number and case attribute in *x* and *y* which are in the intersection of the values of each of the respective attributes for *x* and *y*. The optional argument *CONDITIONALLY* makes unification conditional (i.e. the *UNIFY* operation is performed only if for each respective attribute the intersection is non-empty). It is clear that this function can be used mainly for the identification of agreement or saturation of valency requirements.

The *report* part contains a message describing the action performed.

#### EXAMPLE 3

The following example shows a very simple but extremely effective syntactic disambiguation rule:

##### Rule 1

```
/* No verbal form can immediately follow a part-of-speech unambiguous preposition */
rule PrepVerb1 {
safeprep = ITEM IsSafe Preposition;
/* this is a simple context which specifies one corpus position occupied by a part-of-speech
unambiguous (IsSafe) preposition, i.e. the word form safeprep has no other part-of-speech
interpretation */
possverb = ITEM Possible Verb;
/* the disambiguation area is identified with one corpus position specified as possverb, i.e. at
least one of the interpretations of the word-form possverb must be interpretable as a verbal
form */
REPORT("The verbal form possverb cannot immediately follow the unambiguous preposition
safeprep!");
/* this report describes the disambiguation action given below, referring to the actual word-
forms in the text being processed */
/* the following disambiguation actions are variants resulting in identical modification of the
current data – verbal interpretation (tag) in possverb is discarded */
DELETE Verb FROM possverb;
/* or */
LEAVE ONLY not Verb IN possverb;
}; // end of rule PrepVerb1
```

The rule can be successfully applied e.g. to the following sentence:

(3) *Jsem pro(Prep) rozhodnou(Adj | Verb) odpověď válečným štváčům.*  
(E. lit.: I am for a decisive action against the warmongers.)

Here the incorrect verbal reading of the word form *rozhodnou* (3rd person singular present tense of the verb *rozhodnout*, E. decide) is correctly removed by the rule, the adjectival reading being left intact.

The Possible and IsSafe quantifiers can change places and thus we obtain the dual Rule 1’

#### Rule 1’

```
/* No proposition can immediately precede a part-of-speech unambiguous verbal form */
rule PrepVerb2 {
  possprep = ITEM Possible Preposition;
  /* this is a disambiguation area identified with a Possible preposition, i.e. at least one of the
  interpretations of the word form possprep must be interpretable as a preposition */
  safeverb = ITEM IsSafe Verb;
  /* this is a simple context specifying one corpus position occupied by a part-of-speech
  unambiguous (IsSafe) verbal form, i.e. the word-form safeverb cannot have some other part-
  of-speech interpretation */
  REPORT("The preposition possprep cannot immediately precede the unambiguous verbal form
  safeverb!");
  /* this report describes the disambiguation action given below, referring to the actual word-
  forms in the text being processed */
  /* the following disambiguation actions are again variants resulting in identical modification
  of the current data – prepositional interpretation (tag) in possprep is discarded */
  DELETE Preposition FROM possprep;
  /* or */
  LEAVE ONLY not Preposition IN possprep;
}; // end of rule PrepVerb2
```

The rule can be successfully applied e.g. to the following sentence:

(4) To místo(Noun | ~~Prep~~ | Conj) bylo(Verb) *velmi pěkné*.

(E. lit.: The *place* was very nice.)

Here the incorrect prepositional reading of the word form *místo* (E. instead of) is correctly discarded by the rule, the nominal and conjunctive readings being left intact by the rule.

The configuration formed by the ordered pair (Preposition, Verb) is syntactically incorrect in Czech (and in many other languages of the world) – it is a classical example of the negative bigram concept. As we have seen, two different disambiguation rules can result from this fact: one of the elements of the pair is fixed (as *context*), the other (as *disambiguation area*) is operated on and duly changed. Possible subsequent rule applications will then operate on the data modified by Rule 1 and Rule 1’, respectively. More on that in the next section.

#### 2.3.3 NEGATIVE APPROACH TO THE LANGUAGE SYSTEM

As I have just demonstrated, it may be highly appropriate and productive to look at the language system from the negative point of view. Thus, our point of departure is what the system of language *does not admit*, i.e. it is appropriate to make use of negative constraints in language on all of its levels (primarily the syntactic one). So the traditional positive view of the language system should be reversed, although it is clear that the negative view is only derived from the positive one as its negation. In the solution of the task in hand, the negative approach to the language system results in the search for negative n-grams that form the basis for the development of disambiguation rules. It is precisely here where the rule-based approach I am describing differs from the stochastic one (cf. Oliva et al. 2002) which can use positive evidence only based on the “positive” training data (as no “negative corpora” have been developed so far).

Negative n-grams, i.e. (properties of) word-forms in complementary distribution from the syntactic viewpoint, can be automatically extracted from existing disambiguated corpora (although

these corpora contain errors). The impact on disambiguation of these n-grams can be further extended (cf. Oliva 2001; Oliva et al. 2002; Oliva 2005). If we have an adjacent negative bigram  $(x1, x2)$ , this implies that if element  $x1$  is *immediately* followed by the element  $x2$  the structure is ungrammatical. It may be the case that the presence of another element  $x3$  in between or outside the bigram (in the word order sense) does not change the original ungrammaticality of the bigram  $(x1, x2)$ , i.e. the trigram  $(x1, x3, x2)$  is also ungrammatical and so the original ungrammaticality of  $(x1, x2)$  remains preserved. It is a very important task of the linguists to specify (a) how many elements can stand in between the original bigram  $(x1, x2)$ , so that the ungrammaticality of the resulting structure is preserved, and (b) what properties these elements must have.

It is clear that the concept of negative n-gram is not limited to *non-adjacent* ungrammatical structures as invariants. This is reflected in the rule structure depicted in Sect 2.3.2 above.

Let us now show instances of negative bigrams and trigrams in Czech, as extracted from the Czech National Corpus. For reasons of simplicity, all these instances represent adjacent n-grams.

#### EXAMPLE 4. EXAMPLES OF NEGATIVE BIGRAMS

- Preposition having no locative valency immediately followed by a word in the locative case
- Vocalized preposition immediately followed by a word beginning with a vowel
- Clitic at the beginning of sentence
- Non-prepositional word form immediately followed by a personal/relative pronoun beginning with *-n* (*něho, němu, ...*)
- Word form *velmi* immediately followed by an adjective in the comparative or superlative degree of comparison (there are no exceptions even in collocations!)
- Present form of a verb different from *být* (E. be) immediately followed by a past participle form

#### EXAMPLE 5. EXAMPLES OF NEGATIVE TRIGRAMS

- Three adjacent prepositions
- A triple formed by: transitive verb, adjective in nominative, noun in accusative (*VT, A1, N4*)
- A triple formed by: adjective in locative, nominal word-form in case different from locative and instrumental, adjective in locative:  
(*A6, nominalform[67], A6*) (cf. Hajič 2004)
- Noun in dative, noun in instrumental, noun in dative (*N3, N7, N3*)
- Noun in accusative, adjective in dative, noun in accusative (*N4, A3, N4*)
- Noun in nominative, adjective in genitive, noun in nominative (*N1, A2, N1*)

As I have already indicated, negative n-grams can be automatically transformed into rules. An example of the rule deduced from a negative trigram is shown below.

#### EXAMPLE 6

The following example shows a very simple but extremely effective syntactic disambiguation rule:

##### Rule 2

*rule ThreePrepositions {*

*/\* The rule is based on the negative trigram:*

Preposition Preposition Preposition.

It is a negative trigram because such a sequence can never occur in any Czech sentence. The corresponding rule has three variants:

PossSafeSafe, SafePossSafe, SafeSafePoss, in each of which two prepositions (Safe) are fixed as context, the remaining one (Poss) being discarded. The rule seems to be valid for the vast majority of the languages of the world which have prepositions

\*/

*RuleVariant PossSafeSafe {*

*/\* possible preposition immediately followed by two safe prepositions \*/*

*possib = ITEM Possible Preposition;*

*/\* possible preposition \*/*

*ITEM IsSafe Preposition;*

*/\* safe preposition \*/*

*ITEM IsSafe Preposition;*

*/\* safe preposition \*/*

*DELETE Preposition FROM possib;*

*/\* discard a prepositional reading in the possible preposition \*/*

*}; // end of the PossSafeSafe variant*

*/\*\*\*\*\*/*

*RuleVariant SafePossSafe {*

*/\* possible preposition between two safe prepositions \*/*

*ITEM IsSafe Preposition;*

*/\* safe preposition \*/*

*possib = ITEM Possible Preposition;*

*/\* possible preposition \*/*

*ITEM IsSafe Preposition;*

*/\* safe preposition \*/*

*DELETE Preposition FROM possib;*

*/\* discard a prepositional reading in the possible preposition \*/*

*}; // end of the SafePossSafe variant*

*/\*\*\*\*\*/*

*RuleVariant SafeSafePoss {*

*/\* possible preposition immediately preceded by two safe prepositions \*/*

*ITEM IsSafe Preposition;*

*/\* safe preposition \*/*

*ITEM IsSafe Preposition;*

*/\* safe preposition \*/*

*possib = ITEM Possible Preposition;*

*/\* possible preposition \*/*

*DELETE Preposition FROM possib;*

*/\* discard a prepositional reading in the possible preposition \*/*

*}; // end of the SafeSafePoss*

*}; // end of the ThreePrepositions rule*

The rule may be successfully applied to the following sentence:

EXAMPLE 7

(5) *Sedl si na místo* (Subst | **Prep** | Conj) *s ustaraným výrazem v tváři.*

(E. lit. He sat down on the place with a concerned expression in face.)

Here the prepositional reading is discarded because it is placed in between two other safe prepositions *na* and *s*.

The negative approach to disambiguation based on the positive knowledge of the language system in question may lead to considerations that can sound very surprising for a traditional syntactician; these considerations can reveal a very different thinking about language. Certain facts about the language system can be evidenced from very different angles, and one can then use the simplest one (*simplest* from the viewpoint of its encoding in a formal language, or of its very simple identifiability etc.). This can be demonstrated by many types of examples; I present two simple examples below.

#### EXAMPLE 8

(6) *Tetě jsme se rozhodli dát dárek.*

(E. lit. To the *aunt* we decided to give a present.)

Our task is to properly identify the case of the noun *tetě* (morphological analysis assigns the lemma *teta* (E. aunt) to this word-form and two possible cases: *dative* and *locative*). At first glance, it is evident that *tetě* is in the dative case because it is the indirect object of the verb *dát* (E. give). However, it is relatively difficult to state with 100% certainty that *tetě* really is an indirect dative object because this would mean parsing the whole sentence. It is incomparably simpler to state that *tetě* is not in the locative case (i.e. we apply a negative approach from the viewpoint of the dative identification) because there is no locative-requiring preposition in front of the word *tetě* in sentence (6). We end up with the dative case because we have rejected the locative one. From the two solutions to our problem we have chosen the simpler one.

In the following example, the range of possibilities of the part-of-speech disambiguation is even broader.

#### EXAMPLE 9

(7) *Včera jsme se vážně snažili nakoupit nějaké jídlo.*

(E. lit. Yesterday we seriously tried to buy some food.)

Here the word form *se* is (at least) two-way POS ambiguous: it is either a reflexive pronoun/particle *se*, or the vocalized preposition *s* (the reflexive interpretation of *se* can be further distinguished but I shall not discuss its possible refinements here). In the present example, *se* is the reflexive pronoun for the following reasons:

- *se* as the preposition never vocalizes in front of the word starting with *v* that is followed by a vowel
- no preposition can stand in front of a verb (even with the adverb *vážně* standing in between)
- there is no potential genitive or instrumental (the only cases the preposition *s/se* theoretically requires) behind *se* in the given sentence
- the verb *snažit se* is reflexive-only, i.e. its form *snažili* obligatorily requires the reflexive pronoun/particle *se* and therefore *se* is not a vocalized preposition.

As we see, there are four reasons for identifying *se* as the reflexive pronoun/particle which are mainly based on the rejection of the prepositional interpretation of *se*. Thus, we can select the most simple one here but, in fact, the rule-based disambiguation system contains separate rules for each of the four phenomena mentioned, i.e. each of these rules independently discards the prepositional interpretation of *se*, whichever comes first.

The third example shows different reasons for the other interpretation of *se*, i.e. the prepositional one.

EXAMPLE 10

(8) *Včera jsme po procházce městem byli se starým strýcem a se zámožnou tetou v kině.*

(E. lit. Yesterday we were after a walk through the city with the old uncle and the well-to-do aunt in the cinema.)

The analysis of the sentence leads us to the following reasons for identifying both occurrences of *se* as prepositions:

- each reflexive pronoun/particle *se* in every correct Czech sentence must be associated with an (obligatorily or optionally) reflexive verb or an (obligatorily or optionally) reflexive adjective or an (optionally) reflexive postverbal noun as its free morpheme; moreover, *se* can also express the passive, which means that the presence of a verb in the sentence is obligatory. In the whole sentence, there is no appropriate candidate for *se* to be associated with the verb forms since *byli* and *jsem* of the lemma *být* are neither reflexive, nor can they form the passive voice. Thus, *se* cannot be a reflexive pronoun/particle and so it must be the vocalized form of the preposition *s*;
- *se* as a reflexive pronoun/particle must stand on the second (i.e. Wackernagel's) syntactic position in a clause, which is, however, not the case in sentence (8);
- *se* can be a preposition because it stands immediately in front of a possible instrumental case (weak positive reason).

Thus, the first two reasons are sufficient for us to identify both occurrences of *se* as non-reflexives. The third statement allows for the prepositional interpretation of *se*; for instance, if there were no word either in the genitive case, or in the instrumental case following at least one occurrence of *se*, the sentence would be syntactically wrong.

#### 2.3.4 RULE-BASED SYSTEM AND NOTES ON PART-OF-SPEECH AND MORPHOLOGICAL AMBIGUITY OF CZECH

One of the main problems which the rule-based disambiguation discussed above must cope with is the fact that the morphological (as well as the syntactic) system of Czech is extremely complex. By morphological complexity, I mean a very high ambiguity rate in Czech. The majority of word-forms in Czech are part-of-speech and morphologically ambiguous (homonymous) with respect to the standard tagset for Czech which accounts for all morphological categories in a relatively detailed way (cf. Hajič 2004). From the morphological point of view, Czech is probably the most complex language (at least in the family of Slavic languages). As far as disambiguation is concerned, there are two major kinds of ambiguity in Czech (cf. Oliva et al., 2000):

- systemic ambiguity
- accidental ambiguity.

Systemic ambiguity concerns primarily case syncretism in declension paradigms. In this regard, Czech is the most complex of the Slavic languages. This syncretism presents the most complicated problem for any automatic disambiguation of Czech. Because of its complexity, I will mention only the most complicated types of syncretism the disambiguation task has to cope with:

- syncretism of the nominative/accusative of all masculine inanimate and neuter nouns and adjectives and also of some feminine nominal paradigms (this is a typical syncretism in those Indo-European languages which have a declension system)
- syncretism of the nominative/accusative of all nouns in plural except for masculine animate nouns
- syncretism of all the cases (except for the instrumental case) of the neuter paradigm *stavení* in singular
- soft adjectives, i.e. those adjectives in the positive degree of comparison whose lemma ends in *-í*, and all adjectives in the comparative and superlative degree of comparison.

The first two types of syncretism make it difficult to automatically distinguish subject and object (i.e. they do not differ in form). Due to the free word order in Czech, syntax is of no avail here; only very fine-grained semantic considerations can identify the subject and object here. The following example is typical:

(9a) *Soud*(nom | acc) *vynesl rozsudek*(nom | acc).

(9b) *Rozsudek*(nom | acc) *vynesl soud*(nom | acc).

(E. lit. The court delivered the judgement./The judgement was delivered by the court.)

In both sentences either *soud* (E. court) is the subject in nominative and *rozsudek* (E. judgement) is the object in accusative, or vice versa: *rozsudek* is the subject in nominative, and *soud* is the object in accusative. From the given structure we can only infer that both nouns definitely differ in case. The semantic properties of all three components of the structure, i.e. subject, object and finite verb, should be taken into consideration but no study has been devoted to the solution of this problem to date.

Stochastic methods commit many errors in the identification of cases; this is one of their weakest points. This concerns not only the most difficult problems listed above but far simpler problems, such as the identification of the locative case, cases of elements in prepositional groups, let alone nominative and accusative in those subject-predicate-object structures in which subject or object are easy to identify.

Accidental ambiguity concerns non-systemic ambiguity which manifests itself especially if a word-form can be interpreted as belonging to different parts of speech or a word-form is in the intersection of paradigms of two different lexemes. The first type of accidental ambiguity is demonstrated by Example 11:

#### EXAMPLE 11

(10) *Dělníci šli*(Verb | Noun) *podle*(Prep | Adv) *řeky*.

(E. lit. The workers *went by* the river.)

where:

*šli* is:

- verb: *past participle pl. masc. anim.* of the lemma *jít* (E. go)
- noun: *dat. sg. fem., acc. sg. fem., loc sg. fem.* of the lemma *šle* (E. brace)

*podle* is:

- preposition (taking genitive): of the lemma *podle* (E. by, along)
- adverb: of the lemma *podle* (E. meanly, wickedly)

The second type of accidental ambiguity is demonstrated by Example 12:



#### EXAMPLE 12

(11) *Česká republika je dobrá v tancích.*

(E. lit. The Czech Republic is good at tanks/dances.)

As the English gloss suggests, the word form *tancích* (locative plural masc. inanimate) can either belong to the paradigm of the lemma *tank* or of the lemma *tanec* (E. dance). Every disambiguation system of Czech should be able to identify the correct lemma but the correct identification of the proper lemma seems to be based on the semantics of the context only.

Stochastic disambiguation of Czech (cf. Hajič et al. 1997; Hladká 2000; Hajič 2004) performs relatively well in handling the first type of accidental ambiguity problem (with several notable exceptions), i.e. it can identify with relatively satisfactory accuracy the part-of-speech of the given word-form. The second type of ambiguity that concerns the proper identification of the lemma can, as a matter of fact, hardly be solved by stochastic methods for Czech, because they do not account for semantics.

#### 2.3.5 THE RULE-BASED SYSTEM AND SYNTACTIC COMPLEXITY OF CZECH

As our experience based on the analysis of Czech corpora and on linguistic introspection shows, the syntactic complexity of Czech requires a very sophisticated disambiguation system so that the success rate could be much higher than the actual maximum (94,5%) achieved by the too approximative stochastic methods. The rule system expressed in the formal programming language *LanGr* (Květoň 2003; Květoň in prep.). I have presented has to reflect and capture the following general facts about the structure of a Czech sentence expressed in Statement 1 and Statement 2.

#### STATEMENT 1

The majority of syntactic relations in a Czech sentence have a local character but a very non-local context is required for identifying both local and non-local syntactic relations in it. No disambiguation methods which do not respect these characteristics of a Czech sentence can ever be successful.

For instance, no methods which only use a narrow context (i.e. a fixed window) can be appropriate, due to the free word order in Czech. The following examples demonstrate this clearly:

#### EXAMPLE 13

(12) *Na(acc | loc) její dva až tři roky trvající absenci se podepsalo zranění.*

(E. lit. On her two and a half year's lasting absence signed the injury.

Transl.: Her absence lasting two years and a half was caused by the injury.)

Here the preposition *na* requires a nominal group *její ... trvající absenci* in the *locative case* which is caused by the valency of the verb *podepsat se* requiring the preposition *na* taking the locative case. The adverbial temporal accusative phrase *dva až tři roky* modifying the adjective *trvající* is embedded in the locative nominal group *její ... trvající absenci*. If a narrow context only were taken into account, i.e. the context which did not take the distant verb into consideration, then, due to the adjacent position of the embedded temporal accusative phrase, the entire prepositional group would be incorrectly classified as the accusative one.

Due to the free word order in Czech, there are types of structures in which two syntactically related elements can be almost arbitrarily distant from each other. A typical instantiation of this phenomenon is presented in the following example.



#### EXAMPLE 14

(13) *Tyto problémy se skutečně novými metodami poté, co jsme přijali příslušné usnesení, které bylo z řady hledisek přijatelné, konečně se značným úsilím dařilo řešit.*

(E. lit. These problems ... with really new methods, after we adopted the appropriate resolution, which was from many viewpoints acceptable, eventually with a lot of effort succeeded to solve. Transl.: After we adopted the appropriate resolution which was acceptable from many viewpoints, we eventually managed to solve the problems after much labour/effort.)

In this sentence, the underlined words constitute one lexeme, i.e. the reflexive only verb *dařit se* (E. manage, succeed). As you can see, they are separated by 22 positions (including commas). If a limited narrow window only were used, the underlined *se* would be tagged as a vocalized preposition because all the conditions in favour of this in the near right context are fulfilled (*se* is followed by the nominal group *skutečně novými metodami*, which is in the instrumental case and the first word, i.e. *skutečně*, begins with the sibilant in front of which the preposition *s* must vocalize. Therefore, the context of the rules should extend to the whole sentence at least.

The second general statement about the system of Czech concerns many deviations from the system; for disambiguation, this is especially crucial in morphology and syntax.

#### STATEMENT 2

Czech has a very “dismembered” (full of exceptions, idiosyncrasies) character in both its paradigmatic and syntagmatic subsystems. Again, no disambiguation methods which do not respect these characteristics of a Czech sentence can ever be successful.

This fact complicates both morphological analysis and subsequent disambiguation. Morphological analysis must account for all the complexity of the nominal declension system and the verbal conjugation system. This leads to many specific paradigms in both these systems. In disambiguation, syntactic idiosyncrasies in the system are handled by the collocation component which is, as mentioned above, invoked before and after the rule component.

The above statements lead me to the following conclusions:

- successful part-of-speech and morphological disambiguation can be achieved only if both statements are respected; therefore I come down in favour of a rule-based system
- the rules need to be complicated because they must reflect the inherent complexity of the Czech sentence
- hundreds of rules need to be developed to cover at least a considerable part of the Czech syntax, the rules ranging from very general ones to very specific ones (e.g. those concerning particular word forms, especially the most frequent ones, and those taking up the pivotal positions in a sentence)
- the classes of ambiguity must be construed in as much detail as possible and the rules should use them.

#### 2.3.6 EXISTING EXPERIENCE WITH THE RULE-BASED DISAMBIGUATION SYSTEM

During the development of the rule-based tagger, the need for which was formulated in Oliva et al. 2000, the following experience was collected:

- the conclusions mentioned above have been confirmed
- a very fine-grained analysis of language based on deep linguistic intuition and corpus data is unavoidable
- if possible, the rules must be as general as possible (otherwise even thousands of rules would have to be written) but at the same time they have to be, in the main, very sophisticated and *error-free*

- notwithstanding the maximum generality of the rules, the system of Czech requires hundreds of rules to be formulated
- a very powerful and sophisticated programming language for writing the rules had to be developed (cf. Květoň 2003; Květoň in prep.) so that it could cope with the intricacies of Czech
- the rule-based method is very sensitive to any errors in input, especially to the following ones:
  - unknown words (morphological analysis does not know them)
  - error in morphological analysis
  - missing comma or other punctuation
  - wrong sentence segmentation
  - idiosyncrasies of the system.

Here *sensitive* means that the system can successfully detect errors so that it can also serve as the basis for a grammar-checker.

The following problems prove the most difficult for the rule-based disambiguation:

- systemic case syncretism in Czech declension paradigms (cf. Sect. 2.3.4)
- actual ellipses
- the nominative of nomination (*ve městě Praha*, E. *in the city of Prague*)
- accidental ambiguity of adverbial and particle word-forms belonging to also to a different part of speech
- the presence of various varieties of the same language in the same text
- the administration and application of a database of collocations which is continuously being enhanced.

## BIBLIOGRAPHY

- BRANTS T. (2000): TnT – A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000. Seattle.
- BRILL E. (1992): A Simple Rule-Based Part-of-Speech Tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing. Trento.
- CHANOD J. P., TAPANAINEN P. (1995): Tagging French – comparing a statistical and a constraint-based method. In: Proceedings of EACL-95. ACL, Dublin, 149–157.
- Czech National Corpus (2000). Faculty of Arts, Charles University, Prague. <http://ucnk.ff.cuni.cz>
- Český národní korpus (2000). Úvod a příručka uživatele. Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy, Praha. <http://ucnk.ff.cuni.cz>
- HAJIČ J. (2004): Disambiguation of Rich Inflection (Computational Morphology of Czech). Univerzita Karlova v Praze, nakladatelství Karolinum, Praha.
- 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. Washington D.C.
- HAJIČ J., HLADKÁ B. (1998): Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: Proceedings from COLING-ACL'98. Montreal, 483–490.
- HAJIČ J., KRBEČ P., KVĚTOŇ P., OLIVA K., PETKEVIČ V. (2001): Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001). CNRS – Institut de Recherche en Informatique de Toulouse and Université des Sciences Sociales, Toulouse, 260–267.
- HLADKÁ B. (2000): Czech Language Tagging. PhD thesis. Faculty of Mathematics and Physics, Charles University, Prague.

- KARLSSON F., VOUTILAINEN A., HEIKKILÄ J., ANTILLA A. (eds.) (1995): *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, New York.
- KVĚTOŇ P. (2003): *Language for Grammatical Rules*. Technical Report TR-2003-17. Matematicko-fyzikální fakulta UK, Prague.
- KVĚTOŇ P. (in prep.): *Rule-Based Morphological Disambiguation (Towards a Combination of Linguistic and Stochastic Methods)*.
- NEGRA Corpus: [www.coli.uni-sb.de/sfb378/negra-corpus](http://www.coli.uni-sb.de/sfb378/negra-corpus).
- OLIVA K., HNÁTKOVÁ M., PETKEVIČ V., KVĚTOŇ P. (2000): The Linguistic Basis of a Rule-Based Tagger of Czech. In: Sojka P., Kopeček I., Pala K. (eds.): *Proceedings of the Conference "Text, Speech and Dialogue 2000"*, Lecture Notes in Artificial Intelligence 1902. Springer-Verlag, Berlin – Heidelberg, 3–8.
- OLIVA K. (2001a): The Possibilities of Automatic Detection/Correction of Errors in Tagged Corpora: A Pilot Study on a German Corpus. In: Matoušek V., Mautner P., Mouček R., Taušer K. (eds.): *Proceedings of the Conference "Text, Speech and Dialogue 2001"*. Lecture Notes in Artificial Intelligence 2166. Springer-Verlag, Berlin – Heidelberg, 39–46.
- OLIVA K. (2001b): On Retaining Ambiguity in Disambiguated Corpora. In: *Traitement Automatique des Langues vol. 42 No. 2*, Hermes Science Publications, Paris.
- OLIVA K., KVĚTOŇ P. (2002): (Semi-)Automatic Detection of Errors in PoS-Tagged Corpora. In: Shu Chuan Tseng (ed.): *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002, Taipei)*. Morgan Kaufmann Publishers, San Francisco, 509–515.
- OLIVA K. (2005): Úvahy nad teoretickými základy lingvisticky adekvátní disambiguace jazykových korpusů. In: Blatná R., V. Petkevič (eds.), *Jazyky a jazykověda. Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy, Praha*, 229–245.
- PETKEVIČ V. (2001): Grammatical Agreement and Automatic Morphological Disambiguation of Inflectional Languages. In: Matoušek V., Mautner P., Mouček R., Taušer K. (eds.): *Proceedings of the Conference "Text, Speech and Dialogue 2001"*. Lecture Notes in Artificial Intelligence 2166. Springer-Verlag, Berlin – Heidelberg, 47–53.
- PETKEVIČ V. (2001): Automatic Detection of Subject and Verbal Predicate in the Czech Translation of G. Orwell's '1984'. In: Zybatov G., Junghanns U., Mehlhorn G., Szucsich L. (eds.): *Current Issues in Formal Slavic Linguistics. Proceedings of the Third European Conference on Formal Description of Slavic Languages (FDSL 1999)*. Peter Lang, Frankfurt am Main, 506–518.
- PETKEVIČ V., HNÁTKOVÁ M. (2002): Automatická morfologická disambiguace předložkových skupin v Českém národním korpusu. In: Hladká Z., Karlík P. (eds.): *ČEŠTINA – univerzália a specifika 4*. Nakladatelství Lidové noviny, Praha, 243–252.
- PETKEVIČ V. (2003): Subject-Predicate Agreement and Automatic Morphological Disambiguation of the Czech National Corpus. In: Kosta P., Błaszczak J., Frasek J., Geist L., Žygis M. (eds.): *Investigations into Formal Slavic Linguistics. Contributions of the Fourth European Conference on Formal Description of Slavic Languages (FDSL 2001)*. Peter Lang, Frankfurt am Main, 315–328.
- PETKEVIČ V. (2004): Využití pravidel pro negaci v automatickém značkování českých korpusů. In: Hladká Z., Karlík P. (eds.): *ČEŠTINA – univerzália a specifika 5. Sborník konference v Brně, listopad 2003*. Nakladatelství Lidové noviny, Praha, 143–150.
- Prague Dependency Treebank (2002): <http://ufal.ms.mff.cuni.cz/>
- SAMUELSSON Ch., VOUTILAINEN A. (1997): Comparing a Linguistic and a Stochastic Tagger. In: *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, Madrid.
- TAPANAINEN P., VOUTILAINEN A. (1994): *Tagging accurately – Don't guess if you know*. Technical Report, Xerox Corp.
- VOUTILAINEN A. (1995): Morphological disambiguation. In: Karlsson F. et al. (eds.): *Constraint Grammar*. Berlin – New York, 165–285.

## ABSTRACT

Článek se zabývá automatickou slovnědruhovou a morfologickou disambiguací češtiny jako zřejmě nejsložitějšího slovanského jazyka jak z morfologického, tak syntaktického hlediska. Hlavním tenorem článku je ukázat, že zejména pro jakoukoli spolehlivou disambiguaci českých textů shromážděných zejména v elektronických textových korpusech češtiny a pro účely automatické syntaktické analýzy českých textů je nutné zvolit řešení lingvistické. Tímto řešením je konkrétně systém syntaktických a jiných pravidel, nikoli náhodný přístup statistický, s nímž se článek mj. kriticky vypořádává.

Po stručném nastínění problémů a kritickém pohledu na stochastické metody disambiguace, jimiž je dosud značkován synchronní korpus SYN2000 v rámci projektu Český národní korpus, je charakterizován systém založený na pravidlech, na němž autor spolu s dalšími lingvisty a programátory sám pracuje. Především jsou zdůrazněny hlavní ideové pilíře systému, dále doložené ilustrativními příklady. Těmito pilíři jsou zvláště tyto vlastnosti systému pravidel:

- systém je založen na spolupráci sady *disambiguačních pravidel* odrážejících jazykový systém češtiny a frazémového komponentu, který zpracovává různé odchylky od systému a výjimky
- pravidla zachycují systém češtiny tak, jak se projevuje v *parole*, reprezentovaném korpusovými texty
- pravidla se vyvíjejí na základě lingvistické intuice a autorů a jsou prověřována na datech korpusu
- pravidla využívají neomezeného kontextu
- pravidla využívají negativních i pozitivních jazykových faktů, přičemž základ tvoří fakta negativní v podobě takzvaných negativních *n*-gramů
- disambiguační metoda je metodou redukční, tj. snaží se udržovat maximální pokrytí (*recall*) a postupně zvyšovat přesnost (*precision*)
- pravidla buď odstraňují nesprávné morfologické interpretace (značky/tagy) u jednotlivých slovních tvarů v textech, nebo určují jediné správné interpretace
- pravidla jsou vzájemně nezávislá, a proto neuspořádaná a spolupracují s frazémovým komponentem
- pravidla jsou psána ve speciálním programovacím jazyce, který umožňuje zachytit velice složité syntaktické a slovosledné vztahy v jazyce a který umožňuje snadno lokalizovat chyby v pravidlech.

Každé pravidlo systému se skládá ze čtyř částí: opěrného kontextu, disambiguačního místa, kde dochází k disambiguaci (tj. ke změně dat), zprávy o uskutečněné akci a konečně akce samé, kterou je buď odstranění nesprávných značek na disambiguačním místě, nebo ponechání značek správných.

Po uvedení názorných příkladů se rozebírají vlastnosti a výhodnost pojmu negativní *n*-gram a možnost odvozovat z negativního *n*-gramu příslušné disambiguační pravidlo, což je doloženo příkladem. Dále se probírají výhodné vlastnosti negativního pohledu na jazyk, zejména to, že jistá syntaktická skutečnost platí z několika důvodů (nejen z jednoho), ale že je správné zachytit všechny takové důvody náležitými formálními pravidly.

Text poté pokračuje úvahami o specifických komplikacích v morfologickém a syntaktickém plánu češtiny z hlediska pravidly řízené disambiguace: jde tu zejména o bohatý a nepříjemný pádový synkretismus v deklinačním systému češtiny. Rovněž jsou představeny základní typy homografie v češtině a syntaktická složitost češtiny je vyjádřena dvěma klíčovými tvrzeními o celkovém rázu této složitosti. Konkrétně se praví, že *většina syntaktických vztahů v češtině má lokální ráz, ale k určení lokálních i nelokálních vztahů ve větě je zapotřebí značně nelokální kontext*. A dále: *Čeština má velmi "rozdrobený" ráz ve svém paradigmatickém i syntagmatickém podsystému*. Obecný závěrem těchto tvrzení je toto: Pokud má být disambiguace úspěšná, musí obsah těchto tvrzení bezpodmínečně respektovat.

Na závěr autor uvádí dosavadní zkušenosti s vývojem pravidly řízeného systému a celý text zakončuje stručným výčtem největších problémů popsané disambiguační metody zjištěných na základě dosavadních zkušeností s vývojem pravidel.

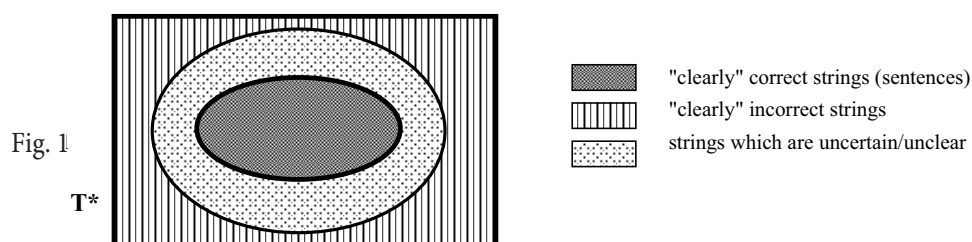
# Discovering and Employing Ungrammaticality

KAREL OLIVA

## INTRODUCTION

Apart from deciding on the membership of a particular string  $\sigma$  in a particular language  $L$ , a formal grammar is usually assigned an additional task: to assign each string from the language  $L$  some (syntactic) structure. The idea behind this is that the property of having a structure differentiates the strings  $\sigma \in L$  from all “other” strings  $\omega \notin L$ , i.e. having a structure differentiates sentences from “non-sentences”. Due to this, the task of identifying the appurtenance of a string to a language (the set membership) and the task of assigning the string its structure are often viewed as in effect identical. In other words, the current approach to syntactic description supposes that any string  $\omega \in T^*$  which cannot be assigned a structure by the respective grammar is to be considered (formally) ungrammatical. Closely linked to this also is the assumption that the borderline between grammatical and ungrammatical strings is sharp and clear-cut.

Even elementary language practice (e.g., serving as a native speaker – informant for fellow linguists, or teaching one’s mother tongue) shows that this assumption does not hold. The realistic picture is much more like the one in Fig. 1: there are strings which are considered clearly correct (“grammatical”) by the native speakers, there are other ones that are indubitably incorrect (out of the language, “informally ungrammatical”, unacceptable for native speakers), and there is a non-negligible set of strings whose status with regard to correctness (acceptability, grammaticality) is not really clear and/or where the opinions of native speakers differ (some possibly tending more in this, others more in the other direction, etc.).



Assuming the better empirical adequacy of the picture in Fig. 1, the objective of this paper will be to propose that a syntactic description of (any natural) language  $L$  should consist of:

- a formal grammar  $G$  defining the set  $L(G)$  of indubitably grammatical strings ( $L(G) \subseteq L$ ). Typically, the individual components of  $G$  (rules, principles, constraints,...) are based on a structure assigned to a string, either directly (mentioning e.g., the constituent structure) or indirectly, operating with other syntactically assigned features (such as

subject, direct object, etc.). Since the description of the “clearly correct” strings via such a grammar is fairly standard, it will not be treated any further here.

- a “formal ungrammar”  $U$  defining the set  $L(U)$  of indubitably ungrammatical strings. Typically, any individual component (“unrule”) of  $U$  would be based on lexical and morphological characteristics only, i.e. it would have no direct recourse to the structure of a string or to other syntactic characteristics (such as being a subject etc.).

Unlike the standard approach, such a description also allows for a non-empty set of strings belonging neither to clearly grammatical nor to clearly ungrammatical strings – more formally, such a description allows for a non-empty set  $T^* - (L(G) \cup L(U))$ . Besides this, the explicit knowledge of the set  $L(U)$  of ungrammatical strings allows for the straightforward development of important applications (cf. Sect. 3).

### 1 THE UNRULES OF THE UNGRAMMAR

The above abstract ideas call for methods for discovering and describing the “unrules” of the “ungrammar”. In the search for such methods, the following two points can be postulated as starters:

1. grammaticality/ungrammaticality is defined for whole sentences (i.e. not for subparts of sentences only, at least not in the general case)
2. ungrammaticality occurs (only) as a result of the violation of some linguistic phenomenon or phenomena within the sentence.

Since any «clear» error consists of the violation of a language phenomenon, it seems reasonable that the search for incorrect configurations be preceded by an overview and classification of the phenomena suited to the current purpose.

From the viewpoint of the manner of their manifestation in the surface string, (syntactic) phenomena can be divided into three classes:

- **selection phenomena:** in a rather broad understanding, selection (as a generalized notion of sub-categorisation) is the requirement for a certain element (a syntactic category, sometimes even a single word)  $E_1$  to occur in a sentence if another element  $E_2$  (or: set of elements  $\{E_2, E_3, \dots, E_n\}$ ) is present, i.e. if  $E_2$  (or:  $\{E_2, E_3, \dots, E_n\}$ ) occur(s) in a string but  $E_1$  does not, the respective instance of selection phenomena is violated and the string is to be considered ungrammatical.

**Example:** in English, if a non-imperative finite verb form occurs in a sentence, then a word functioning as its subject must also occur in the sentence (cf. the contrast in grammaticality between *She is at home.* vs. *\*Is at home.*). ♦

- **(word) order phenomena:** word order rules are rules which define the mutual ordering of two (or more) elements  $E_1, E_2, \dots$  occurring within a particular string; if this ordering is not maintained, then the respective word order phenomenon is violated and the string is considered to be ungrammatical.

**Example:** in an English *do*-interrogative sentence consisting of a finite form of the auxiliary verb *do*, of a subject position filled by a noun or a personal pronoun in nominative, of a base form of a main verb different from *be* and *have*, and of the final question mark, the order must necessarily follow the pattern just used for the listing of the elements, or, in an echo question, it must follow the pattern of a declarative sentence. If this order is not maintained, the string is ungrammatical (cf. *Did she come? She did come?* vs. *\*Did come she?*, etc.). ♦

- **agreement phenomena:** broadly understood, an agreement phenomenon requires that if two (or more) elements  $E_1, E_2, \dots$  co-occur in a sentence, then some of their



morphological characteristics have to be in a certain systematic relation (most often, identity); if this relation does not hold, the respective instance of the agreement is violated and the string is ungrammatical. (The difference to selection phenomena consists thus of the fact that the two (or more) elements E1, E2, ... need not co-occur at all – that is, the agreement is violated if they co-occur but do not agree, but it is not violated if only one of the pair (of the set) occurs – which is the difference to a violation of selection.)

**Example:** the string *\*She does it himself* breaks the agreement relation in gender between the anaphora and its antecedent (while the sentences *She does it herself* and *She does it* are both correct – “note here the difference to selection”). ♦

This overview of classes of phenomena suggests that each string violating a certain phenomenon can be viewed as an extension of some **minimal violating string**, i.e. as an extension of a string which contains only the material necessary for the violation. For example, the ungrammatical string *The old woman saw himself in the mirror yesterday*, if considered a case of violation of the anaphora-agreement relation, can be viewed as an extension of the minimal string *The woman saw himself*, and in fact as an extension of the string *Woman himself* (since for the anaphora-agreement violation, the fact that some other phenomena are also violated in the string does not play any role).

This means that a minimal violating string can be discovered in each ungrammatical string, and hence each “unrule” of the “formal ungrammar” can be constructed in two steps:

- first, by defining an (abstract) minimal violating string, based on a violation of an individual phenomenon (or, as the case might be, based on a combination of violations of a “small number” of phenomena)
- second, by defining how the (abstract) minimal violating string can be extended into a full-fledged (abstract) violating string (or to more such strings, if there are more possibilities of the extension), i.e. by defining the material (as to quality and positioning) which can be added to the minimal string without making the resulting string grammatical (not even contingently).

The approach to discovering/describing ungrammatical strings will be illustrated by the following example where the sign ‘{’ will mark sentence beginning (an abstract position in front of the first word), and ‘}’ will mark sentence end (i.e. an abstract position “after the full stop”).

**Example:** As already reasoned above, the abstract minimal violating string of the string *The old woman saw himself in the mirror yesterday* is the following configuration (in the usual regular expression notation, using feature structures for the individual elements of the regular expression, ‘∨’ for disjunction, the sign ‘⊕’ for concatenation, and brackets ‘(’ and ‘)’ in the usual way for marking off precedence/grouping).

$$(1) \quad \{ \oplus \left( \left[ \begin{array}{l} \text{cat : n} \\ \text{gender : fem} \end{array} \right] \vee \left[ \begin{array}{l} \text{cat : pron} \\ \text{pron\_type : pers} \\ \text{gender : fem} \end{array} \right] \right) \oplus \text{himself} \oplus \}$$

This configuration states that a string consisting of two elements (the sentential boundaries do not count), a feminine noun or a feminine personal pronoun followed by the word *himself*, can never be a correct sentence of English (cf., e.g., the impossibility of the dialogue *Who turned lo into a cow? \*Hera himself*).

Further, such a violating (abstract) string can be generalized into an incorrect configuration of unlimited length using the following linguistic facts about the anaphoric pronoun *himself* in English:

- a bound anaphor must co-occur with a noun or nominal phrase displaying the same gender and number as the pronoun (with the binder of the anaphor); usually, this binder precedes the pronoun within the sentence (and in this case it is a true anaphor) or, rarely, it can follow the anaphor (a case of cataphoric relation: *Himself, he bought a book.*).
- occasionally, also an overtly unbound anaphor can occur; apart from imperative sentences (*Kill yourself !*), the anaphor must then closely follow a *to*-infinitive (*The intention was only to kill himself.*) or a gerund (*Killing himself was the only intention.*).

Taken together, these points mean that the only way to give the configuration from string (1) at least a chance to be grammatical is to extend it with an item which

- either, is in masculine gender and singular number
- or, is an imperative or an infinitive or a gerund and stands to the left of the word *himself*.

This further suggests that – in order to keep the string ungrammatical also after the extension – no masculine gender and singular number item must occur within the (extended) string, as well as no infinitive or gerund appearing to the left of the word *himself*.

This can be captured in a (semi-)formal way (employing the Kleene-star ‘\*’ for any number of repeated occurrences, and ‘¬’ for negation) as follows.

In the first step, the requirement of no singular masculine in the whole sentence is to be added (2), in the second step, the prohibition on the occurrence of an imperative or an infinitive (represented by the infinitival particle *to*) or a gerund to the left of the word *himself* will be expressed as in (3). This is then the final form of the description of an abstract violating string. Any string matching this description is guaranteed to be ungrammatical in English.

$$\begin{aligned}
 (2) \{ & \left( \neg \left[ \begin{array}{l} \text{number : sg} \\ \text{gender : masc} \end{array} \right] \right)^* \oplus \left( \left[ \begin{array}{l} \text{cat : n} \\ \text{gender : fem} \end{array} \right] \vee \left[ \begin{array}{l} \text{cat : pron} \\ \text{pron\_type : pers} \\ \text{gender : fem} \end{array} \right] \right) \\
 & \oplus \left( \neg \left[ \begin{array}{l} \text{number : sg} \\ \text{gender : masc} \end{array} \right] \right)^* \oplus \text{himself} \oplus \left( \neg \left[ \begin{array}{l} \text{number : sg} \\ \text{gender : masc} \end{array} \right] \right)^* \} \\
 (3) \{ & \left( \neg \left( \left[ \begin{array}{l} \text{number : sg} \\ \text{gender : masc} \end{array} \right] \vee \left[ \text{v\_form : (imp} \vee \text{ger)} \right] \vee \left[ \begin{array}{l} \text{cat : part} \\ \text{form : to} \end{array} \right] \right) \right)^* \oplus \left( \left[ \begin{array}{l} \text{cat : n} \\ \text{gender : fem} \end{array} \right] \vee \left[ \begin{array}{l} \text{cat : pron} \\ \text{pron\_type : pers} \\ \text{gender : fem} \end{array} \right] \right) \\
 & \oplus \left( \neg \left( \left[ \begin{array}{l} \text{number : sg} \\ \text{gender : masc} \end{array} \right] \vee \left[ \text{v\_form : (imp} \vee \text{ger)} \right] \vee \left[ \begin{array}{l} \text{cat : part} \\ \text{form : to} \end{array} \right] \right) \right)^* \oplus \text{himself} \oplus \left( \neg \left[ \begin{array}{l} \text{number : sg} \\ \text{gender : masc} \end{array} \right] \right)^* \}
 \end{aligned}$$

## 2 UNGRAMMAR AND THE THEORY OF GRAMMATICALITY

An important case – mainly for the theory of grammaticality – of a minimal violating string is three finite verbs following each other closely, i.e. the configuration *VFin + VFin + VFin*. Such a configuration appears, e.g., in the sentence *The mouse the cat the dog chased caught survived* which is a typical case – frequently discussed in its time – of a multiple centre self-embedding construction. The important point concerning this construction is that it became the issue of discussions since:

- on the one hand, this construction is – (almost) necessarily – licensed by any “reasonable” grammar of English, due to the necessity of allowing in this grammar for the possibility of (recursive) embedding (incl. centre self-embedding) of relative clauses,



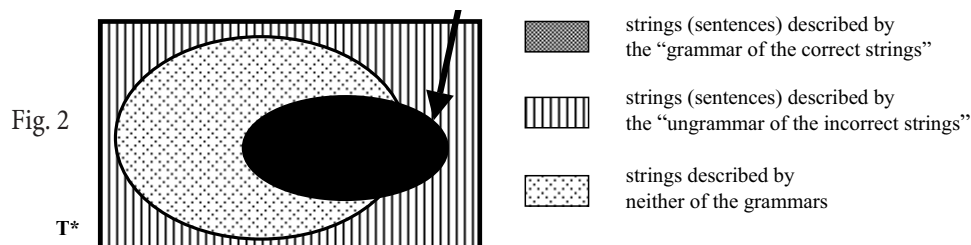
- on the other hand, such sentences are unanimously considered unacceptable by native speakers of English (with the contingent exception of theoretical linguists).

The antagonism between the two points is traditionally “explained” by a tension between the langue (grammar, grammatical competence) and the parole (language performance) of the speakers; that is, by postulating that the speakers possess some internal system of the language but that they use the language in a way which deviates from this system. Such an assumption is generally a good explanation for such (unintentional) violations of langue (i.e. of grammaticality) in speech as, e.g., slips of the tongue, hesitations and/or repetitions, etc., but it can hardly be sensibly employed where there are no extra-linguistic factors and, above all, where the sentences in question correspond to the langue (to the grammatical description). This demonstrates that what is really at stake here is the correctness in the general understanding of the langue (and not a problem of a particular grammar of a particular language).

The difference in methods of ruling sentences with multiple centre self-embedding out of the language moves us on to the fact that the standard view of the langue – and hence that of a grammar – and the view advocated in this paper differ considerably:

- the standard approach to the langue, which allows for the specification of the set of correct strings only (via the grammar), has no means available for ruling out constructions with multiple centre self-embedding constructions (short of ruling out recursion of the description of relative clauses, which would indeed solve the problem; however, it would also have serious negative consequences elsewhere),
- the approach proposed, by allowing for explicit and, most importantly, independent specifications of the sets of correct and of incorrect strings as two autonomous parts of the langue, allows for ruling out constructions involving multiple centre self-embedded relative clauses (at least in certain cases); this is achieved without consequences to any other part of the grammar and the language described, simply by stating that strings where three (or more) finite verbs follow each other immediately belong to the area of “clearly incorrect” strings.

By solving the problem of unacceptability of the strings involving three (and more) finite verbs following each other via the formal ungrammar, the approach proposed enforces a refinement of perspective of the general description of grammaticality and ungrammaticality. In particular, from now on, Fig. 1 above has to be understood as depicting the situation in the language (understood as a set of strings) only, i.e. without any recourse to the means of its description (i.e. without any recourse to a grammar and, in particular, to the coverage of a grammar). The coverage of the two grammar modules introduced above (the “grammar of the correct strings” and the “ungrammar of the incorrect strings”), i.e. the string-sets described by the components of the grammar describing the “clearly correct” and the “clearly incorrect” strings, should rather be described as in Fig. 2.



The crucial point is the part of this diagram determined by the arrow (where the dense dots and vertical bars overlap). This area of the diagram is the one representing strings which are described by both components of the grammar, i.e. strings which are covered both by the description (grammar) of the correct strings and by the description (ungrammar) of the incorrect strings. At first glance, this might seem to be a contradiction (seemingly, some strings are simultaneously considered correct and incorrect), but this is not the case, since the true situation described in this diagram is the partitioning of the set of strings  $T^*$  by two **independent** set description systems, each of which describes a subset of  $T^*$ . Viewed from this perspective, it should not be surprising that some strings are described by both of the systems (while others are described by neither of them). The fundamental issue here is the relation of the two description systems (the grammar and the ungrammar) to the pre-theoretical understanding of the notion of grammaticality as the acceptability of a string for a native speaker of a language. Traditionally, all those strings were considered grammatical which were described by the grammar of the correct strings. In the light of the current discussion, and particularly from the evidence provided by the multiple centre self-embedding relative constructions, this definition of grammaticality should be weakened by adding the proviso that strings which are covered by the description of incorrect strings (by the ungrammar) should not be considered grammatical (not even in cases where they are simultaneously covered by the grammar of the correct strings). This changes the perspective (compared to the standard one), by giving the ungrammar the “veto right” over the grammaticality of a string, but obviously corresponds more closely to the language reality than the usual approach.

Viewed from the perspective of a grammatical description considered as a model of a linguistic competence, the fore-going discussion can be summed up as follows:

- (formally) grammatical strings are strings described by the grammar but not by the ungrammar,
- (formally) ungrammatical strings are strings described by the ungrammar,
- strings whose grammaticality is (formally) undefined are strings which are described neither by the grammar nor by the ungrammar.

### 3 APPLICATIONS

In the previous sections, rather theoretical issues concerning the general view of grammaticality and a means of description of grammatical/ungrammatical strings were dealt with. However, the task of finding the set of strictly ungrammatical strings also has a practical importance, since for certain applications it is crucial to know that a particular configuration of words (or of abstractions over strings of words, e.g., configurations of part-of-speech information) is guaranteed to be incorrect.

The most prominent (or at least: the most obvious) among such tasks is **robust parsing**, including its applications such as **grammar-checking** etc. (for recent relevant references cf., e.g., Schneider and McCoy 1998, Holan et al. 2003, Bender et al. 2004), where ungrammatical input is usually dealt with by means of rules describing ungrammatical constructions (e.g., a rule  $S \rightarrow NP[sg] VP [pl]$ , describing a case of subject-verb agreement violation). The important difference between these “mal-rules” (as they are often dubbed) and the approach presented in this paper is that a successful application of a “mal-rule” during the parsing process does not automatically imply the ungrammaticality of the string (i.e. as long as there exists an alternative parse which does not contain any application of a “mal-rule”). On the other hand, if an unrule matches a sentence (as described above), then this sentence

can safely be considered ungrammatical. In this respect, it can be expected that the application of ungrammars would result in grammar-checkers truly capable of reliably recognizing that a string is ungrammatical, which in turn would result in systems with considerably more user-friendly performance than our present ones (based mostly on simple pattern-matching techniques, and hence producing a lot of false alarms over correct strings on the one hand while leaving unflagged many strings whose ungrammaticality is obvious to a human, but which cannot be detected as incorrect, since their inner structure is either too complex or does not correspond to any of the patterns for any other reason).

Another practical task where the knowledge of the ungrammar of a particular language may turn into the central expertise needed is *part-of-speech tagging*, i.e. assigning morphological information (such as part-of-speech, case, number, tense, ...) to words in running texts. The main problem for (automatic) part-of-speech tagging is morphological ambiguity, i.e. the fact that words might have different morphological meanings (e.g., the English word-form *can* is either a noun (“a food container”) or a modal verb (“to be able to”); a more typical – and much more frequent – case of ambiguity in English is the noun/verb ambiguity in such systematic cases as *weight*, *jump*, *call*, ...). The knowledge of ungrammatical configurations can be employed in the build-up of a part-of-speech tagger based on the idea of (step-wise) elimination of those individual readings which are ungrammatical (i.e. impossible) in the context of a given sentence. In particular, each extended violating string with  $n$  constituting members (i.e. a configuration which came into being by extending a minimal violating string of length  $n$ ) can be turned into a set of disambiguation rules by stipulating, for each resulting rule differently,  $(n - 1)$  constituting members of the extended violating string as unambiguous and issuing a deletion statement for the  $n$ -th original element in a string which matches the constituting elements as well as the extension elements in between them. Thus, each extended violating string arising from a minimal violating string of length  $n$  yields  $n$  disambiguation rules.

**Example:** The minimal violating string ARTICLE + VERB, after being extended into the configuration (in the usual Kleene-star notation) ARTICLE + ADVERB\* + VERB, yields the following two rules:

**Rule 1:** *find\_a\_string* consisting of (from left to right):

- a word which is an unambiguous ARTICLE (i.e. bears no other tag or tags than ARTICLE)
- any number of words which bear the tag ADVERB (but no other tags)
- a word bearing the tag VERB

*delete\_the\_tag* VERB *from* the last word of the string

**Rule 2:** *find\_a\_string* consisting of (from left to right):

- a word bearing the tag ARTICLE
- any number of words which bear the tag ADVERB (but no other tags)
- a word which is an unambiguous VERB (i.e. it bears only a single tag verb or it bears more than one tag, but all these tags are VERB)

*delete\_the\_tag* ARTICLE *from* the first word of the string

The (linguistic) validity of these rules is based on the fact that any string matching the pattern part of the rule on each position would be ungrammatical (in English), and hence that the reading to be deleted can be removed without any harm to any of the grammatical readings of the input string.

It is important to realize that the proposed approach to the “discovery” of disambiguation rules yields the expected results – i. e. rules corresponding to the Constraint Grammar rules given in standard literature (e.g., it brings in the rule for English saying that if an unambiguous ARTICLE is followed by a word having a potential VERB reading, then this VERB reading is to be discarded, cf. Karlsson et al. 1995, p. 11, and compare this to the example above). The most important innovative feature (with regard to the usual ad hoc approach to writing these rules) is thus the *systematic linguistic method* of discovering the violating strings, supporting the development of all possible disambiguation rules, i.e. of truly powerful Constraint Grammars. It is also worth mentioning that the method as such is language-independent – it can be used for the development of Constraint Grammars for most different languages (even though the set of rules developed will, of course, be language-specific and will depend on the syntactic regularities of the language in question).

Yet another task – but closely related to the above two – which can profit greatly from the ability to recognize ungrammaticality is parsing. If the part-of-speech (i.e. morphological) information of the words on the input can be determined prior to syntactic parsing, or at least the morphological ambiguity of the input be reduced, then, obviously, the parsing process can be considerably more effective. Also, if the input can reliably be claimed to be ungrammatical, its parsing by standard methods need not even be started, and hence a lot of time-consuming processing can be avoided (either by giving up the parsing task completely, or by employing some techniques for dealing with ungrammatical input straightforwardly from the start of the processing).

#### ACKNOWLEDGEMENT

The work described in this paper has been supported in part by Grant No. 16614 of the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)* of the Republic of Austria. The *Austrian Research Institute for Artificial Intelligence (OeFAI)* is supported by the *Austrian Federal Ministry for Education, Science and Culture* and by the *Austrian Federal Ministry for Transport, Innovation and Technology*.

#### BIBLIOGRAPHY

- BENDER E. M., D. FLICKINGER, S. OEPEN, A. WALSH AND T. BALDWIN. Arboretum: Using a Precision Grammar for Grammar Checking in CALL. In: Proceedings of the InSTIL/ICALL Symposium 2004: NLP And Speech Technologies in Advanced Language Learning Systems, Venice.
- KARLSSON F., A. VOUTILAINEN, J. HEIKKILÄ AND A. ANTILLA (eds.) (1995) *Constraint Grammar – A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin & New York.
- HOLAN T., V. Kuboň, M. Plátek and K. Oliva (2003). A Theoretical Basis of an Architecture of a Shell of a Reasonably Robust Syntactic Analyser. In: Proceedings of the Conference on Text, Speech and Dialogue TSD 2003, Lecture Notes in Artificial Intelligence vol. 2807, Springer, Berlin.
- SCHNEIDER D. and K. F. McCoy (1998). Recognizing Syntactic Errors in the Writing of Second Language Learners. In: Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and the Seventeenth International Conference on Computational Linguistics (COLING-ACL), Montreal.

## ABSTRACT

A natural language is usually modelled as a subset of the set  $T^*$  of strings (over some set  $T$  of terminals) generated by some grammar  $G$ . Thus,  $T^*$  is divided into two disjoint classes: into grammatical and ungrammatical strings (any string not generated by  $G$  is considered ungrammatical). This approach brings with it the following problems:

- on the theoretical side, it is impossible to rule out clearly unacceptable yet “theoretically grammatical” strings (e.g., strings with multiple centre self-embeddings, cf. *The cheese the lady the mouse the cat the dog chased caught frightened bought cost £ 10*),
- on the practical side, it impedes the systematic build-up of such practical applications of computational linguistics as, e.g., grammar-checkers.

In an attempt to lay a theoretical foundation facilitating the solution of these problems, the paper first proposes a tripartition of the string-set  $T^*$  into:

- clearly grammatical strings,
- clearly ungrammatical strings,
- strings with unclear (“on the verge”-) grammaticality status

and, based on this, it concentrates on

- techniques for the systematic discovery and description of clearly ungrammatical strings,
- the impact of the approach on the theory of grammaticality,
- an overview of simple ideas about applications of the above in building grammar-checkers and rules-based part-of-speech taggers.



# Complex Corpus Annotation: The Prague Dependency Treebank

JAN HAJIČ

## 1 INTRODUCTION

Let us now reveal the truth: the idea for the Prague Dependency Treebank did not really come from Prague. First, the original inspiration came from Philadelphia (where else?): in the early 90s, the availability of the Penn Treebank (Marcus et al., 1993) was an object of fascination (to us at least). Then, at the European ACL Conference in Dublin in 1995, a small group of us “Praguians” met to discuss the feasibility of such a treebank (based on the dependency framework, of course – what else!). We had no money and, therefore, no people to carry it out, but we decided to push the idea through the national Czech Grant Agency (even though it was clear we could not really call it a “treebank”<sup>1</sup>, since that was quite a “dirty” word, then), proposing at the same time another large grant for a Czech National Corpus together with several other colleagues from the country and a project called the Laboratory for Language Data (with the idea that it would be in this Laboratory where the annotation would in fact take place). Fortunately enough, we were awarded grants for all three of these projects<sup>2</sup> and, in the fall of 1996, the project was able to go ahead at full speed.

In present-day computational linguistics (CL), the availability of annotated data (spoken utterances, written texts) is becoming a more and more important factor in any new development. Apart from speech recognition, where statistical methods are almost exclusively *the* solution and where the data is a *conditio sine qua non*, textual data are being used for the training phase of various statistical methods solving many other problems in the field of CL. While there are many methods which use texts in their plain (or raw) form (for unsupervised training), (much) more accurate results may be obtained if annotated corpora are available. It is believed that syntax (and, therefore, syntactic annotation) helps for subsequent processing in the direction of “language understanding” (or “comprehension”).

With the increasing complexity of such tasks, data annotation in itself is a complex task. While tagged corpora (pioneered by Henry Kučera in the 60’s) are now available for English and other languages, syntactically annotated corpora are rare. We decided to develop a similarly sized corpus of Czech with a very “deep” and rich annotation scheme.

---

<sup>1</sup> At that time we called it “validation of a theory”, without giving any figures regarding the number of words or sentences for which such “validation” would be performed.

<sup>2</sup> The project was started with support from the grant GAČR No. 405/96/0198 (“Formal specification of language structures”), and the annotation effort has been made possible by grant GAČR No. 405/96/K214 and by the project of the Ministry of Education of the Czech Republic No. VS96151. Later, the work continued under the project called Center for Computational Linguistics (2000-2004), MSMT CR Project LN00A063. The development of some software tools used in this project has been supported by grant GAČR No. 405/95/0190 and by the individual author’s grant OSF RSS/HESP 1996/195.

The textual data used for the task consists of general newspaper articles (40%; including but not limited to politics, sports, culture, hobbies, etc.), economic news and analyses (20%), popular science magazines (20%), and information technology texts (20%), all selected from the early collection of the Czech National Corpus.

## 2 THE PRAGUE DEPENDENCY TREEBANK STRUCTURE

The Prague Dependency Treebank (PDT) has a three-level structure (with tokenized text being taken as the input to the whole system). Full *morphological* annotation has been performed on the lowest (first) *level*. The middle level deals with syntactic annotation using dependency syntax; it is called the *analytical level*. The highest level of annotation is the *tectogrammatical level*, or the level of linguistic meaning. We annotate the same text on all three levels, but the amount of annotated material decreases with the complexity of the levels<sup>3</sup>.

## 3 THE MORPHOLOGICAL LEVEL

On the morphological level, a tag and a lemma is assigned to each word form as identified in the input text. The annotation contains no (syntactic) structure; no attempt is even made to put together analytical verb forms, for example.

### 3.1.1 THE CZECH TAG SYSTEM

Czech is an inflectionally rich language. The full tag set currently contains 4712 tags (including morphological variants, which are distinguished). We are using a positional tag system, a full description of which can be found in (Hajič, 2004).

We use 13 grammatical categories in the tag. For each category, one symbol is used at a fixed position in the tag string.

Cat.	Cat. Name	Description	Example values
1	POS	Part of Speech	A – adjective, R – preposition
2	SUBPOS	Detailed part of speech	s – passive participle, V – vocalized prep., Q – rel. pronoun
3	GENDER	Gender (grammatical, agreement)	I – masc. inanimate, N – neuter
4	NUMBER	Number (grammatical)	S – sing., D – dual
5	CASE	Case (or required case, for prep.)	1 – Nom., 3 – Dat., 7 – Instrumental
6	POSSGENDER	Possessive gender (owner's gender)	F – fem, M – masc. anim.
7	POSSNUMBER	Possessive number (owner's number)	S – singular, P – plural
8	PERSON	Person (verbs, pronouns)	1, 2, 3
9	TENSE	Tense (for participles, some exceptions)	R – past, F – future, P – present
10	GRADE	Degree of comparison (adjectives, adv.)	1 – positive, 3 – superlative
11	NEGATION	Negation prefix present	N – negated
12	VOICE	Voice (verbs)	A – active, P – passive
13	RESERVE1	Unused	
14	RESERVE2	Unused	
15	VAR	Variant, style, register, abbreviation, ...	1 – variant, 6 – colloquial, 8 – abbr.

<sup>3</sup> For various reasons, mainly technical: it has been experimentally proved (Zeman, 1998) that serially applied machine learning and statistical methods perform better if every step is trained on the true automatic output of the previous step rather than the manual one. In order to achieve this, there must be separate (additional) training data available for the preceding step, resulting in the greatest quantity of data being necessary for the beginning (the first step) of the analysis, namely, morphology, and the least for the last, the tectogrammatical analysis.



A brief example<sup>4</sup> now presents a simple sentence as a sequence of annotated words:

Form (Czech)	(Lit.)	Tag
,	,	<b>Z</b> :-----
že	<i>that</i>	<b>J</b> ,-----
litera	<i>the-letter</i>	<b>NNFS1</b> ----- <b>A</b> ----
výše	<i>above</i>	<b>Dg</b> ----- <b>2A</b> --- <b>1</b>
uvedené	<i>of-mentioned</i>	<b>AAFS2</b> ----- <b>1A</b> ----
mezinárodní	<i>of-international</i>	<b>AAFS2</b> ----- <b>1A</b> ----
smlouvy	<i>of-agreement</i>	<b>NNFS2</b> ----- <b>A</b> ----
mezi	<i>between</i>	<b>RR</b> -- <b>7</b> -----
ČR	<i>Czech Rep.</i>	<b>NNFXX</b> ----- <b>A</b> --- <b>8</b>
a	<i>and</i>	<b>J</b> ^-----
SR	<i>Slovakia</i>	<b>NNFXX</b> ----- <b>A</b> --- <b>8</b>
bude	<i>will</i>	<b>VB-S</b> --- <b>3F-AA</b> ---
mít	<i>have</i>	<b>Vf</b> ----- <b>A</b> ----
co	<i>pretty</i>	<b>TT</b> -----
nevidět	<i>soon</i>	<b>Vf</b> ----- <b>N</b> ----

Special symbols are used for combinations of values that are not easily distinguished, or the processing of which was simply left for the future. In most cases, we use the symbol 'X' for 'any value' in the particular grammatical category.

The lemma represents a unique identification of the word in the morphological dictionary. Usually, the customary dictionary base form (headword) is used as the identification string, extended (if necessary) by a dash and a number distinguishing it from its homographs. We use the following convention: all forms of a lemma must have the same part of speech, and for nouns, they also have to have the same gender. (This is, obviously, in accordance with the conventions of the morphological dictionary we use – see below in 3.1.2 Morphological Analysis).

### 3.1.2 MORPHOLOGICAL ANALYSIS

Morphological analysis is a process of which the input is a word form as found in the text, and the output is a set of possible lemmas which represent the form in the dictionary, with each lemma accompanied by a set of possible tags (as defined in the previous section). For example, for the word form *ženu* the morphological analysis returns the following results:

Lemma	tag(s)
žena ( <i>woman</i> )	<b>NNFS4</b> ----- <b>A</b> ----
hnát ( <i>to rush</i> )	<b>VB-S</b> --- <b>1P-AA</b> ---

This example exhibits an ambiguity at the lemma level, but no ambiguity within the lemmas. On the other hand, the word form *učení* displays both types of ambiguity:

<sup>4</sup> Example from the weekly journal Českomoravský profit, 10/1994.

Lemma	tag(s)
učení ( <i>theory</i> )	NNNS1-----A-----, NNNS2-----A-----, NNNS3-----A-----, NNNS4-----A-----, NNNS5-----A-----, NNNS6-----A-----, NNNP1-----A-----, NNNP2-----A-----, NNNP4-----A-----, NNNP5-----A-----
učený ( <i>educated</i> )	AAMP1-----1A-----, AAMP5-----1A-----

There could be as many as five different lemmas for a given word form and as many as 27 different tags for one lemma.

Morphological analysis currently covers about a million Czech lemmas (including derivations), and is based on about 520,000 stems. It can recognize about 25 million word forms and their tags.

### 3.1.3 THE PROCESS OF MANUAL MORPHOLOGICAL ANNOTATION

Morphological analysis is the first step towards the first level of annotation (morphological tagging) in the Prague Dependency Treebank. It can proceed fully automatically and very quickly (about 20000 word forms per second on today's average machine). We have developed a special software tool (called *sgd* on a Unix platform, and *DA* under Windows) which enables easy manual disambiguation of the morphological output. It also helps the annotators to edit the output of the morphology, thus facilitating the identification of possible problems and unknown words in the morphology itself.

The morphological annotation has been performed on every sentence in the PDT twice, with a third person resolving the differences between the two annotators. Inter-annotator agreement has been around 97% (measured as a percentage of input tokens receiving the same tag by both annotators). After the adjudication process, errors still remain, though as we are currently preparing version 2.0 of the PDT, we are better able to identify those errors (based on the upper levels of annotation) and we are correcting them.

A total of 1,800,000 words (tokens) is now available with manually annotated lemmas and tags.

### 3.2 THE ANALYTICAL LEVEL

The analytical (surface-syntactic) level of annotation is a newly designed level to more easily use (and compare) the results achieved in English parsing to Czech, and to have a preliminary analysis of a sentence structure before proceeding to the most detailed level, the tectogrammatical one. We have chosen the dependency structure to represent the syntactic relations within the sentence. Thus, the basic principles can be formulated as follows:

- The structure of the sentence is an oriented, acyclic graph with one entry (root) node; the nodes of the tree are annotated by complex symbols (attribute-value pairs);
- The number of nodes of the graph is equal to the number of words in the sentence plus one for the extra root node;
- The annotation result is only
  - 1. the *structure* of the tree,
  - 2. the *analytical function* of every node.

An analytical function determines the relation between the (dependent) node and its governing node (which is the node one level up the tree). All the other node attributes (see the table below)

are used as guidance for the annotators, or they are used as input or intermediate data for various automatic tools which play a part in the annotation process, but are not considered to be the result of analytical annotation. In particular, the tags and lemmas are taken from the morphologically annotated data, and they are merged into the resulting data structure.

The first 10 node attributes are summarized in the following table (there are 8 more “technical” attributes used for macro programming as intermediate data holders etc.):

Attribute name	Brief description
lemma	lemma (see sect. 3.1, The Morphological Level)
tag	morphological categories, or tag (see sect. 3.1, The Morphological Level)
form	word-form, after minor changes in some cases (error correction)
afun	the analytical function, or the type of dependency relation (towards the governing node)
origf	original word-form as found in the text
origap	formatting (preceding the original word-form)
gap1, gap2, gap3	formatting info preceding form, parts 1,2,3
ord	sequence no. of the word-form in a sentence

The annotation rules are described in the manual (Bémová et al., 1997), the final version of which is available together with the annotated data (and much more) on the Prague Dependency Treebank v1.0 CD (Hajič et al., 2001).

These rules follow, where possible, the traditional grammar books, but are both extended (where no guidance has been found in the books) and modified (where the current grammars are inconsistent). They are intentionally as independent of any formal theory as possible (even though the decision to use the traditional – at least in Prague – dependency representations is certainly not quite theory-independent – but in fact, this decision made our lives easier because of several phenomena occurring inherently in Czech (non-projective constructions, see e.g. Hajičová et al., 2004), which would otherwise result in the well-known “crossing brackets” problem.

In the following table, all the possible values of the analytical function attribute (*afun*) are described briefly. The existence of a “suffixed” version (*\_Co* for coordination, *\_Ap* for apposition, *\_Pa* for parenthetical expressions) is marked by an x.

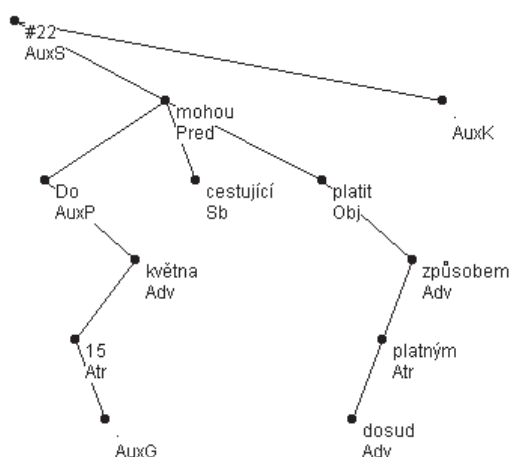
afun	_Co	_Ap	_Pa	Description
Pred	x	x	x	Predicate if it depends on the tree root (#)
Sb	x	x	x	Subject
Obj	x	x	x	Object
Adv	x	x	x	Adverbial (without a detailed type distinction)
Atv	x	x	x	Complement; technically depends on its non-verbal governor
AtvV	x	x	x	Complement, if only one governor is present (the verb)
Atr	x	x	x	Attribute
Pnom	x	x	x	Nominal predicate’s nominal part; depends on the copula “to be”
AuxV	x	x	x	Auxiliary Verb “to be” ( <i>být</i> )
Coord	x	x	x	Coordination, main node
Apos	x	x	x	Apposition, main node

afun	_Co	_Ap	_Pa	Description
AuxT	x	x	x	Reflexive particle <i>se</i> , lexically bound to its verb
AuxR	x	x	x	Reflexive particle <i>se</i> , which is neither Obj nor AuxT (passive)
AuxP	x	x	x	Preposition, or part of a compound preposition
AuxC	x	x	x	Conjunction (subordinate)
AuxO	x	x	x	(Superfluously) referring particle or emotional particle
AuxZ	x	x	x	Rhmatizer or other node acting to stress another constituent
AuxX				Comma (but not the main coordinating comma)
AuxG				Other graphical symbols not classified as AuxK
AuxY	x	x	x	Other words, such as particles without a specific (syntactic) function, parts of lexical idioms, etc.
AuxS				The (artificially created) root of the tree (#)
AuxK				Punctuation at the end of a sentence or direct speech or citation clause
ExD	x	x	x	Ellipsis handling (Ex-Dependency): function for nodes which “pseudo-depend” on a node on which they would not depend if there were no ellipsis.
AtrAtr, AtrAdv, AdvAtr, AtrObj, ObjAtr	x	x	x	A node (analytical function: an attribute) which could also depend on its governor’s governor (and have the appropriate other function). There must be no semantic or situational difference between the two cases (or more, in the case of several attributes depending on each other). The order represents the annotator’s preference, but is largely unimportant.

As an example of an analytical-level annotation of a sentence, we present here the representation of the sentence

*Do 15. května mohou cestující platit dosud platným způsobem.*  
*Till 15<sup>th</sup> May can passengers pay hither to valid way.*

(Until May 15, the passengers can pay in the way currently used.)



The original word forms as well as the attribute values of the analytical functions are displayed. This example shows

- the extra root node of the tree (showing the number of the sentence within a file)
- the handling of an analytical verb-form (modal verb *mohou* + infinitive *platit*)
- the fact that the verb is the governing node of the whole sentence (or of every clause in compound sentences), as opposed to the complex subject – complex predicate distinction made even in the otherwise dependency-oriented traditional grammars of Czech, such as (Šmilauer 1969)
- attachment of a manner-type adverbial to an analytical verb-form
- handling of a date expression
- prepositional phrase structure (preposition on top)

and, of course, all the analytical functions assigned to these nodes.

### 3.3 THE TECTOGRAMMATICAL LEVEL

The tectogrammatical level of annotation is based on the framework of the Functional Generative Description (FGD) as developed in Prague by Petr Sgall and his collaborators since the beginning of the 1960's (for a more detailed and integrated formulation, see Sgall, Hajičová and Panevová 1986). The basic principles of annotation are different from those on the analytical level. Instead of requiring every word to become a node, we require that only every autosemantic word become a node. On the other hand, all nodes deleted on the surface – and thus on the analytical level – are added.

The tectogrammatical level is the most developed, complicated but also the most theoretically-based level of semantico-syntactic (or “deep syntactic”) representation. The tectogrammatical level annotation scheme is divided into four “sublevels” (or perhaps better, sub-areas, since they are all intertwined and do not form separate levels):

- dependencies and functional annotation,
- topic/focus and deep word-order annotation,
- coreference, and
- “deep” grammatical information.

As an additional data structure we use a syntactic lexicon, mainly capturing the notion of *valency*. The lexicon is not needed for the interpretation of the tectogrammatical representation itself,<sup>5</sup> but is helpful when working on the annotation since it defines when a particular node that is missing on the surface should be created. In other words, the notion of (valency-based) ellipsis is defined by the dictionary. But before describing the dictionary, let us talk first about the core sublevel of annotation.

#### 3.3.1 DEPENDENCIES AND FUNCTORS

The tectogrammatical level goes beyond the surface structure of the sentence, replacing notions such as “subject” and “object” by notions like “actor”, “patient”, “addressee” etc. The representation itself still relies upon the language structure itself rather than on world knowledge. The nodes in the tectogrammatical tree are *autosemantic words* only<sup>6</sup>.

<sup>5</sup> Nor for further analysis (say, a logical one) based on it, nor (in the other direction) for generation (synthesis) of surface sentences.

<sup>6</sup> By “autosemantic” we mean words that have lexical meaning, as opposed to just grammatical function.

Dependencies between nodes serve as the relations between the (autosemantic) words in a sentence, for the predicate as well as any other node in the sentence. The dependencies are labeled by *functors*<sup>7</sup>, which describe the dependency relations. Every sentence is thus represented as a dependency tree, the nodes of which are autosemantic words, and the (labeled) edges name the dependencies between a dependent and a governor.

The dependency edge labels (functors) are much more detailed than the analytical functions (see the analytical function table in Sect. 3.2). They can be distinguished in several ways; here we use a rather technical classification:

1. the separate root of the tree,
2. verbal and other complementations,
3. coordination, apposition and other functors for other “grouping” nodes,
4. other functors that can be classified as describing neither autosemantic nor “grouping” nodes.

We use over 80 different functors. In the following table, only the most important ones are described.

Functor class	Functor type	Description and examples
Root	Technical	SENT – Technical root of the tree
	Utterance root	PRED – Predicate of main clause in sentence DENOM – Nominal head of nominal expression
Dependency	Verbal Inner Participants	ACT – Actor PAT – Patient ADDR – Addressee ORIG – Origin EFF – Effect
	Time	TWHEN – When? TTILL – Till when? TSIN – Since when? TFHL – For how long? THL – How long? TFRWH – From when? TOWH – To when? TPAR – Parallel events THO – How often?
	Location	LOC – Location (non-directional) DIR1 – From where? DIR2 – Through where? DIR3 – To where?
	Manner	MANN – General manner MEANS – Means to achieve something RESL – Result REG – “with regard to”, “according to” CRIT – Criterion or norm EXT – Extent ACMP – Accompaniment DIFF – Difference CPR – Comparison

<sup>7</sup> At two levels of detail; here we ignore so-called *subfunctors*, which provide the more detailed subclassification.

Functor class	Functor type	Description and examples
Dependency	Implication	CAUS – Cause COND – Condition AIM – Aim INTT – Intention
	Other	BEN – Benefactor SUBS – Substitution HER – Heritage CONTRD – Contradiction RSTR – General attribute (of nouns) AUTH – Authorship APP – Appurtenance or property MAT – Material, container ID – Identity (name or description) COMPL – Complementizer (verb-noun “double dependency”)
Grouping	Coordination	CONJ – Conjunction DISJ – Disjunction CONFR – Confrontation (clauses) CONTRA – Contrariety (expressions) GRAD – Gradation ADVS – Adversative CSQ – Consequence REAS – Reason OPER – Operand (mathematical-like expr.)
	Parenthesis	PAR – Root of parenthesis
	Rhematizer	RHEM – rhematizer (negation, only, also, ...)
Other non-dependency		ATT – attitude PREC – Loose backward reference VOCAT – Addressing vocative expression PARTL – Unidentified particle, interjection INTF – Intensifier DPHR – Part of fixed phrase, idiom CPHR – Semantic part of light verb construct FPHR – Foreign language phrase CM – Part of conjunction

Many nodes found at the morphological and analytical levels disappear<sup>8</sup> (such as function words, prepositions, subordinate conjunctions, etc.). The information carried by the deleted nodes is not lost, of course: the relevant attributes of the autosemantic nodes they belong to now contain enough information to reconstruct them (even though such a reconstruction is not trivial, since it amounts to natural language generation from a semantic representation).

Ellipsis is being resolved at this level. Insertion of nodes is driven by the notion of *valency* (see the section on Dictionary below) and completeness (albeit not in its mathematical sense): if a word is deemed to be used in a context in which some of its valency frames apply, then all the frame's slots are to be “filled” (using regular dependency relations between nodes) by either existing nodes or by newly created nodes, and these nodes are annotated accordingly. Actual ellipsis (often found in coordination, direct speech etc.)<sup>9</sup> is resolved by creating a new node and copying all relevant information from its origin, keeping the reference as well.

<sup>8</sup> Based on the principle of using only autosemantic words in the representation.

<sup>9</sup> Nominal phrases, as used in headings, sports results, artefact names etc. are not considered incomplete sentences, even though they do not contain a predicate; they are rather marked as denominalizations.



Every node of the tree is further annotated by a set of grammatical features that makes it possible to fully capture the meaning of the sentence (and therefore, to recover – at least in theory, see above: the note of the NL generation problem – the original sentence or a sentence with synonymous linguistic meaning). The types of grammatemes belonging to individual nodes are defined by the notion of a *word class* (for autosemantic words, it corresponds to a “semantic class” of the word in question, i.e. semantic noun, verb, adjective or adverb). For example, a (semantic) number is necessary to correctly form a sentence where no numeric expression is attached to a (semantic) noun. Another (obvious) example is (semantic) time: since auxiliaries are no longer present in the sentence structure, we have to have some means of determining present, past or future tense (both relative to the time when the sentence was uttered and between clauses). Verbs do have other grammatemes, such as aspect, iterativeness, modalities of several types (related to modals such as “must” or “may”, or to sentence modality: positive, interrogative, imperative sentence, etc.). Types of pronouns are also recorded where necessary.

### 3.3.2 THE (SYNTACTIC) DICTIONARY (VALENCY LEXICON)

The tectogrammatical level dictionary is viewed mainly as a valency dictionary of Czech (as theoretically defined in: Panevová, 1974, Panevová, 1994; for recent accounts of the computational side and the actual dictionary creation, see: Lopatková et al., 2002, Lopatková, 2003, Lopatková et al., 2003, Hajič et al., 2004, Žabokrtský and Lopatková, 2004). By valency, we mean the necessity and/or ability of words to take other (autosemantic) words as their dependents, as defined below.

Every dictionary entry may contain one or more (*valency*) *frames*. A frame consists of a set of (*valency*) *slots*. Each slot contains a *function* section (the actual *functor*, and an indication whether the functor is obligatory<sup>10</sup>), and an associated *form* section. The form section has no direct relation to the tectogrammatical representation, but it is an important link to the analytical level of annotation: it contains an (underspecified) analytical tree fragment that conforms to the analytical representation of a possible surface expression (or surface “realization”, or simply “form”) of the particular slot. Often, the form section is as simple as a trivial (analytical) subtree with a single (analytical) dependency only, where the dependent node has a particular explicitly specified morphosyntactic case;<sup>11</sup> equally often, it takes the form of a two-edge subtree with two analytical dependencies: one for a preposition (together with its case subcategorization) as the dependent for the surface realization of the root of the frame itself, and one for the preposition’s dependent (which is completely underspecified). However, the form section can be a subtree of any complexity, as might be the case for phrasal verbs with idiomatic expressions etc.

Moreover, the form section might be different for different expressions (surface realizations) of the frame itself. For example, if the frame is a verb and its surface realization is in the passive voice, the form of the (analytical) nodes corresponding to its (tectogrammatical) valency slots will be different than if realized in the active voice. However, relatively simple

<sup>10</sup> By “obligatory” we mean that this functor (slot) must be present at the tectogrammatical level of annotation; this has immediate consequences for ellipsis annotation, cf. below.

<sup>11</sup> Czech has seven morphosyntactic cases: nominative, genitive, dative, accusative, vocative, locative, and instrumental, usually numbered 1 to 7. In the example in section 3.1.1, the case takes the 5<sup>th</sup> position in the positional representation of the morphological tag.

rules do exist to “convert” the active forms into the passive which work for most verbs; therefore, for such verbs, only the canonical (active) forms (by “form” we mean the analytical tree fragment as defined above) are associated with the corresponding valency slots. For irregular passivization problems, there is always the option to enter the two (or more) different realizations explicitly into the dictionary.

Many more rules have to be included, since passivization is not the only process that changes the form of a valency frame; most often, various expressions of modalities (or “near-modalities”, that are not really treated as “true” modalities) have this effect.

A similar mechanism could be defined for nominalizations. Verbal nouns typically share the function section of the valency frame with their source verbs, but the form section might be a regular or an irregular transformation of the corresponding form section. In the current version of the annotation valency lexicon, however, nouns (including verbal nouns) are given in full with their particular valency frame and its form.

Other issues are important in the design of the valency lexicon as well, such as reciprocity etc., but they are outside the scope of this rather brief discussion.

The issue of word sense(s) is not really addressed in the valency dictionary. Two entries might have exactly the same set of valency frames (as defined above, i.e., including the form section(s) of the slot(s)); in such a case, it is assumed that the two words have different lexical meanings (polysemy)<sup>12</sup>. It is practical to leave this possibility in the dictionary (however “dirty” this solution is from the puristically syntactic viewpoint), since it makes feasible linking the entries by a single reference to e.g., the Czech WordNet senses (Pala and Smrž, 2004). The lexical (word sense) disambiguation problem is, however, currently being solved beyond the tectogrammatical level of annotation, even though, for obvious reasons, we plan to link the two, eventually. Then it will be possible to relate the entries for one language to another in their respective (valency) dictionaries (at least for the majority of entries). From the point of view of machine translation, this can be viewed as an additional source of syntactically-based information of form correspondence between the two languages.

For more on valency dictionaries, see: Panevová and Lopatková, 2006, Cinková and Kolářová, 2006, and Urešová, 2006, in this volume.

### 3.3.3 TOPIC, FOCUS AND DEEP WORD ORDER

Topic and focus (Hajičová, 2003, Hajičová et al., 2003) are marked, together with deep word order of the nodes of the tectogrammatical tree. The ordering of nodes is, in general, different from the surface word order, and all the resulting trees are projective by the definition of deep word order.

By *deep word order* (sometimes referred to as “contextual boundness”) we mean a (partial) ordering of nodes at the tectogrammatical level that puts the “newest” information on the right, and the “oldest” information on the left, and all the rest in between, in the order of a discourse-related notion of “newness”. Such an ordering is fully defined at each single-level subtree of the tectogrammatical tree; i.e., all sister nodes *together with their head* are fully ordered left-to-right. The order is relative to the immediate head only; therefore, there exists such a total ordering of the whole tectogrammatical tree that the tree is projective. We believe that the deep word order is language-universal for every utterance in the same context, unless,

---

<sup>12</sup> On the other hand, it is clear that two entries that do *not* share the same set of frames must have different lexical meanings as well, unless truly synonymous at a higher level of analysis.

roughly speaking, the structural differences are “too big” (or, in the case of translation, the corresponding translation is “too free”).

In written Czech, the surface word order roughly corresponds to the deep word order (with the notable systematic exception of adjectival attributes to nouns, and some others), whereas the grammar of English syntax dictates in most cases a fixed order, and therefore the deep word order is often different; (though not always; even English has its means to shuffle words around to make the surface word order closer to the deep one, such as extraposition).

#### 3.3.4 CO-REFERENCE

Grammatical and some textual co-reference relations are resolved and marked. Grammatical co-reference (such as the antecedent of “which”, “whom”, etc., control etc.) is simpler than the textual one (personal pronoun reference resolution etc.).

### 4 THE MANUAL ANNOTATION OF THE PDT

#### 4.1 ORGANIZATION

The manual tagging effort (level 1 annotation, see sect. 3.1) was coordinated by Barbora Vidová Hladká. She supervised a team of 5-7 students who double-tagged<sup>13</sup> the texts selected for the Prague Dependency Treebank. Each annotator was given a description of the tag system (see sect. 3.1.1). Given that Czech morphology is taught extensively in Czech high schools (both junior and senior), that is all they required from the linguistic point of view.<sup>14</sup> The discrepancy rate between any two annotators working on a single text is on average 5%, and there are virtually no opinion-type disagreements – the differences are human performance errors (typos, misunderstandings, etc.). The manual corrections of the annotated text revealed, however, that there are substantial differences among the annotators – ranging from 0.8 to 5% of errors. Other errors (about 1%, apart from missing words) were caused by errors made by the morphological analyzer during preprocessing. About 1,800,000 words have been annotated for PDT 1.0. The tools used for annotation are *sgd* (on Unix) and *DA* (for MS Windows), mutually compatible disambiguation programs with character-based window interface (see sect. 4.2.1).

Not surprisingly, the effort of organizing the structural annotation (sect. 3.2) appeared to be a more complicated task than the organization of the manual morphological annotation. There was little experience to help with such a task: we learned from the LDC's experience with Penn Treebank, but there was no other description available of similar projects. The annotation itself began in November 1996 by constituting a working group of 8 people, 5 of them hired solely for the annotation of the data (the remaining three were faculty members). However, all the newly hired linguists were quite computer-literate, as were the computer science majors. Their background, therefore, allowed us virtually to skip any introduction to computational linguistics and we were able to start immediately with the annotation process itself.

The process of annotation has been (and still is) viewed as a cyclical process where the rules for annotation are constructed on the basis of the evidence found in the data. Thus, we

---

<sup>13</sup> Double-tagging means that the same text is processed twice by different annotators and the results are automatically compared and manually adjudicated to get a single (and presumably better) version.

<sup>14</sup> Ultimately, a slim annotator's handbook has also been developed, to solve certain technically difficult cases (such as foreign names, abbreviations, incomplete sentence with errors, etc.), mostly according to convention.

explained the basic principles of annotation to the annotators and asked them to use existing grammar books, most notably (Šmilauer, 1969), an old but still the best Czech grammar description. This description also builds on a dependency framework, although there are some (easily identifiable and replaceable) deviations. We were aware of the fact that there are many gaps in such a traditional grammar from the point of view of an explicit annotation based on the above basic principles: mainly, the request to have each input word represented by a node in the tree (a request quite natural from the computational point of view) is largely not reflected in any human-oriented grammar description. Nevertheless, before starting to write authoritative guidelines based on such a grammar, we believed that a final version could be constructed on-the-fly with annotation corrections made later, should the rules change.

The key software tool used was the GRAPH program, developed initially as an undergraduate thesis in 1995/96, and substantially enhanced afterwards (see also below, sect. 4.2). This tool allows for graphical viewing and editing of the dependency representation of annotated sentences.

All the annotators have helped to formulate the final wording in the Guidelines, and each of them is responsible for a certain section of the Guidelines (for example, for subject, or rhematizers and multiword units, etc.). Given their effort in this respect, and also their contribution to the formulation of the annotation rules during the first phase of the project, they have all become not only the annotators, but also the authors of the Guidelines (Bémová et al., 1997).

Ultimately, 90,000 sentences (1.3 mil. word tokens) are available as part of the Prague Dependency Treebank at the end of the project. There were also other non-trivial tasks connected to the project: for example, tagged data (level 1) had to be merged with the structurally annotated data, changes in morphology had to be incorporated, the resulting format had to be converted to SGML, etc. The PDT version 1.0 which contained the manually annotated data on the morphological and analytical levels was published in the fall of 2001 at the Linguistic Data Consortium in Philadelphia (Hajič et al., 2001).

Annotation at the tectogrammatical-level commenced in 2001. The preliminary guidelines were used (already published as part of the PDT 1.0 CDROM). The annotators did not start from scratch this time: the analytical-level trees selected for tectogrammatical annotation had been preprocessed by a set of rules to decrease the annotation effort in cases where such rules can be formulated unambiguously, or for technical transformations of the tree that have been in conventional use (Böhmová, 2001 and Böhmová and Hajičová, 2003). Later, after a certain volume of the annotated data was at our disposal, the functor assignment was rewritten to use a decision tree mechanism to further ease the task of manual functor assignment.

Based on the division of work into sublevels (see 3.3 above), the actual annotation also proceeded along the four lines, with four groups (teams) working in parallel (some people participated in more than one effort). Also, a new platform-independent tool was developed, called TrEd (Hajič, Hladká and Pajas, 2001), described in more detail below.

First, we concentrated on the dependencies and functors, together with developing the valency dictionary and linking it to the corpus. Separately, exploratory work started for topic/focus and deep word order annotation, and for co-reference annotation. The work on grammemes was postponed until 2003.

The corpus has been annotated only once (50,000 sentences in total), with every fourth sentence double annotated (structure and functors) for inter-annotator agreement evaluation purposes. The valency dictionary has been developed by the annotators, sharing the dictionary

amongst themselves during the course of the annotation. The structural annotation was finished by mid-2003, and an 18-month checking and correction period ensued.

The newly developed annotation tool, data markup and sophisticated organization of the technical work made it possible to work in parallel not only along the four major lines of annotation, but also within each line, to make changes and corrections relatively independently.<sup>15</sup> Those changes involve corrections after various automatic checks, merging the data from the four lines of annotation, corrections at the morphological and analytical levels (involving errors that were discovered during the tectogrammatical annotation and, sometimes, because of it), and many more things. The valency dictionary has also been “unified” by a single person, with changes mapped back to the data and manually corrected. Grammatemes have mostly been filled in automatically, based on quite sophisticated rules, even though some simplifications to their definitions had to be made to avoid the most time-consuming annotation tasks.

## 4.2 TOOLS

Manual annotation does not mean that people are typing complicated formal representations by hand into a computer. Even the first annotation attempts in the times when graphical editing was resource-demanding and therefore not feasible were guided by software tools. These tools allowed the annotators to assign a formally correct entry only, avoiding expensive checking-and-correction processes afterwards.

On the basis of the computing power available today, we decided that for the annotation of the PDT we should use tools that are as advanced as possible.

### 4.2.1 MORPHOLOGICAL DISAMBIGUATION: SGD AND DA

We use a special purpose tool for morphological annotation, which allows for an easy disambiguation of lemmas and tags as output by the morphological analyzer. The tool was first implemented under the Linux operating system under the name `sgd` (and is capable of running also on Solaris and other operating systems of the Unix type). It has been re-implemented also for the Windows platforms (under the `DA` name), to allow for annotators who were not able to install Linux on their home machines. The user interface is identical. The `sgd` tool is text-terminal-based so it can be relatively easily (character coding problems aside) used from any `vt100`-capable terminal, as well as from `xterm` or similar programs.

The tools work full screen on texts in a SGML format (as defined by the Czech National Corpus’s standard data type definition, namely, the `csts.dtd`) preprocessed by a morphological processor (see sect. 3.1.2 above). The annotators are presented with a list of ambiguous words as found in the input text (expandable to full text list, with ambiguous words marked by an asterisk). The full text context is also displayed in a separate window, with the active word marked by reverse video. The largest part of the screen is devoted to the disambiguation process itself. The annotator first chooses the correct lemma, and then, if needed (as is usually the case, since more than 45% of words (tokens) are morphologically ambiguous in Czech), the correct tag. S/he has also the option to edit both the lemma and the tag, in case the morphological processor did not know the word at all or made an error. The text is then saved with the lemmas and tags chosen by the annotators marked appropriately. There are other tools related to morphological annotation, but these are mostly standard

---

<sup>15</sup> Otherwise we would have needed a lot more time than those 18 months to finish the work.

Unix tools (diff, flex, awk, perl etc.). These help to resolve differences between two annotators on the same text and to do other conversions of the material.

#### 4.2.2 THE ANALYTICAL LEVEL ANNOTATION TOOL: GRAPH

The analytical level, even though we are interested in the structure and one attribute (analytical function) “only”, is a major challenge because of its inherently non-linear nature. We have used a program called, rather unimaginatively, GRAPH. This program works under Microsoft Windows (3.1 and 95) and was developed as an undergraduate thesis based on an initial specification developed long before the annotation project actually began. It has changed a lot since then – there were about 40 versions of it with bug fixes, minor and major updates. The program allows for drag-and-drop style editing of trees with annotated nodes. It is not just for dependency-based formal representations, even though it has special features (such as visual node ordering) which were inspired by such formalisms. Several files can be opened concurrently; (sub)trees may be copied among them using multiple-buffer clipboard, and files may be searched for node annotations. The display of trees (attributes to be displayed, colors, fonts, line thickness, etc.) is fully configurable to suit the task in hand as well as the annotator’s preferences, which might depend on the hardware or other differences. The program can be completely mouseless driven, too.

One of the major features of the GRAPH program is the possibility to use macros – in other words, the program is programmable. The programming language is similar to C but contains only those constructs necessary for the annotation tasks. The functions can be invoked interactively (by a keypress) or from the command line when starting the GRAPH program. These macros have been used so far for two different purposes:

- as shortcuts, requested by the annotators, to avoid opening 2 or 3 menu windows when selecting the appropriate analytical function for a node in the tree;
- for a preliminary assignment of analytical functions to nodes when the tree structure is built, but before the manual node annotation.

The programming facility is not intended to be used by the annotators, but they are able to use the macros prepared by programmers. These macros can also be used for tree checking and transformations, if necessary e.g. after changes made in the annotation rules. The programming language facilitates almost all the editing operations made normally by the annotators, including tree restructuring. Thus, in principle, they could also be used for the initial tree structure assignment.

The shortcuts allow the annotators to assign an analytical function to an active node by a simple keypress, or Ctrl and/or Shift plus a key in case of functions “suffixed” by `_Co`, `_Ap` or `_Pa`. These macros also store the value previously assigned to this node, and another macro function, when activated, can thus revert to the previous value, should the annotator decide that s/he has made a mistake. There are also macros for node swap, for assignment of the `Attr` function to all nodes in a subtree (a frequent case near the leaves of the tree), and for special coordination and apposition handling.

The initial analytical function assignment was performed by an 800+ line-long function which tried to assign the most plausible analytical function to every node of a tree. The assignment was based on relatively simple hand-crafted rules. They were far from perfect, and sometimes intentionally ignored some complicated contexts, but as the feedback from the annotators showed, they were correct in almost 80% of cases. The initial assignment function could also be used (under a different name) on a file as a whole, which meant that



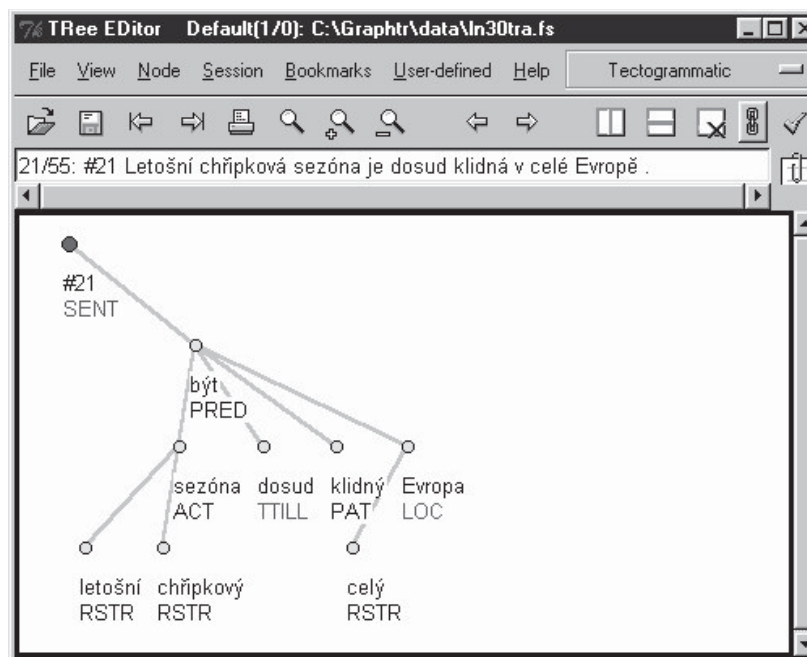
the annotators did not have to run the macro on every tree. The batch feature of the GRAPH program also allowed the same macro to be run on many files using a single command.

#### 4.2.3 TRED: THE TECTOGRAMMATICAL ANNOTATION TOOL

The graphical tree editor Tred was developed originally when the final corrections and changes were made to the analytical-level annotation before it was published. However, due to its advanced properties, easy extensibility and modularity and platform independence it has eventually been chosen as the main tool for tectogrammatical annotation.

Tred is written in the **perl** programming language, it uses the **perlTk** extension for its graphical interface, and its basic functionality is the same as for the GRAPH tool (see above). It has been extensively used on both the Linux and Windows platforms. It can be customized for both manual annotation work and for batch processing of the annotated data. Thanks to its **perl** roots, it can be easily extended, and additional modules can equally easily be added for online<sup>16</sup> data processing, substantially extending the original idea of macros of the GRAPH editor. One of the extensions that has been heavily used is its lexical interface to the valency lexicon, which facilitates both lexicon maintenance (adding, deleting, modifying entries and their associated valency frames) and linking the lexical entries to the annotated data.<sup>17</sup> Tred also contains a general search interface that can be used by the annotators as well as during the subsequent checking of data; both simple and sophisticated searches (again, using **perl** expressions) can be launched.

The example below shows an open editing window with a tectogrammatical representation of the sentence “*This-year flu season is so-far quiet in [the] whole Europe.*”



<sup>16</sup> By “online” we mean during the manual annotation.

<sup>17</sup> The valency lexicon maintenance module can also be used outside of Tred as a stand-alone application.



Tred can also display additional links that are not part of the basic tree structure, in various graphic forms. It is used e.g. for co-reference annotation, which links the consequent to the antecedent by a coloured dashed arrow.

Two files can be displayed at the same time in two windows, side-by-side, with differences automatically highlighted. This is used for visual checking of the double-annotated data or different versions of the data. Also, the same sentence can be displayed on the analytical and tectogrammatical levels, facilitating a comparison between the annotations of a particular sentence at these two levels.

## 5 TREEBANK USAGE: TAGGING AND PARSING UNRESTRICTED TEXT

The treebank can obviously be used for further linguistic research, as it contains a lot of material annotated in a way directly usable by original linguistic research, readily searchable using different criteria. However, in the present contribution, we will discuss a more “computational” use of the treebank, namely, as a basis for creating a statistically-based tagger and a parser of unrestricted written text.

### 5.1 FULL MORPHOLOGICAL TAGGING

We have developed a statistical model which has been successfully used for tagging (full morphological disambiguation), where it improved accuracy by 5 percentage points, from 80% (Hladká, 1994, Hajič and Hladká, 1997a) to 93% (Hajič and Hladká, 1998, Hajič, 2004) to 95% (Krbec et al., 2001). The statistical models are based on both the “classic” HMM:

$$p(T|W) = \prod_{i=1..n} p(t_i | t_{i-2}, t_{i-1}) p(w_i | t_i) / p(W)$$

where we use the Bayes formula to reverse the conditioning (simulating the well-known source-channel paradigm) and the trigram approximation for the tag language model, or the exponential probabilistic model of the form

$$p(y|x) = e^{\sum_{i=1..n} \lambda_i f_i(y,x)} / Z_{\lambda}(x)$$

where  $f_i(y,x)$  is a feature selector function which returns 1 or 0 depending on the value of  $y$  and the context  $x$ ,  $\lambda_i$  is its weight, and  $Z_{\lambda}(x)$  is a normalization factor making the distribution a probabilistic distribution which adds to 1.

The crucial property of this model, used successfully for many applications in tagging as well as in machine translation, is the set of  $n$  features (typically in the order of hundreds or thousands). These features are selected automatically, based on objective criteria, from a much larger “pool” of available features. The selection of features may be guided by two different principles: a “minimal cross-entropy” principle, which compares the probability distribution constructed to the training data (using the cross-entropy measure, or simply the probability of training data), or “minimal error rate” (again, on training data). We have chosen the second principle, as it addresses the problem in hand more directly.

The selection of features, however, depends also on the values of  $\lambda_i$ . The basic method for feature weight computation is the Maximum Entropy method. Unfortunately, this method involves several numerical iterative algorithms, which makes it rather slow. We believe, based on our experience with similar models, (and with smoothing, which displays a similar “weighting” issue, in general) that exact weight computation is not so important to the resulting model performance, and thus that the values of  $\lambda_i$  may be roughly – and quickly –

approximated instead. This would allow us to select features from larger pools, thus making it possible for more sophisticated features to be selected.

## 5.2 PARSING

There are many attempts to parse sentences of natural language at various levels (Brill, 1993a, Brill, 1993b, Collins, 1996, Collins, 1997, Charniak, 2000, Ribarov, 1996). We aim here at syntactico-semantic parsing of unrestricted text. It is a well-known fact that hand-crafted rules work well for restricted domains and vocabularies, whereas they generally fail for unrestricted text parsing. So far the (partial and imperfect, but still the best available) answer to this problem has been statistical parsing based on training on manually annotated data.

Having such a resource available for Czech (the Prague Dependency Treebank as described in the previous sections), we have successfully applied the Collins parsing model to Czech (Hajič et al., 1998, Collins et al., 1999). The Collins parser currently achieves 82% dependency accuracy when trained on the PDT 1.0 analytical level training data. We also have at our disposal a modified version of Charniak's parser for Czech (unpublished), which achieves a slightly better performance (84% dependency accuracy when trained on the same data). Several other parsers have been developed since then, but none of them surpassed these two, except that Zeman, (2004) constructed a combined "superparser" that shows the best results so far by combining several of the available parser outputs (having almost 85% accuracy for the best parsing method). These parsers are complemented by a decision-tree implementation of function assignment that performs with much the same accuracy.

For tectogrammatical parsing, we currently use a set of manually written rules (Bohmová et al., 2003) that in fact requires that the analytical parse be completed by either the Collins or Charniak parser, and then it transforms the analytical level tree to the tectogrammatical one. The result is worse than that on the analytical level, but we believe that it will improve once statistical methods are employed, once the manual annotation at the tectogrammatical level is completed. The functor assignment is being performed by a mechanism similar to the analytical one, namely, a decision-tree functor classifier (implemented using the C5.0 software tool), with accuracy of over 80%.

## 6 CONCLUSIONS

Building a treebank is an expensive and organizationally complicated task, especially when a rich annotation scheme is adopted such the one used in the Prague dependency treebank, where (roughly speaking) each word token from the selected text needs over 30 attribute-value pairs to be completed.

Everybody would certainly agree that to build a treebank is a difficult task. Our belief is, however, that all the hard work will pay off – in that not only we who are building it, but all the computational linguists interested in the morphology and syntax of natural languages in general, and of Czech or other inflectional and free word order languages in particular, will benefit from its existence. The building of the treebank has already been very fruitful even now, halfway through the whole treebank annotation: we have effectively been forced to describe the syntactic behaviour of Czech more explicitly and more widely (in the sense of overall coverage, including also "peripheral" phenomena) than ever.

## 7 REFERENCES

BĚMOVÁ et al. (1997): Anotace na analytické rovině – příručka pro anotátory [Annotation on the Analytical Level – Annotator's Guidelines], Technical Report #4 (draft), LJD ÚFAL MFF UK, Prague, Czech Republic (in Czech).

- BÖHMOVÁ, ALENA (2001): Automatic Procedures in Tectogrammatical Tagging. In: PBML 76. MFF UK Prague.
- BÖHMOVÁ, ALENA; HAJIČOVÁ, EVA (2003): Large Language Data and the Degrees of Automation. In: Proceedings of XVII<sup>th</sup> International Congress of Linguists, CD-ROM. Matfyzpress, MFF UK Prague.
- BRILL, E. (1993a): Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach. In: Proceedings of the 3<sup>rd</sup> International Workshop on Parsing Technologies, Tilburg, the Netherlands.
- BRILL, E. (1993b): Transformation-Based Error-Driven Parsing. In: Proceedings of the 12<sup>th</sup> National Conference on Artificial Intelligence.
- CINKOVÁ, S.; KOLÁŘOVÁ, V. (2006): Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. This volume.
- COLLINS, M. (1996): A New Statistical Parser Based on Bigram Lexical Dependencies. In: Proceedings of the 34<sup>th</sup> Annual Meeting of the ACL'96, Santa Cruz, CA, USA, June 24-27, pp. 184-191.
- COLLINS, M. (1997): Three Generative, Lexicalised Models for Statistical Parsing. In: Proceedings of the 35<sup>th</sup> Annual Meeting of the ACL/EACL'97, Madrid, Spain, pp. 16-23.
- HAJIČ, JAN (2004): Disambiguation of Rich Inflection. Karolinum, Charles University Press, Prague. 332pp.
- HAJIČ, JAN; COLLINS, MICHAEL; RAMSHAW, LANCE; TILLMANN, CHRISTOPH (1999): A Statistical Parser for Czech. In: Proceedings of ACL'99, Maryland, USA.
- HAJIČ, J., AND HLADKÁ, B. (1997a): Probabilistic and Rule-based Tagger of an Inflective Language – A Comparison. In: Proceedings of the 5<sup>th</sup> Conference on Applied Natural Language Processing, ACL, Washington, DC, USA, pp. 111-118.
- HAJIČ, J., AND HLADKÁ, B. (1998): Morfologické značkování korpusu českých textů stochastickou metodou [Morphological tagging of Czech corpora using stochastic methods]. In: Slovo a Slovesnost, Vol. 58, No. 4, ÚJČ AV ČR, Prague.
- HAJIČ, JAN; VIDOVÁ-HLADKÁ, BARBORA; PAJAS, PETR (2001): The Prague Dependency Treebank: Annotation Structure and Support. In: Proceeding of the IRCS Workshop on Linguistic Databases University of Pennsylvania, Philadelphia, USA, pp. 105-114.
- HAJIČ, J., AND RIBAROV, K. (1997): Rule-Based Dependencies. In: Proceedings of the Workshop on Empirical Learning of Natural Language Processing Tasks, MLNet, Prague, Czech Republic, April 23-25, pp. 125-136.
- HAJIČ, JAN; PANEVOVÁ, JARMILA; UŘEŠOVÁ, ZDEŇKA; BÉMOVÁ, ALEVTINA; KOLÁŘOVÁ, VERONIKA; PAJAS, PETR (2003): PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: Proceedings of the Second Workshop on Treebanks and Linguistic Theories, pp. 57-68. Vaxjo University Press.
- HAJIČ, JAN; VIDOVÁ-HLADKÁ, BARBORA; PANEVOVÁ, JARMILA; HAJIČOVÁ, EVA; SGALL, PETR; PAJAS, PETR (2001): Prague Dependency Treebank 1.0. CDROM. CAT: LDC2001T10, ISBN 1-58563-212-0. Linguistic Data Consortium, Univ. of Pennsylvania, Philadelphia, USA. Also at <http://ufal.mff.cuni.cz/pdt>.
- HAJIČOVÁ, EVA (2003): Information structure and syntactic complexity. In: Investigations into formal Slavic linguistics, pp. 169-180. Peter Lang.
- HAJIČOVÁ, EVA; SGALL, PETR; VESELÁ, KATEŘINA (2003): Information structure and contrastive topic. In: Formal approaches to Slavic linguistics. The Amherst Meeting 2002, pp. 219-234. Michigan Slavic Publications.
- HAJIČOVÁ, EVA; HAVELKA, JIŘÍ; SGALL, PETR; VESELÁ, KATEŘINA; ZEMAN, DANIEL (2004): Issues of Projectivity in the Prague Dependency Treebank. In: Prague Bulletin of Mathematical Linguistics MFF UK.
- HLADKÁ, B. (1994): Programové vybavení pro zpracování velkých českých textových korpusů [Software for Large Czech Corpora Annotation], MSc thesis, MFF UK, Prague, Czech Republic.
- LOPATKOVÁ, MARKÉTA (2003): Valency in the Prague Dependency Treebank: Building the Valency Lexicon. In: Prague Bulletin of Mathematical Linguistics, pp. 37-60. MFF UK.

- LOPATKOVÁ, M., PANEVOVÁ, J. (2006): Recent developments in the theory of valency in the light of the Prague Dependency Treebank. This volume.
- LOPATKOVÁ, MARKÉTA; ŘEZNÍČKOVÁ, VERONIKA; ŽABOKRTSKÝ, ZDENĚK (2002): Valency Lexicon for Czech: from Verbs to Nouns. In: Text, Speech and Dialogue. 5<sup>th</sup> International Conference, TSD 2002, pp. 147-150. Springer.
- LOPATKOVÁ, MARKÉTA; ŽABOKRTSKÝ, ZDENĚK; SKWARSKA, KAROLINA; BENEŠOVÁ, VÁCLAVA (2003): VALLEX 1.0 Valency Lexicon of Czech Verbs. MFF UK.
- MARCUS, M.P., SANTORINI, B. AND MARCINKIEWICZ, M. (1993): "Building a large annotated corpus of English: the Penn Treebank," Computational Linguistics, vol. 19, pp. 313-330.
- PANEVOVÁ, J. (1974), On Verbal Frames in Functional Generative Description. Part I, Prague Bulletin of Mathematical Linguistics 22, 3-40, Part II, Prague Bulletin of Mathematical Linguistics 23, 1975, 17-52.
- PANEVOVÁ, J. (1994), Valency Frames and the Meaning of the Sentence. In: The Prague School of Structural and Functional Linguistics (ed. by Ph. L. Luelsdorff), Linguistic and Literary Studies in Eastern Europe 41, Amsterdam-Philadelphia: John Benjamins, 223-243.
- RIBAROV, K. (1996): Automatická tvorba gramatiky přirozeného jazyka [The Automatic Creation of a Grammar of a Natural Language], MSc thesis, MFF UK Prague.
- ŘEZNÍČKOVÁ, VERONIKA (2003): Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora. In: Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003), pp. 88-97. UCREL, Lancaster University.
- PALA, K., SMRŽ, P. (2004): Building Czech Wordnet. Romanian Journal of Information Science and Technology Special Issue. Ed. By D. Tufis. Vol. 7, No. 1-2. pp. 79-88.
- SGALL, P. et al. (1986): The Meaning of the Sentence and Its Semantic and Pragmatic Aspects, Reidel Publishing Company, Dordrecht, Netherlands; Academia, Prague, Czech Republic.
- ŠMILAUER, V. (1947): Novočeská skladba [Syntax of Contemporary Czech]. 1<sup>st</sup> ed., Prague.
- ŠMILAUER, V. (1969): Novočeská skladba [Syntax of Contemporary Czech]. 3<sup>rd</sup> ed., SPN, Prague, 574 pp.
- UREŠOVÁ, Z. (2006): The verbal valency in the Prague Dependency Treebank from the annotator's point of view. This volume.
- ZEMAN, D. (2004): Parsing with a statistical dependency model. PhD Thesis. MFF UK Prague.
- ŽABOKRTSKÝ, ZDENĚK; LOPATKOVÁ, MARKÉTA (2004): Valency Frames of Czech Verbs in VALLEX 1.0. In: Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference, pp. 70-77.

## ABSTRAKT

Pražský závislostní korpus (PDT, Hajič et al., 2001) obsahuje bohatou morfológickou, syntaktickou a syntakticko-sémantickou informaci ve formě manuálně provedené anotace. V tomto článku představujeme stručný popis celého PDT včetně seznamů hlavních značek užitých pro anotaci na jednotlivých rovinách. Rovněž jsou uvedeny některé zkušenosti z průběhu anotace, a jsou popsány i nástroje, které byly při anotaci použity. Na závěr článku uvádíme možnosti využití manuálně anotovaných korpusů pro vytváření automatických programových nástrojů pro analýzu jazyka na morfológické a syntaktické rovině.

## ABSTRACT

The Prague Dependency Treebank (Hajič et al., 2001) is approaching the publication of its second version in which tectogrammatical annotation is being added to morphological and analytical (surface-syntactic) annotation. In this article, the Prague Dependency Treebank is described as a whole, including its brief history. In this volume, there are three more papers with detailed accounts of some of the most recently tackled phenomena occurring at the tectogrammatical level of annotation (Panevová and Lopatková, 2006, Cinková and Kolářová, 2006, and Uřešová, 2006).



# Towards the Underlying Structure Annotation of a Large Corpus of Texts

EVA HAJIČOVÁ

## 1 INTRODUCTION

Present day linguistic research is not possible without an intensive exploitation of large corpora of texts. The collection of a large corpus and making it available on computers is in itself a very demanding task; the corpus has to be supported by friendly and clever computerized tools to make it possible to search in the corpus, to formulate appropriate requests and to get the required answers. The use of a corpus broadens the linguists' horizons and is an important starting point for the development of more ambitious language resources for linguistic research, namely for the creation of corpora annotated on different levels of language.

The most topical, urgent and very ambitious task is now an underlying level annotation (see Uszkoreit 2004; Sgall et al. 2004). A necessary requirement of such an annotation is to have a well developed scenario based on a solid and well-tested conception of a theoretical syntactic framework, which, of course, is being gradually complemented by further research. On the other hand, corpus annotation offers a highly effective, reliable and useful way to test, complement or modify the existing theoretical framework and to develop it further on the basis of language data supplied by real texts.

## 2 PRAGUE DEPENDENCY TREEBANK

2.1 The Prague Dependency Treebank (PDT in the sequel) is conceived as a collection of 2000 samples each containing 50 continuous sentences from current Czech texts (samples are taken at random from the Czech National Corpus), annotated – besides a complex scheme of morphemic tags – on two layers of dependency-based sentence syntax, the first of which – the analytic one (for a detailed description, see Hajič 1998) – has no theoretical status and is considered to be an auxiliary, intermediate step towards the underlying (deep syntactic) level of annotation, the so-called tectogrammatical tree structures (TGTs), in which nodes are also reconstructed for items deleted in the surface shape of the sentences (Hajičová 2000). Besides rendering a theoretically substantiated, detailed view of sentence structure, these representations are designed in a way that allows for an inclusion of information on the topic-focus articulation of the sentence and on both grammatical and textual co-reference relations.

2.2 The conception of the tectogrammatical layer of annotation is based on the Functional Generative Description of Language (FGD) as proposed by Petr Sgall in the 1960s and further developed by the group of theoretical and computational linguistics at Charles University in Prague (see e.g. Sgall et al. 1986). The FGD approach can be briefly characterized by the following five characteristics and claims:

- (a) syntactic relations are dependency-based;
- (b) on the underlying (tectogrammatical) level, the dependency tree should meet the condition of projectivity; deviations from projectivity at the surface shape of the sentence can best be described as differences between the surface (morphemic) word order and the deep word order that are contextually restricted (for a most recent discussion, see Hajičová et al. 2004);
- (c) a formal description of language should also include a description of the topic-focus articulation (information structure);
- (d) the number of layers of the description should be minimized; it is believed that the following layers should be distinguished: phonological, morphemic (the representation in terms of strings), and tectogrammatical (the representation in terms of projective dependency trees, with complex symbols as the labels of the nodes);
- (e) coordination can be understood as a “third” dimension of the structure, which is no longer a tree but an acyclic graph.

2.3 In FGD, as well as in PDT, the focus of attention is on the representation of the sentence at the tectogrammatical level. It is a matter of course that the “deeper” one goes into the annotation scheme, the smaller the share of automatic procedures and the larger the proportion of intellectual work, i.e. hand-made corrections, complementations and modifications. In the present stage of tectogrammatical annotation, we proceed in a cascaded way:

- (i) the first phase consists of automatic pre-processing, in the course of which the output of the analytic annotation – the ‘analytic’ tree structures – is automatically modified for the structure (deletion of the nodes for prepositions and auxiliaries and addition of the information these nodes carry to their head nodes and some other changes that can be done automatically);
- (ii) in the second phase, the output of this pre-processing module is checked and modified for the tree structure and for the labels indicating the kind of dependency; one of the crucial differences between the analytic and tectogrammatical tree structures is the fact that, while in the analytic trees there is a node for each word (even a punctuation mark) of the sentence but no node can be added, in the tectogrammatical tree structures the nodes of auxiliaries and prepositions are deleted (hidden) but new nodes should be established for units that are deleted on the surface level but should be present in the underlying structure of the sentence;
- (iii) in the third phase, the nodes of the checked and corrected tree structures are assigned one of the values of the attribute of topic-focus articulation;
- (iv) in the fourth phase, basic co-referential relations are marked.

An invaluable and most substantial precondition for successful and effective annotation consists in the development and implementation of friendly software tools (the tree editor TRED, see the References below).

2.4 In the present contribution, we would like to briefly characterize two aspects of the annotation mentioned above under (iii) and (iv), namely the annotation of topic-focus articulation (Sect. (3)) and the assignment of the basic co-reference relations (Sect. (4)). For some more detailed aspects of the PDT annotation, see other papers on PDT in this volume and for the overall scenario see the PDT web page.

### 3 ANNOTATION OF THE TOPIC-FOCUS ARTICULATION

3.1 As has been documented in the course of theoretical research and is now commonly accepted by the linguistics community, the topic-focus articulation of the sentence (its



information structure, TFA in the sequel) is semantically relevant (even for the truth conditions) and as such should be an inherent part of the description of the underlying structure of the sentence (for the most recent discussion, see Hajičová, Partee and Sgall 1998). To support this claim, let us observe the following pairs of sentences (the intonation centre is denoted by capitals):

- (1) Český se mluví na MORAVĚ. – Na Moravě se mluví ČESKY.  
Transl.: Czech is spoken in Moravia. – In Moravia, one speaks Czech.
- (2) Dogs must be CARRIED. – DOGS must be carried. (=Carry DOGS).
- (3) Staff behind the COUNTER. – STAFF behind the counter.
- (4) Nejde o to, že Janouch koupil LEKSELLŮV GAMMA NŮŽ, ale že Leksellův gamma nůž koupil JANOUCH.  
Transl. The matter is not that Janouch bought LEKSELL GAMMA KNIFE, but that Leksell gamma knife was bought by JANOUCH.
- (5) Dobrá zpráva: Češi udělali REVOLUCI. Špatná zpráva: revoluci udělali ČEŠI.  
Transl.: Good news: Czechs made revolution. Bad news: Revolution was made by Czechs.

The semantic relevance of TFA can best be illustrated by sentences with negation (Hajičová 1984); cf. the possible interpretations of (6):

- (6) Nepřišel, protože byl nemocen.  
Transl.: (He) did not come because he was ill.
- (6')(a) he did not come, and the reason why he did not come is that he was ill
- (6')(b) he came – not because he was ill but for some other reason

In interpretation (6')(a), the negation is included in the topic of the sentences (the sentence “is about” his not-coming, i.e. his not-coming is in the topic of the sentence and as such it triggers a presupposition), while in interpretation (6')(b) the sentence “is about” his coming and the negation concerns the relation between the topic and the focus; under this interpretation it is not necessarily the case that he was ill (he might have been ill but it need not be so, the matter negated is just that the reason for his coming was not that he was ill).

The articulation of the sentence into topic (what the sentence is about) and focus (what the sentence says about the topic) is based on the notion of contextual boundness; in the prototypical case, contextually bound (cb) nodes are parts of the topic and the contextually non-bound (nb) ones are in the focus; this is the case for nodes directly depending on the main verb; the more deeply embedded nodes may be either cb or nb according to their relation to their governors.

The cb nodes may be either non-contrastive or contrastive; the main operational criteria of this opposition are as follows: (i) a contrastive cb node may be substituted (in the surface shape of the sentence) by a long form of the pronoun (see (7)), and (ii) in the spoken form of the given sentence, the contrastive cb node prototypically carries a rising accent (see Veselá et al. 2003).

3.2 It is evident that TFA should also be captured in the deep level annotation of a large corpus such as PDT (for a more detailed description, which we briefly summarize here, see Veselá and Havelka, 2003; the results of the evaluation of the annotators' agreement are presented in Veselá et al. 2004).



For the representation of TFA, a special attribute has been established with three values, one of which has to appear with every node in a TGTS:

- (i) T: a non-contrastive CB node (standing to the left of its governor in the TGTS);
- (ii) F: an NB node (if different from the main verb, then following after its head word in the TGTS);
- (iii) C: a contrastive CB node.

Example (8) and the corresponding (rather sketchy) TGTS in Figure 1 illustrate the result of the TFA assignments:

(8) Už první pohled na atypickou karosérii potvrzuje, že se jim podařilo tento záměr naplnit.

Lit. E. transl.: Already first look at atypical car-body confirms, that Refl. them succeeded this intention to-fulfil.

E. transl.: Already the first look at the atypical car-body confirms that they have succeeded in meeting the intention.

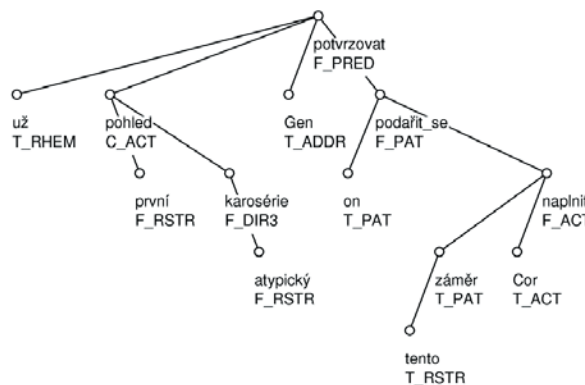


Figure 1: A sample TGTS

Note: Gen and Cor are formal lemmas of nodes restored in the TGTS's (i.e. their correlates are absent in the surface shape of the sentences).

Thus, the division of the sentence into topic and focus can be derived on the basis of the assignments of the TFA values and it corresponds to the context in which the sentence occurs in the annotated text.

3.3 As discussed in Veselá et al. (2004), the evaluation experiments have documented that there is an overall agreement among human annotators of approximately 80%. This indicates that the annotation of TFA is feasible, and that the perception of contextual boundness is not too subjective to disallow a sufficiently reliable annotation of texts. The substantial increase in agreement towards the end of the evaluation also indicates that the completion of the manual for annotation helped to increase the reliability of annotation and that a further elaboration of hypotheses and their applications in Functional Generative Description helped the annotators to understand the subject matter more deeply and to make the annotations of TFA more consistently.

#### 4 ANNOTATION OF BASIC RELATIONS OF CO-REFERENCE IN PDT

4.1 In the Prague Dependency Treebank, co-reference is understood as an asymmetrical binary relation between the nodes of a TGTS (not necessarily the same TGTS), or, as may be the case, as a relation between a node and an entity that has no corresponding counterpart in the TGTS(s). The node from which the co-referential link leads is called an anaphor, and the node to which the link leads is called an antecedent.

The current scenario of PDT provides three co-referential attributes *coref*, *cortype* and *corlemma* (for a more detailed description, see Kučová et al. 2003, and Kučová and Hajičová 2004). The attribute *coref* contains the identifier of the antecedent; if there is more than one antecedent of the anaphor in question, the attribute *coref* includes a sequence of identifiers of the relevant antecedents. The attribute *cortype* includes the information on the type of co-reference (the possible values are *gram* for grammatical and *text* for textual co-reference), or a sequence of the types of co-reference, where each element of *cortype* corresponds to an element of *coref*. The attribute *corlemma* is used for cases of a co-reference between a node and an entity that have no corresponding counterpart in the TGTS(s): for the time being, there are two possible values of this attribute, namely *segm* in the case of a co-referential link to a whole segment of the preceding text (not just a sentence), and *exoph* in the case of an exophoric relation.

In order to facilitate the task of the annotators and to make the resulting structures more transparent, the co-reference relations are captured by arrows leading from the anaphor to the antecedent and the types of co-reference are distinguished by the different colours of the arrows. Certain notational devices are used in cases when the antecedent is not within the co-text (exophoric co-reference) or when the link should lead to a whole segment rather than to a particular node. If the anaphor co-refers to more than a single node or a sub-tree, the link leads to the closest preceding co-referring node (sub-tree). If a possibility to choose between a link to an antecedent or to a postcedent exists, the link always leads to the antecedent.

Manual annotation is made user-friendly by a special tool in the TRED editor used for tree-structure assignment (Kučová et al. in prep.); the values of the attributes of co-reference with each node of the tree will be assigned by an automatic procedure.

In our project, two types of co-reference are distinguished: grammatical co-reference (i.e. with verbs of control, with reflexive possessive pronouns, and with relative pronouns) and textual (which may cross sentence boundaries), both endophoric and exophoric. For the time being, the PDT annotation of textual co-reference covers only cases in which a demonstrative or anaphoric pronoun (also in its zero form) is used (with the demonstrative pronoun, we take into consideration only its use as a noun, not as an adjective). We do not include cases of exophoric co-reference rendered by a pronoun of the 1<sup>st</sup> and 2<sup>nd</sup> persons (be they expressed explicitly or by a zero form, i.e. deleted in the surface shape of the sentence). For the time being, we also leave out of consideration a cataphoric reference (exemplified by (9)) and the so-called bridging anaphor.

- (9) “Vidím ho.” Velitel: “Oddělej ho.” Čečen se hroutí.  
“I see him.” (The) Commander: “Kill him.” (The) Chechen falls down.

Notational remark: The elements of the sentences referred to are printed in bold; if the anaphor is deleted on the surface and restored in the underlying structure of the sentence (i.e. if the node representing the anaphor has been reconstructed in the tectogrammatical representation), it is included in brackets and printed in capitals.

4.2 The following types of textual co-reference links are distinguished:

(i) A link to a particular node if this node represents the single antecedent of the anaphor; with a co-referential chain, all links (in the backward direction) are established, as in ex. (10); the link would lead from THEY to “them” and from there to “protestants” :

(10) Dohoda pochopitelně nic nevyřešila – pouze prohloubila v protestantech pocit, že je Londýn nechává na holičkách. Dnes tento pocit, že jsou (ONI) pro Británii pouze břemenem, s nímž si neví rady, v ulsterských protestantech pouze zesílil.

The agreement of course has not solved anything – it only deepened the feeling in the protestants that London leaves them in the lurch. Today this feeling, that (THEY) are only a burden for Great Britain they do not know how to deal with, has strengthened in Ulster protestants.

(ii) A link to the governing node of a sub-tree if the antecedent is represented by this node plus (some of) its dependents; this is also the way in which a link to a whole previous sentence (ex. (11)) or to a previous clause (ex. (12)) is established:

(11) Generál kromě toho připravuje nařízení, podle něhož se na něj budou moci obrátit všichni, kteří se domnívají, že se jim děje bezpráví. Hodlá **tím** předejít tomu, aby se redukce armády stala záminkou k vyřizování účtů.

The general also prepares an order, according to which all who think that harm is being done to them can turn to him. By **this** he intends to avoid a reduction of the army being a pretext for paying off old scores.

(12) Ale je něco jiného, když je někdo podnikatel a pak jde do politiky, anebo jestli někoho politické změny vynesou na špičku a on **toho** pak využívá k hospodářské činnosti a zastává vysoké funkce ve velkých firmách.

But it is a different thing when someone is an entrepreneur and then goes into politics than when political changes elevate somebody to the top and he then uses **this** in his economic activities and attains a high position in a big firm.

In (11), the pronoun form *tím* “(by) this” refers to the whole preceding sentence and the link is thus led to the root of the tree, i.e. to the main verb; in (12) the link points to the governing verb of the second conjunct, namely the verb *vynesou* ‘elevate’.

(c) A specifically marked link (SEGM for segment as one value of the attribute *corlemma*) denotes that the referent is a whole segment of (previous) text larger than one sentence (ex. (13)):

(13) Podle Kohla nelze zapomenout na to, že Německo přepadlo 22. června 1941 Sovětský svaz. Němci jménem Německa přivodili ruskému lidu nesmírné utrpení. Stejně tak nelze zapomenout, co Rusové způsobili Němcům. Z **toho** všeho si chceme vzít společné poučení.

According to Kohl it should not be forgotten that on June 22, 1941 Germany attacked the Soviet Union. Germans on behalf of Germany caused the Russians to suffer immensely. It also cannot be forgotten what the Russians did to Germans. From all **this** we should learn.

(d) A specifically marked link (EXOPH for exophor as one value of the attribute *corlemma*) denotes that the referent is ‘out’ of the co-text, it is known only from the situation; a rather clear instance of an exophor is in (14): one should know, if only from school history lessons, that the antecedent of the demonstrative pronoun is the Munich Treaty.

(14) V období vrcholícího léta roku 1939 již málokdo v Evropě mohl uvěřit nadějeplným slovům ... Chamberlaina, proneseným ... po návratu z Mnichova: Myslím, že je **to** mír na celou naši dobu.

In the height of the summer of 1939, only a few people could believe the hopeful words ... Chamberlain uttered ... after the return from Munich. I think that **this** is peace for our time.

(e) Cases of reference difficult to identify even if the situation is taken into account are marked by the assignment of *Unsp* as the lemma of the anaphor. This does not mean that a decision is to be made between two or more referents but that the reference cannot be specified precisely, even within a broader context. Thus e.g. in (15) it is not clear by what action the field has been prepared: by the minister's admission or by his opening the possibility? The difference, however, is not relevant. The co-referential link would lead to the nearest preceding possible antecedent, i.e. in case of (15) to the node representing the main verb of the preceding sentence *otevřel* 'opened'.

(15) Slovenský ministr kultury ... připustil, že zápůjčky obrazů nemusí být jednosměrné ... Otevřel tedy možnost, o které se dosud nemluvalo. Ředitelům obou galerií **tím** zároveň připravil pole, na němž si mohou vzájemně ustoupit ...

The Slovak minister of culture ... admitted that the loans of pictures need not be unidirectional. He thus opened a possibility which has not yet been discussed. By **this**, he prepared the field for the directors of both galleries so that they can make mutual concessions.

4.3 The annotation process has revealed several interesting phenomena concerning co-reference in Czech; many of them are rather complex cases in which a decision should not only be made on the co-referential links but also on the restoration of the nodes deleted in the surface shape of the sentence. For instance, the Czech verbs *říkat* [tell], *zapomenout* [forget], *ukázat* [show], *zapamatovat* [remember smth], *pochopit* [understand smth] have a semantically obligatory complementation of Patient (in other terminology, Objective); in the translation of the Chinese proverb in (16), the Patient of these verbs is deleted in the surface shape of the sentence and should be reconstructed in the TGTS with the lemma '*Unsp*' (unspecified) assigned to the reconstructed item in the first clause in each pair (except for the last pair where the demonstrative is not reconstructed but is present in the outer shape of the sentence), and with the lemma '*ten*' (that) in the second clause of the respective pair; this difference seems to be the most appropriate way to distinguish that the first reference is a general one, while the others are of a more "demonstrative kind", pointing to the first; the co-referential link would lead to the first occurrence.

(16) Každá kultura má svá rčení, která popisují zkušenosti lidstva s učením. České sděluje: Opakování, matka moudrosti. Čínské praví: Řekni mi (UNSP) a já zapomenu (TO); ukaž mi (UNSP) a já si (TO) zapamatuji; nech mne **to** dělat a já (TO) pochopím.

Every culture has its own sayings describing mankind's experience with learning. The Czechs say: Repetition is the mother of wisdom. The Chinese say: Say (it) and I will forget (that); show me (it) and I will remember (that); Let me do **it** and I will understand (that).

In many cases it is difficult to decide between an exophoric co-reference as a co-reference to an unspecified element somehow deducible from the preceding context as e.g. in (16), and a co-reference to a segment (perhaps of the "inferential" kind, see ex. (17)):

(16) Na churáňovských svazích se **to** zelená, běžkaři na kvildských pláních masově krouží na posledních zbytcích vlhkého sněhu.

On the hills of Churáňov (**it**) looks green, the cross-country skiers on Kvilda plains make big circles on the last remains of wet snow.

(17) Děkuji za sérii povídání o Osvětimi. Jsem rád, že se konečně píše o tom, jak **to** skutečně bylo.

Thanks for the series of writings about Auschwitz. I am glad that finally one writes about how **it** really was.

Special attention is to be paid to constructions which include a demonstrative pronoun and which represent phrasemes or "frozen" collocations. In those cases, no co-referential links are

established; actually the form ‘to’ (neuter form of the demonstrative ‘ten’) does not function as a pronoun here, see (18):

- (18) Nevím, **čím to je**, ale absolutně se mi tady nedaří.  
I do not know **what’s the matter**, but I am absolutely unsuccessful here.

## 5 OPEN QUESTIONS AND CONCLUSIONS

One of the advantages of a corpus-based study of a language phenomenon is that the researchers become aware of subtleties and nuances that are not apparent. Of course it is necessary, for those attempting a corpus annotation, to collect a list of open questions which have a temporary solution but which should be studied more intensively and in greater detail in the future. Also, there are whole fields that have not yet been systematically covered by any annotation scheme that we know of (PDT included), such as the formulation of a still deeper (‘logical’, ‘cognitive’ or other) layer of annotation, the interlinking of the underlying layer of annotation with the word sense disambiguation module, as well as the interlinking of the annotation of written texts and spoken discourses.

However, the annotation of the PDT has confirmed that a multi-layered annotation based on a solid and broadly tested theoretical framework makes it possible also to create reliable language resources for languages with a relatively “free” word order (i.e. order not determined grammatically) and with rich inflection (such as Czech). At the same time, a well-conceived annotation scenario helps to penetrate into the many details of language structure and thus to validate or, as the case may be, to complement the existing framework. This leads to the desired balance between the exactness of annotation and the requirements laid down by the very large amount of linguistic data available.

## ACKNOWLEDGEMENTS

The research reported on in this paper has been supported by the project of the Czech Ministry of Education LN00A063.

## REFERENCES

- HAIJČ, JAN (1998): Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning. Studies in honour of Jarmila Panevová (ed. by Eva Hajičová), Karolinum: Prague, 106-132.
- HAIJČOVÁ, EVA (1984): Presupposition and allegation revisited. *Journal of Pragmatics* 8: 155–167.
- HAIJČOVÁ, EVA (2000): Dependency-based Underlying-structure Tagging of a Very Large Corpus. In: *Les grammaires de dependence*, ed. Sylvain Kahane, Paris, 57-78.
- HAIJČOVÁ, EVA, HAVELKA, JIŘÍ, SGALL, PETR, VESELÁ, KATEŘINA AND DANIEL, ZEMAN (2004): Issues of Projectivity in the Prague Dependency Treebank. *Prague Bulletin of Mathematical Linguistics* 81, 5-23.
- HAIJČOVÁ, EVA, PARTEE, BARBARA H. AND SGALL, PETR (1998): *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer: Dordrecht.
- KUČOVÁ, LUCIE AND HAIJČOVÁ, EVA (2004): Prague Dependency Treebank: Enrichment of the Underlying Syntactic Annotation by Co-referential Mark-Up, *Prague Bulletin of Mathematical Linguistics* 81, 23-34.
- KUČOVÁ, LUCIE, KOLÁŘOVÁ, VERONIKA, PAJAS, PETR, ŽABOKRTSKÝ, ZDENĚK AND ČULO, OLIVER (2003): Anotování koreference v Pražském závislostním korpusu [Annotation of co-reference in the Prague Dependency Treebank]. Tech. Rep. of the Center for Computational Linguistics, Charles Univ., Prague.
- Prague Dependency Treebank: <http://ckl.mff.cuni.cz>

- SGALL, PETR, HAJIČOVÁ, EVA, AND PANEVOVÁ, JARMILA (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel: Dordrecht.
- SGALL, PETR, PANEVOVÁ, JARMILA AND HAJIČOVÁ, EVA (2004): *Deep Syntactic Annotation: Tectogrammatical Representation and Beyond*. In: *Proceedings of ACL 2004, Workshop Frontiers of Annotation*. Boston.
- TRED: <http://ckl.mff.cuni.cz/pajas/tred>
- USZKOREIT, HANS (2004): *New Chances for Deep Linguistic Processing*. In: Chu-Ren Huang and Winifried Lenders, eds. *Computational Linguistics and Beyond*. Academia Sinica: Taipei, 111-134.
- VESELÁ, KATEŘINA, NINO PETEREK, AND EVA HAJIČOVÁ (2003): *Topic-Focus Articulation in PDT: Prosodic Characteristics of Contrastive Topic*. *Prague Bulletin of Mathematical Linguistics*, 79-80: 5-22.
- VESELÁ, KATEŘINA, AND JIŘÍ HAVELKA (2003): *Anotování aktuálního členění věty v Pražském závislostním korpusu (Annotation of TFA in Prague Dependency Treebank)*. ÚFAL/CKL Technical Report TR-2003-20.
- VESELÁ, KATEŘINA, HAVELKA, JIŘÍ AND EVA HAJIČOVÁ (2004): *Annotators' Agreement: The Case of Topic-Focus Articulation (2004)*. In: *Proceedings of Int. Conf. on Language, Resources and Evaluation (LREC 2004)*, Vol. VI, ELRA, Paris, 2191-2194.

## RESUMÉ

V současné době je v korpusové lingvistice aktuálním tématem zachycení hloubkové struktury věty (viz např. pozvaná přednáška prof. H. Uszkoreita na COLINGu 2002 v Taipei). Koncepce Pražského závislostního korpusu (dále jen PDT) s takovou úrovní anotace od počátku počítala (projekt vznikl v letech 1995-6). Tektogramatická (hloubková) úroveň anotace vychází z Funkčního generativního popisu navrženého Petrem Sgallem v 60. letech a dále rozpracovávaná pražskou univerzitní skupinou teoretické a počítačové lingvistiky. V našem příspěvku se soustředíme především na anotaci informační struktury věty (aktuálního členění), na zachycení základních aspektů gramatické a textové koreference. V závěru se zamýšlíme nad úskalími anotačního procesu (především pokud jde o shodu anotátorů), konstatujeme však, že hloubková anotace jazykových korpusů je nesmírně přínosná jak pro lingvistickou teorii, tak pro nejrůznější účely aplikační.



# Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank\*

MARKÉTA LOPATKOVÁ, JARMILA PANEVOVÁ

## 1 THE FRAMEWORK

The Functional Generative Description (FGD, see Sgall, 1967, Sgall et al., 1986) was applied as a general framework for the development of the valency theory (see Panevová, 1974-75, 1980, 1994) as well as for the design of the Czech syntactically annotated corpus (PDT, see Hajič, 1998, Hajičová et al., 2001).

Valency is understood as a lexico-syntactic attribute of a word – more precisely, of a particular lexical sense of the lemma, called here *lexis* (“*lexie*” in Czech terminology, see Filipec and Čermák, 1985). More precisely, we can understand a *lexis* as a pair formed by a lexical unit and one of its meanings.<sup>1</sup> A valency frame (VF) is assigned to every auto-semantic lexical unit (*lexis*). This, however, may be empty, e.g. with the Czech verb *pršet* [*to rain*], with nouns such as *stůl* [*the table*], adjectives as *hezký* [*beautiful*]. The labels used for the valency slots belong to the underlying structure (tectogrammatrics) and, together with the lexical unit (*lexis*), they constitute a tectogrammatical representation of the lexical entry. With regard to the applied tasks, we include the morphemic counterparts of the particular valency slots as a part of the (complex) frame of the given unit.

Valency is prototypically connected with verbs. We have distinguished two main classes of verbal complements:

- (i) inner participants, IP in the sequel (ACT(or), PAT(ient), ADDR(essee), ORIG(in) and EFF(ect)),
- (ii) free modifications, FM in the sequel.

The criteria for the distinction between these two classes are given in Panevová (quoted above).

Valency frames of *lexes* are constituted by their respective inner participants (either obligatory or optional) and by their obligatory free modifications.<sup>2</sup>

---

\* The work reported on in this paper has been carried out under the project of “Centers of Excellence” supported by MŠMT, grant No LN00A063. It has been partly supported from the grant GAČR No 405/04/0243.

<sup>1</sup> The formal representation of *lexis* in FGD has not yet been specified. The surface shape (lemma) of the lexical item is used instead (with a differentiating subscript, if necessary).

<sup>2</sup> We prefer this terminology rather than the terminology used in Daneš et al., 1981 and “Mluvnice češtiny 3”, 1987. There the term “potenciální” (potential) is used for optional as well as for obligatory positions of VF omitted on the surface. Moreover, the difference between the VF as a part of lexicon and its application for the concrete utterance is not reflected in the terminology common in Czech handbooks.



We share Tesnière's (1959) approach as to the one-argument and two-argument verbs: the first slot is structured as ACT(or) (though it corresponds to different semantic (ontological) roles, such as Bearer, Processor, Stimulus etc.); with two-argument verbs the inner participants are structured as ACT(or) and PAT(ient). The relation between the syntactic arguments and their cognitive roles is called a "shifting of participants", see Panevová, 1980. If the verb has three (or more) valency slots, the semantics of them is taken into account. This strategy agrees with the theory of case meanings, distinguishing between syntactic (grammatical) cases and semantic (concrete) cases (see Kuryłowicz, 1949): the valency slots of ACT and PAT are occupied mostly by syntactic cases (Nominative and Accusative, respectively), while the other participants and free modifications are expressed mostly by cases with concrete (semantic) meanings.

## 2 AN INTRODUCTION OF QUASI-VALENCY COMPLEMENTS

In section 1 we briefly summarized the basic features of our valency theory of verbs. However, in the course of empirical studies of material, especially in connection with the building of the valency lexicon of verbs VALLEX (see Lopatková, Žabokrtský, 2003 and section 5 below) and with a tectogrammatical annotation of PDT (see Uřešová, this volume), some unresolved problems appeared. Firstly, it was necessary to introduce some additional functors (types of syntactic-semantic relations) for newly discovered semantically relevant distinctions (namely OBST(acle) and MED(iator)). In analyzing their semantic and syntactic distribution, we observed that they share partly the features of inner participants, and partly the features of free modifications. Secondly, revisiting the list of verbal complements introduced earlier, we discovered that some complements (namely DIFF(erence) and INT(ent)) also share important features of inner participants (see (i), (ii) and (iii)), although they also have some of the characteristic features of free modifications (see (iv), (v) and (vi)):

- (i) they are governed (their morphemic shape is determined) by their verbal heads
- (ii) they occur with a limited class of verbs
- (iii) they cannot be repeated,  
however
- (iv) as to their meaning, they are semantically homogeneous
- (v) they do not underlie the "shifting"
- (vi) they are mostly optional.

We also reconsidered the complements ADDR, ORIG (and perhaps EFF) from this point of view. The complements ADDR and ORIG undoubtedly fulfill (i), (ii), (iii) characteristics for IP, but also (iv),<sup>3</sup> which is typical of FM; they do not meet (v) and (vi). The features of EFF shared with quasi-valency complements are limited; (i), (ii) and (iii) are present in EFF, but one of the most important quasi-valency features (iv) is missing here. This is the main reason why we still classify EFF as an inner participant. However, we are still undecided as to whether the ADDR and ORIG should not be classified as quasi-valency complements, too.

### 2.1 OBSTACLE

The meaning of OBST(acle) is expressed in Czech by the prepositional group *o* + Accusative with verbs like *zakopnout* [to stumble], *uhodit se* [to strike oneself], *bouchnout se* [to bump oneself], *zranit se* [to injure oneself], *píchnout se* [to prick oneself], *bodnout se* [to prick oneself].

<sup>3</sup> This statement is valid at least for verbal valency features. As for nouns, see Section 4 below.

Their form is governed by their head verbs. In handbooks on Czech syntax they are classified as Means (Instrument), but they undoubtedly have a special instrumental semantics, see (1), (2) and (3):

- (1) Jan zakopl nohou o stůl  
[John stumbled over the table with his leg]
- (2) Matka se píchla nůžkami  
[Mother pricked herself with the scissors]
- (3) Růženka se píchla o trn  
[Sleeping Beauty pricked herself on a thorn]

In (1) *noha* [leg] is a proper means (Instrument), while the construction *o stůl* [about the table] is not. In (2) *nůžky* [scissors] refers to a device used as an Instrument proper, its semantics includes the semantics of movement with this instrument. In (2) the manipulation with scissors is presumed, while in (3) the noun *trn* [thorn] (with an instrumental semantics) is fixed (see also Apresjan, 2001). The feature of an unconscious action is typical of (3), while in (2) the action can be either conscious or unconscious. For the semantics of “fixed” Instrument (expressed by the prepositional group *o* + *Accusative*) the new label **Obstacle** was proposed (initially in Panevová, 2003). All the verbs listed in this sample imply their unconsciousness. The verbal modification of **Obstacle** shares the features of the group of inner participants (i), (ii) and (iii), but also all the features listed above as free modification attributes (iv), (v), and (vi)<sup>4</sup>.

## 2.2 MEDIATOR

Also, the Czech prepositional group *za* + *Accusative* is described in syntactic handbooks as a kind of Instrument, see e.g. (4), (5), (6):

- (4) Otec přitáhl kluka levou rukou za ucho  
[Father has drawn boy's ear by his left hand]
- (5) Když jsem odcházel, zatahal mě soused za rukáv  
[When I was leaving, the neighbor pulled my sleeve]
- (6) Jan přivedl psa za obojek  
[John brought the dog by its collar]

Examples (4) to (6) demonstrate that the semantics of this prepositional group is different from the pure Instrument. Pure Instrument is usually used by the Actor of the action directly, while in (4) to (6) the instrument is a part of another entity (the ear belongs to the boy in (4) and as a part of a boy it is used for drawing the boy). In (4) the Instrument proper is present (*ruka* [hand]). The Actor uses his own hand as a means to reach the boy, and he uses the boy's ear as a **Mediator** for reaching him. Like the Obstacle, the Mediator shares some features of IP and some of the class of FM. Unlike the Obstacle, we have not yet found any verb with an obligatory Mediator.

## 2.3 DIFFERENCE

The prepositional group *o* + *Accusative*, although it mostly combines with the comparatives of adjectives or adverbs, can also occur with some verbs (see e.g. (7), (8), (9) for verbs, (10) for an adverb):

<sup>4</sup> Feature (vi) has some exceptions: we have found the verbs *zavadit* [to touch], *(za)chytit (o něco)* [to get caught (on st)] with obligatory OBST.

- (7) Inflace se zvýšila proti roku 2000 o několik procent.  
[The inflation has increased in comparison with 2000 by several percent]
- (8) Náš tým zvítězil o dvě branky  
[Our team won by two goals]
- (9) Jan zvítězil v závodě o prsa  
[John won the race by a hair's breadth]
- (10) Postupte o dva schody výš  
[Move two steps higher]

The modification of **DIFF(erence)** can be characterized as a kind of extent, but while the general extent expresses nothing more than a high or low degree, the modification of DIFF specifies the extent more precisely. At least two entities are compared here, although one of them is more or less implicit (inflation in the current year and in 2000 are compared in (7), the score of a match of two teams are compared in (8), John's rivals are understood in (9) as the other entity) and the difference between them is explicitly expressed by the Difference modification.

#### 2.4 INTENT

The modification of **INT(ent)** is compatible mainly with the verbs of motion and it differs from the FM of AIM: an actor of the INT is identical with the person that provides the intended action himself/herself (the action can be transformed into a nominalization, see e.g. (12), contrary to (13), where the FM of AIM is expressed). The actor (mother in the case of (13)) only transfers potatoes from one place to another. The difference between INT and AIM could be exemplified by the acceptability of (14a) and unacceptability of (14b).<sup>5</sup>

- (11) Jan se šel koupat  
[John went to swim]
- (12) Helena šla na jahody  
[Helen went (to pick) strawberries / *lit.* Helen went on strawberries]
- (13) Matka šla do sklepa pro brambory  
[Mother went to the cellar for potatoes]
- (14a) Helena šla do krámu pro jahody  
[Helen went to the shop for strawberries]
- (14b) \*Helena šla do krámu na jahody  
[\*Helen went to the shop (to pick up) strawberries / *lit.* Helen went to the shop on strawberries]

### 3 VALENCY OF ADJECTIVES

Our analysis of adjective valency was aimed at the verification of two hypotheses:

- (i) that the valency slots of adjectives share the roles of verbal complements;
- (ii) that the shifting of participants is here valid in the same manner as with verbs (with one natural exception: one of the valency slots is absorbed by the governing noun in

<sup>5</sup> The introduction of the INT complement is supported by the findings presented in Poldauf, 1959. The prototypical expression of an INT is an infinitive; unprototypically, the prepositional expression is used (see (12)); it implies the active participation of the actor in collecting strawberries. This is the reason why (14b) is meaningless (at least in our actual world), somebody else (other than Helen) has collected the strawberries and delivered them to the shop.

noun phrases or by the subject position in the clauses with the copula *být* [*to be*] so it is excluded from the valency frame of the respective adjective).

In the case of primary adjectives, the position of ACT is absorbed; with deverbal adjectives the absorbed position depends on the type of derivation (with active participles the position of ACT is absorbed as well, with passive participles PAT, ADDR or EFF is absorbed, for details see Panevová, 1998).

Otherwise, the deverbal adjectives share the valency of their source verbs.

The question of the lexical ambiguity of adjectives used for human qualities remains open. This consideration concerns such adjectives as *hrdý* [*proud*], *věrný* [*faithful*] etc. They are used either as the “absolute” attribute of a noun (and they have an empty valency frame), or they are used as relative adjectives with an obligatory PAT (*hrdý na* + Acc, *věrný* + Dat). We have also considered an alternative solution, where we have to deal with a single lexical sense for absolute and relative usage and where the optional PAT enters their valency frame (for more examples, see Panevová, 1998 and Panevová, in prep.).

#### 4 VALENCY OF NOUNS

The set of valency complements of nouns was extended, as proposed by Piřha, 1981, if compared with the set of valency complements of verbs. We have accepted his proposal as to the complements called there **MAT(erial)** (as an obligatory or an optional noun participant) and **APP(ur)tance** (as a free noun modification, obligatory with the listed nouns). We have reconsidered his proposal to classify **ID(entity)** as an optional participant of a noun; it should belong to the class of FM, because any noun can have its name (not only *lod' Titanic* [*boat Titanic*], but also *tuřka Koh-i-nor* [*pencil Koh-i-nor*], *souprava Julie* [*set Julia*]).

In the valency frame of many nouns, the same complements occur as in the VF of verbs. This is obvious for deverbal nouns (for details see Novotný, 1980, Karlík, 2000, Panevová, 2000 and esp. Řezníčková-Kolářová, 2003, Kolářová, in prep.). Moreover, the complements (functors) typical of verbs are compatible with a high number of primary nouns (e.g. PAT in *názor na něco* [*opinion on*], *příklad na něco/něčeho* [*example for*], *kniha o něčem* [*book on*], ADDR in *dárek někomu* [*gift to*], ORIG in *daň z pozemku* [*tax for*]). In the last two cases, we again perhaps have to do with the absorption of one participant built within the head noun (*dárek* and *daň* are patients themselves, a gift is what was given, tax is what is paid).

The functor called ORIG(in) has a special position among noun complements. Although it has its counterpart within verbal inner participants, with nouns it typically behaves as a free modification: it is compatible with any primary noun and it can be repeated (*šaty ze lnu od starší sestry* [*a dress from linen from my elder sister*], *nábytek ze dřeva od našeho hlavního dodavatele* [*furniture from wood from our main provider*]). The interpretation of the inanimate noun expressing an Origin is material, while an animate name (and its equivalents as the names of institutions, human collectives etc.) corresponds to the source. A re-classification of Origin as a FM noun complement – proposed here for the first time within our framework – is based on its syntactic behaviour with nouns (different from its behaviour with verbs, where it cannot be repeated and it is not compatible with every verb).

#### 5 THE BUILDING OF A VALENCY LEXICON BASED ON THE THEORY DESCRIBED

A description of valency is impossible without a good syntactically based framework, and – since valency differs from one lexical item to another – it cannot be described by general rules. Therefore a valency lexicon belongs among the basic language resources indispensable

for any rules-based task of NLP (Natural Language Processing). Here we refer to the valency lexicon VALLEX, which has been created in connection with the annotation of PDT.<sup>6</sup>

The Valency Lexicon of Czech Verbs, Version 1.0 (VALLEX 1.0, <http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/>) is a collection of linguistically annotated data and documentation, resulting from the attempt at formal description of the valency frames of Czech verbs. VALLEX 1.0 contains roughly 1400 verbs in all their senses (app. 4000 frame entries / senses). VALLEX is designed both for human readers and for application tasks in NLP as e.g. machine translation or information retrieval.

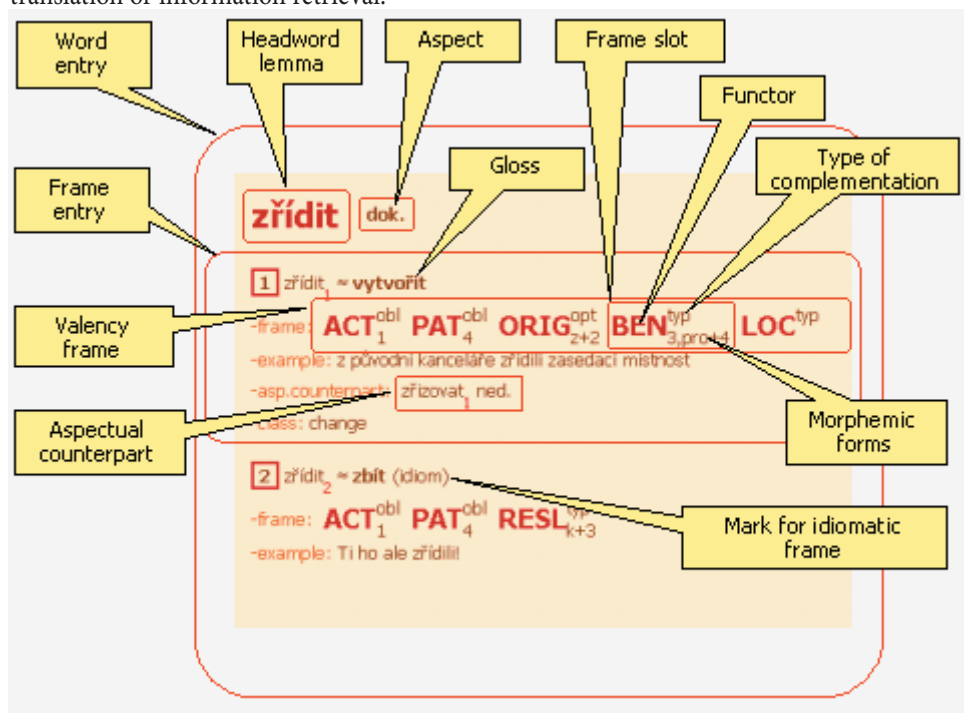


Figure 1: Word entry in VALLEX

A Czech verb as a whole, a verb lexeme (**word entry** in VALLEX) is an abstract unit made up by all the senses of a particular verb. A word entry consists of a (non-empty) sequence of **frame entries**, each of which corresponds to a single sense (“lexis”, see above). Each frame entry describes the valency frame itself, the specification of a sense in question (by gloss(es) and example(s)), and additional information (as e.g. aspect, type of reflexivity, control, (preliminary) semantic class). A **valency frame** itself is a sequence of **frame slots** corresponding to (either required or specifically permitted) complements of a given verb. Each valency slot is characterized by its **functor**, i.e. the name of the syntactic-semantic

<sup>6</sup> Besides VALLEX, a larger valency lexicon (called PDT-VALLEX, see e.g. Hajič et al., 2003, Urešová, this volume) has been created during the annotation of PDT. PDT-VALLEX contains more verbs (5200 verbs), but with only those of their senses that occurred in PDT, whereas in VALLEX the verbs are analyzed in their full complexity, in all their senses. In addition, richer information is assigned to particular valency frames in VALLEX, and stress is laid on the consistency and completeness of annotation.

relation (labels of underlying roles), and the possible morphemic form(s) (specification of morphemic case, prepositional group, infinitive or subordinated verbal construction).

A word entry in VALLEX corresponds to the whole lexeme; it consists of a (non-empty) sequence of frame entries corresponding to a single sense.

We have formulated the following principles and functional criteria for distinguishing particular senses adopted that are connected with their valency. The principles can be characterized by two statements:

- A. any change in valency frame (either in functor, in the combination of functors, or possible form(s) of functor) justifies an introduction of a new frame entry;
- B. any significant change in sense justifies the introduction of a new frame entry.

These fundamental principles imply the following rules.

(i) The difference in the sense is a necessary but not sufficient condition for a postulation of two (or more) valency frames – a (slight) difference in the sense is ignored if lexical units do not differ syntactically.

- (15) *hýbat*<sub>1</sub> [to move]<sup>7</sup> ... ACT(1;obl) PAT(Instr,s+Instr;obl)  
 hýbat rukou; hýbat (s) křeslem  
 [to move (with) sb's hand, to move an armchair]

In Czech lexicons “Slovník spisovného jazyka českého” [The dictionary of Standard Czech] (1964) as well as in “Slovesa pro praxi” [Verbs for Practice] (1997) two distinct senses are distinguished – “uvádět něco v pohyb, pohybovat” [to set st in movement, to move st] and “měnit polohu” [to change position (of st)]. In VALLEX, these two usages of the verb *hýbat* in (15) are described in a single valency frame – the difference in the senses is not taken into account, their syntactic behaviour being the same. The decision to ignore this type of difference is based on the fact that such a “fine-grained” distinction of senses is not reflected in the syntactic behaviour of the given lexical units and they are often not perceived, even by a human reader in real texts.

(ii) Two different senses can have an identical valency frame.

- (16a) *chovat*<sub>1</sub> [to cradle] ... ACT (1;obl) PAT(4;obl)  
 chovat dítě (v náručí)  
 [to cradle a child (in one's arms)]  
 (16b) *chovat*<sub>2</sub> [to keep] ... ACT (1;obl) PAT(4;obl)  
 chovat prasata (na farmě)  
 [to keep pigs (on a farm)]

The indisputable different senses of the verb *chovat* have the same valency frame consisting of two inner participants, Actor and Patient with the same morphemic forms; however, the difference of the sense has to be reflected by distinguishing two different frame entries in VALLEX.

(iii) The change in morphemic realization signalizes the possibility of different senses.

- (17a) *hlásit se*<sub>2</sub> [to be counted among sb] ... ACT(1;obl) PAT(k+3;obl)  
 hlásit se ke komunistům

<sup>7</sup> The lower numeral index attached to the lemma denotes a particular frame entry in VALLEX notation.

- [to be counted among communists]  
 (17b) *hlásit se*<sub>4</sub> [to apply for st] ... ACT(1;obl) PAT(o+4;obl)  
*hlásit se o svá práva*  
 [to apply for own rights]

The change in morphemic realization signalizes different senses and thus two lexical items *hlásit se*<sub>2</sub> and *hlásit se*<sub>4</sub> are distinguished.

(iv) On the other hand, a particular complement in a valency frame can have morphemic variants (if they differ stylistically, rather than in their semantics).

- (18) *učit*<sub>1</sub> [to teach] ... ACT(1;obl) ADDR(4;obl) PAT(3,4,inf,že,zda,aby,jak;obl)  
*Učitel učí žáky matematice / matematiku / pracovat / ...*  
 [Teacher teaches his pupils mathematics<sub>Dat</sub> / mathematics<sub>Acc</sub> / to work / ...]

With this lexical unit there is more than a single possibility to express the obligatory Patient.

(v) A change in valency frame is connected with a change of sense – two valency frames cannot share their senses.

- (19a) *postavit*<sub>1</sub> [to raise] ... ACT(1;obl) PAT(4;obl)  
*postavit sloup*  
 [to raise a column]  
 (19b) *postavit*<sub>2</sub> [to build] ... ACT(1;obl) PAT(4;obl) ORIG(z+2;opt)  
*postavit budovu; postavit model letadla z balzy*  
 [to build up a building; to construct a model of a plane from balsa wood]  
 (20a) *poslat*<sub>1</sub> [to send] ... ACT(1;obl) ADDR(3;obl) PAT(4;obl)  
*poslat matce dárek k narozeninám.*  
 [to send sb's mother a birthday gift]  
 (20b) *poslat*<sub>2</sub> [to send] ... ACT(1;obl) PAT(4;obl) DIR3(obl)  
*poslat zásilku do Konga*  
 [to send a consignment to Congo]

The valency frames in (19a) and (19b) differ in the presence of an optional inner participant ORIG(in) – *postavit*<sub>1</sub> [to raise] cannot be modified by this complement. This distinction entails a clear distinction in the senses of *postavit*<sub>1</sub> and *postavit*<sub>2</sub> (reflected also by different translation equivalents, *to raise* and *to build*).

With some groups of verbs this principle is not obvious at first sight – they have two valency frames and their sense is rather close, e.g. *poslat* in (20a) and (20b). However, the detailed analysis of syntactic and semantic properties of some of these groups given in Benešová, 2004 shows clear syntactic and semantic distinctions in sense between them.

(vi) Different valency frames can reflect a primary and a secondary (figurative) usage of a given verb.

- (20a) *dopadnout*<sub>1</sub> [to fall (down)] ... ACT(1;obl) DIR3(obl)  
*dopadnout na zem*  
 [to fall down to the ground]  
 (20b) *dopadnout*<sub>2</sub> [to strike] ... ACT(1;obl) PAT(na+4;obl)  
*Dopadly na ně starosti.*  
 [Troubles have fallen on them]



Directionality proper and directionality in a metaphorical sense are met in (20a) and (20b). Despite the same morphemic realizations, different functors, namely DIR3 (direction – to where) and PAT, are assigned to the second complement. This distinction is justified by different syntactic-semantic features (*dopadnout<sub>1</sub>* belongs to the “verbs of motion”, unlike *dopadnout<sub>2</sub>*).

Distinguishing the particular senses of a single verb lexeme is amongst the most complicated problems in the domain of constructing a lexicon. We have tried to discuss and exemplify the criteria connected with the valency behaviour of verbs.

## 6 CONCLUSION

The Czech data analyzed during the development of the PDT present some new issues not yet solved within the theoretical background. In confronting these issues, we have made some modifications in the framework: we have introduced new types of functors (syntactic-semantic relations) and we have shifted some functors into another class of valency complements. We have presented here several examples illustrating the methodology used in building up the valency lexicon (VALLEX 1.0). The relations between the lexical meanings of verbal units and their valency frames are illustrated in Section 5. We can conclude, however, that the changes to the framework resulting from the annotation of relatively large data are not substantial, although they have brought some refinements of the theory of FGD.

## REFERENCES

- APRESJAN, J. D. (2001): Znachenije i upotreblenije. Voprosy jazykoznanija 4, pp. 3-22.
- BENEŠOVÁ, V. (2004): Delimitace lexii českých sloves z hlediska jejich syntaktických vlastností. Diplomová práce, FFUK, Praha.
- DANEŠ, F., HLAVSA, Z. a kol. (1981): Větné vzorce v češtině. Academia, Praha.
- FILIPEC, J., ČERMÁK, F. (1985): Česká lexikologie. Academia, Praha.
- HAIJČ, J. (1998): Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. E. Hajičová), Karolinum, Charles University Press, Prague, pp. 106-132.
- HAIJČ, J., PANEVOVÁ, J., UREŠOVÁ, Z., BÉMOVÁ, A., KOLÁŘOVÁ, V., PAJAS, P. (2003): PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pp. 57-68. Vaxjo University Press.
- HAIJČOVÁ, E., PANEVOVÁ, J., SGALL, P. (2001): Manuál pro tektogramatické značkování. Verze IV. Technická zpráva, ÚFAL MFF UK.
- KARLÍK, P. (2000): Valence substantiv v modifikované valenční teorii. In: Čeština – univerzália a specifika 2, Sborník z konference ve Šlapanicích u Brna, Masarykova Univerzita, Brno, pp. 181-192.
- KOLÁŘOVÁ, V. (in prep.): Valence deverbativních substantiv v češtině. (Manuscript of PhD thesis).
- KURYŁOWICZ, J. (1949): Le problème du classement des cas. Biuletyn Polskiego Towarzystwa Językoznawczego, Vol. 9, pp. 20-43.
- LOPATKOVÁ, M., ŽABOKRTSKÝ, Z., SKWARSKA, K., BENEŠOVÁ, V. (2003): Valency Lexicon of Czech Verbs VALLEX 1.0. CKL/UFAL Technical Report TR-2003-18, 2003.
- Mluvnice češtiny 3, Skladba (1987). Akademie, Praha.
- NOVOTNÝ, J. (1980): Valence dějových substantiv v češtině. In: Sborník Pedagogické fakulty v Ústí n. Labem, Praha.
- PANEVOVÁ, J. (1974-75): On Verbal Frames in Functional Generative Description. Part I, The Prague Bulletin of Mathematical Linguistics 22, pp 3-40, Part II, The Prague Bulletin of Mathematical Linguistics 23, pp. 17-52.
- PANEVOVÁ, J. (1980): Formy a funkce ve stavbě české věty. Academia, Praha.

- PANEVOVÁ, J. (1994): Valency Frames and the Meaning of the Sentence. In: *The Prague School of Structural and Functional Linguistics* (ed. Ph. L. Luelsdorff), Amsterdam-Philadelphia, John Benjamins, pp. 223-243.
- PANEVOVÁ, J. (1998): Ještě k teorii valence. In: *Slovo a slovesnost* 59, pp.1-14.
- PANEVOVÁ, J. (2000): Poznámky k valenci podstatných jmen. In: *Čeština – univerzália a specifika* 2, Sborník z konference ve Šlapanicích u Brna, Masarykova Univerzita, Brno, pp. 173-180.
- PANEVOVÁ, J. (2003): Some Issues of Syntax and Semantics of Verbal Modifications. In: *Proceedings MTT 2003, First International Conference on Meaning-Text Theory*, pp. 139-146. Ecole Normale Supérieure.
- PANEVOVÁ, J. (in prep.): Valence vybraných českých adjektiv ve světle ČNK. *Slavistična revija*.
- PIŤHA, P. (1981): On the Case Frames of Nouns. *Prague Studies in Mathematical Linguistics* 7, Academia, Prague, pp. 215-224.
- POLDAUF, I. (1959): Děj v infinitivu. In: *Slovo a slovesnost* 20.
- ŘEZNÍČKOVÁ-KOLÁŘOVÁ, V. (2003): Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora. In: *Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003)*, pp. 88-97. UCREL, Lancaster University.
- SGALL, P. (1967): *Generativní popis jazyka a česká deklinace*. Academia, Praha.
- SGALL, P., Hajičová, E., Panevová, J. (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects* (ed. by J. Mey), Dordrecht: Reidel and Prague: Academia.
- Slovník spisovného jazyka českého* (1964). Praha.
- Slovesa pro praxi* (1997): Svozilová, N., Prouzová, H., Jirsová, A. (autoři), Academia, Praha.
- UREŠOVÁ, Z. (this volume): The verbal valency in the Prague Dependency Treebank.
- TESNIÈRE, L. (1959): *Eléments de syntaxe structurale*. Paříž.



# Verbal Valency in the Prague Dependency Treebank from the Annotator's Viewpoint

ZDEŇKA UREŠOVÁ

## 1 THE CONCEPT OF VALENCY IN PDT

One of the prerequisites of the correct syntactic annotation at the tectogrammatical level (TR) of the Prague Dependency Treebank (see Hajič, this volume) is the knowledge of *valency frames*. The valency theory (see Panevová, 1974-75, 1980, 1994, 1999) as used in the process of annotation of the Prague Dependency Treebank (PDT) corresponds to the concepts of the Functional Generative Description (FGP) (see Sgall, 1967, Sgall et al., 1986). Within this approach, syntactic as well as semantic criteria are used to identify verbal complementations.

The verb is considered to be the core of the sentence (or clause, as the case may be). Its complementations (*dependents*) are classified either as *inner participants* or as *free modifications*. Both types of verbal complementations can be either *obligatory* (semantically always present with a given verb) or *optional* (not necessarily present). Only inner participants (obligatory or optional) and obligatory free modifications belong to the verbal valency frame. Optional free modifications are not listed in the valency frame.<sup>1,2,3</sup> The relation between the dependent and its *governor* at the TR is labelled by a *functor*. The functor must be determined and recorded for all complementations in the actual process of data annotation. Annotators choose<sup>4</sup> this value from a set of functors listed in the manual for tectogrammatical annotation (see Hajičová et al., in prep.) The intersection of the set of functors used for valency modifications and non-valency modifications is not empty. Had we also annotated in the corpus which verbal complementations are obligatory and which are optional, we could then have simply extracted the valency frames of all verbs from the annotated sentences. However, in order to obtain the highest possible mutual agreement among the annotators and to maintain consistency in the course of annotation, such a lexicon is already being built up step-by-step *during* the annotation. It is shared among the annotators and it is gradually being enlarged. This lexicon is called **PDT-VALLEX**.

<sup>1</sup> Neither are the so-called quasi-valency and typical complementations stored in the valency frames of the PDT-VALLEX lexicon (these types of complementations are described by Lopatková et al. (2003), Panevová (2003)).

<sup>2</sup> We discuss here the verbal valency frame in a narrow, strict sense, i.e. the verbal valency frame captured in the lexicon. The verbal valency frame in a broader sense consists of all the complementations which can expand the given verb. The types of all the complementations are captured in the structure of the annotated tree as some of the values of the dependent nodes.

<sup>3</sup> Valency is also considered for many nouns and adjectives, see Řezníčková, V. (2003), Hajič et al. (2003).

<sup>4</sup> If the annotators hesitate about the correct value of the functor, they have the choice of marking this uncertainty through the multiple selection of several functors.

## 2 THE CONCEPT OF THE VALENCY FRAME

Taking the basic principles (see Panevová, 1974-75 and the writings quoted above) as a starting point, we use the criteria for **distinguishing inner participants and free modifications, the concept of shifting** of “cognitive roles” and **the dialogue test** for determining the obligatoriness of inner participants and free modifications.

### 2.1 DISTINCTION BETWEEN INNER PARTICIPANTS AND FREE MODIFICATIONS

This distinction applies to the set of complementation types (functors) as a whole (i.e., if a functor is classified as an inner participant, it will be called an inner participant in any valency frame (of any verb) in which it appears).

If a complementation type modifies the verb only once in any given clause (without regard to possible coordination or apposition) and it occurs just with particular verbs, which can, in principle, be listed, we call it an *inner participant*. Five inner participants are distinguished at the tectogrammatical level in the PDT. (For a detailed discussion about the position of ADDR, ORIG and EFF, see Lopatková, Panevová (this volume); the ideas proposed there have not yet been taken into account in PDT-VALLEX):

- ACT (Actor),
- PAT (Patient),
- ADDR (Addressee),
- ORIG (Origin) and
- EFF (Effect).

Inner participants are determined semantically, except for the Actor (ACT) and (to a certain extent) also the Patient (PAT). The first participant is always the Actor; the second one is always the Patient. Addressee (ADDR) is the semantic counterpart of an indirect object. As a rule, ADDR is animate (*promise something to somebody.ADDR, talk to somebody.ADDR about something, teach someone.ADDR something*). Effect (EFF) is the semantic counterpart of the second object or of the verbal attribute (*break something into something.EFF, appoint somebody as somebody.EFF*). Origin (ORIG) also comes from the second (or third or fourth) object, describing origin or something that is being transformed by the verb into something else (*create something from something. ORIG, translate something (a book) from Czech. ORIG to English, expect something from somebody. ORIG*).

On the other hand, if the same type of verbal complementation can be repeated within the same clause and if it can modify any verb (in principle), we call it a *free modification*. There are approximately 50 distinct free modifications used at the TR. The list comprises modifications of different kinds, such as local and directional (LOC, DIR1, DIR2, DIR3), temporal (TWHEN, TSIN, TTILL, TFL, TFHL, THO, TPAR, TFRWH, TOWH), manner (MANN), intention (INTT) or causal (CAUS), etc. The full list of all functors is given in the annotation manual (see Hajičová et al., in prep.).

A secondary criterion for distinguishing the difference between an inner participant and a free modification is *government* (“*rection*”). If the form of the dependent (morphological case, preposition, particular lexeme to be used etc.) is determined by the governing verb, it is considered to be an inner participant; if the dependent is independent in its form of the governing verb, then it is considered to be a free modification.

Inner participants are subject to *shifting* but free modifications are not.

## 2.2 THE CONCEPT OF SHIFTING OF “COGNITIVE ROLES”

If the valency slots for Actor and Patient are not occupied (for the verb in question), *shifting* of participants takes place. The principle of shifting requires that if a verb has only one inner participant, it is always the Actor and if there are two inner participants of the verb, they are always the Actor and the Patient, regardless of their “semantics”. In the case of three or more inner participants of one particular verb, the first two are always Actor and Patient; for other than the first two slots, the above more or less semantic criteria are taken into account. For instance:

PAT shifted to the position of ACT: *The book.PAT came out*;

ADDR shifted to the position of PAT: *she understood him.PAT*;

ADDR shifted to the position of PAT, EFF stays in its slot: *elect him.PAT as a chairman.EFF*;

EFF shifted to the position of PAT: *to build a group.PAT*;

ORIG shifted to the position of PAT: *to act on the basis of a presupposition.PAT*.

There are only some specific cases where shifting does not apply (see 4.10)

## 2.3 THE DIALOGUE TEST

Inner participants and free modifications can (at the tectogrammatical level) be either obligatory or optional. The semantically obligatory dependent does not have to be present at the syntactic (analytical) level (AR); it can be omitted without the sentence becoming ungrammatical. However, the annotator has to restore this node in the tectogrammatical tree that represents the sentence at the TR. The obligatory complementations are thus always present at the TR despite their omission at the analytical level, which might even be the correct or preferred case: e.g., (pronominal) Actor is always an omissible member in the surface structure of a Czech sentence (Czech is a pro-drop language). (Some obligatory free modifications are also, in general, omissible in the surface realization: for example, in short answers to questions.) Therefore, the semantic obligatoriness cannot be determined by the surface form; but it can be examined by the *dialogue test* (see Panevová, 1999): the answer “I don’t know” is not acceptable (it would disturb the smoothness of the dialogue) if the complementation is semantically obligatory. For instance, the functors DIR3 (directional – where to) with the verb *to come* and DIR2 (directional – from where) with the verb *to leave* are obligatory. As long as the answer “I don’t know” is acceptable without disturbing the smoothness of the dialogue, we speak about an optional complementation (again, we mean optional in the Tectogrammatical Representation). For instance, the functors DIR2 (directional – from where) with the verb *to come* and CAUS (cause) with the verb *to leave* are optional.

It should be pointed out that the application of the dialogue test was, largely, very helpful but for some verbs it merits further discussion. Unfortunately, there was no time during the process of annotation to construct special semantically related groups of verbs (see Levin, 1993) in order to assist the application of the dialogue test (under the assumption that such verbs behave in a similar way with regard to obligatoriness vs. optionality). We assume that, by subsequently using the valency data for various tasks and applications, we can achieve further refinement of the relevant criteria.

## 3 THE PROCESS OF CREATING VERBAL VALENCY FRAMES

### 3.1 VALENCY FRAME AND ITS SURFACE REALIZATION IN THE PDT-VALLEX LEXICON

For each verb, the appropriate functor as well as its surface realization (surface-syntactic and morphological form) is recorded in every slot of its valency frame. In general, the mapping of the valency frame to its surface realization can be quite complex (see Hajič et al., 2003, Hajič,

Urešová, 2003), but with a pinch of salt we can assume that each of the valency members (slot fillers) can be mapped to its surface form independently. The surface realization through the morphemic case, preposition and morphemic case, and subordinate sentence with a conjunction is the most common.

For instance:

***snížit***

- valency frame: ACT(.1) PAT(.4) ?ORIG(z+2) ?EFF(na+4)
- example: *snížit nájem z 8 na 6 tisíc*  
(lit.: *lower the rent from 8 to 6 thousand*)

The question mark in front of the valency member in the above example denotes optionality, the other valency members are obligatory. The valency frame can also be empty, denoting that the valency frame does not contain any valency member. For instance, the verb *pršet* (lit.: *rain*) has an empty valency frame (written as EMPTY).

The surface realization of the valency frames is important information for the automatic generation procedures of the surface structures as well as for the automatic “translation” of the analytic sentence representations to the tectogrammatical ones. The knowledge of the surface-syntactic realization is of course already useful in the course of manual annotation in order to distinguish individual valency members (by being suitably careful; in so doing one should not forget that, during the annotation process, the valency lexicon is simultaneously being created and verified). For polysemic lexemes, the surface realization can indicate more or less subtle semantic differences and thus help the process of manual annotation by distinguishing individual valency frames (and, therefore, the individual senses or at least groups of senses of the lexeme). The surface-syntactic realization is aimed at the analytical level of sentence representation (i.e., at the next level down, where every surface word is represented by one annotation unit; we consider the morphological annotation to be part of the analytical level). All the necessary conditions for a part of speech or the morphemic realization of individual members of verbal frames (or even specific lemmas, such as prepositions) should be specified. The original notation, known from the literature on valency for tectogrammatical tree structures, has been extended and an enriched formalized notation of surface realizations of individual valency members has been proposed. It captures not only the simple cases (such as the requirement for a certain morphemic case of the dependent member, regardless of the part of speech and other characteristics), but also the surface structure of idioms, which is often very complicated.

In order to describe the surface realization of the valency frame, we have first to capture the surface structure of this realization in the way it is represented at the analytical level of annotation (see Bémová et al., 1997). Square brackets are used to denote (analytical-level) dependency and a comma is used for separating sister nodes: the governing node is written first, followed by the opening square bracket ('['), the dependent node<sub>1</sub>, dependent node<sub>2</sub>, etc., then the closing square bracket (']'). The requirements on the part-of-speech and morphemic characteristics of individual nodes are written in a shorthand form (by means of a single character for each category) after the dividing symbol '.' (full stop) or ':' (colon) in the following order: part of speech, gender, number, case, degree of comparison and an agreement. For example, for an accusative requirement we write .4, for a plural locative .P6 etc. If any of these characteristics are missing, then this indicates that the given category can take any value in the annotation (with the exception of the first one, the major part-of-speech category, for which more complicated rules apply if no concrete indication is present). The



lemma (its analytical form, i.e. the form which corresponds to the morphological lexicon) is put, if it is needed, in front of the separator: a requirement for the preposition *s* (lit.: *with*) with instrumental looks like this: *s*[.7]. Some special symbols are used for capturing the omission of the member at the analytical layer. In order to shorten the realization in the most common case (which is the requirement for a preposition and a certain morphological case) an abbreviation “preposition+case” instead of “preposition[.case]” can be used (this is the description method usually used in the literature, such as (Panevová, 1974-75)). The difference between a period and a colon as the separators of the lemma and the morphological part of the realization is as follows: the period determines the node of the corresponding analytical tree on which the nodes corresponding to the verbal complementations at the TR should depend (this difference is particularly important for complex phrasal slot descriptions).

For instance:

- volat** – frame: ACT(.1) PAT(.4) i.e. the Actor in Nominative, the Patient in Accusative
  - example: *volejte telefonní číslo 205338* (vytáčet) (lit.: *call the phone number 205338*) (to dial)
- frame: ACT(.1) PAT(po+6) i.e. the Actor in Nominative, the Patient with the preposition “po” + locative
  - example: *volat po otevřeném trhu* (vyžadovat, usilovat) (lit.: *clamour for the open market*) (to ask for, to cry out for)

Different meanings can have the same morphological realization of the valency frame; this is used only when a clear distinction between the meanings (senses) exists (see Lopatková, Panevová, this volume):

For instance<sup>5</sup>:

- zakládat** – frame: ACT(.1) PAT(.4)
  - example: *zakládat sukni* (zkracovat) (lit.: *to shorten a skirt* (to shorten [by folding]))
- frame: ACT(.1) PAT(.4)
  - example: *zakládat stránky v knize* (označovat) (lit.: *to mark the pages in a book* (to mark))

For obligatory free modifications only, empty parentheses may be used to denote any surface realization usual for the free modification in question. The realization of inner participants is always given in full, since there is no “standard” or “default” realization for any of them.<sup>6</sup>

Examples of realizations:

Simple morphological case (.1, .2, .3, .4, .5, .6, .7)<sup>7</sup>

Prepositional case (preposition without vocalization and the number of the required morphological case): na+4, k+3, o+6, ...; or secondary preposition and the

<sup>5</sup> Please note that the “*zakládat*” valency frames quoted above are only examples and they do not represent all the existing valency frames of this verb.

<sup>6</sup> Frequency-wise, of course, some realizations are more frequent than others – for example, for ACT in an active verbal construction, the nominative case is very often used.

<sup>7</sup> Numbers are used in Czech grammars to denote cases: 1 for nominative, 2 for genitive, 3 for dative, 4 for accusative, 5 for vocative, 6 for locative, and 7 for instrumental.



number of the required morphological case: e.g., *prospěch*[v,2],<sup>8</sup> lit.: *to the benefit of*  
 Infinitive: (.f)  
 Subordinating conjunction: (že, aby, když, zda, jestli, ať, ...; lit.: *that, to, when, whether, if, let, ...*)  
 Subordinate clause without conjunction (.c); (if started for instance with an interrogative pronoun or adverb: *který, proč, kde, kdy, ...*; lit.: *which, why, where, when, ...*)  
 Adjective: (usually with a case, e.g., .a7)  
 Adverb: (.d)  
 Interjection: (.i)  
 Numeral: (.m)  
 Pronoun: (.p)  
 Construction with ‘to be’ (*to be* and the required morphological case, e.g., *být*[.7])  
 Direct speech: (.s)  
 Any common (“standard” for given functor) realization: ()  
 State: (=)  
 Empty frame: (EMPTY)

The annotation of idioms (functor: DPHR) is much more complicated. Almost always, it is necessary to capture a particular lemma with an appropriate morphological case and often also with a number: *jít příkladem*: DPHR(příklad.S7) (lit. *go [by an] example; give an example*). A lemma with a required prepositional case also occurs very often: *lapat po dechu*: DPHR(po[dech.S6]) (lit. *catch [s-one’s] breath*). The phrase is sometimes realized through even more complex (sets of) dependent subtrees: (*někomu*) *běhá mráz po zádech* (lit.: [*a*] *frost runs on [sb] back; a shiver runs down sb’s spine*): DPHR(mráz:S1,po[záda.P6]).

### 3.2 THE PROCESS OF BUILDING THE PDT-VALLEX LEXICON

The annotators work primarily only with those verbs (or their senses) found in the PDT data. On the other hand, every occurrence of a verb in the corpus contains a reference to its valency frame (i.e., to an entry in the valency lexicon). The annotators insert the verbs (senses) found in the course of the annotation and their associated valency frames into the lexicon. They create the particular valency frame and write an example (or more examples) of its usage. If they find it reasonable, they can insert a note that refers to another verb that has one of its valency frames related to the current one (a synonym/antonym, an aspectual counterpart, etc.).

Notes and comments on problems encountered during the creation and/or usage (annotation) of the valency frames can also be recorded.

### 4 PROBLEMS RELATED TO THE VERBAL VALENCY

Naturally, many problems and confusions emerged in the course of verifying and adopting the valency theory to particular verbs during the annotation. Let us focus briefly on some of them.

<sup>8</sup> The conjunction “jako” (*as*) is also included in the list of the prepositions, as it requires a particular morphological case in some valency frames. For instance: *bral to jako problém* (lit: *he considered it as [to be] a problem*).

#### 4.1 MISSING OPTIONAL VALENCY SLOTS

It is natural that the annotators primarily include and describe valency slots according to their surface realization as it occurred in the data. That is why a valency frame in the PDT-VALLEX might sometimes not contain an optional inner participant because it is difficult to determine such inner participants (and the dialogue test is of no help either, because it is not applicable for determining optional slots). For instance, with the verb mentioned above *snížit* (*to lower*), only two inner participants (ACT and PAT) were at first listed in the lexicon and only when the construction *to lower the rent from 8 to 6 thousand* occurred was the frame extended with the optional inner participants ORIG and EFF. Similarly, a valency frame may not capture all the possible morphemic realizations of the given valency slot; however, the valency frame should contain all the morphemic realizations that occur in the annotated data. From this point of view, the complex Valency Lexicon VALLEX (see Straňáková-Lopatková and Žabokrtský, 2002) is more complete in describing valency frames in full, using the much bigger (yet syntactically unannotated) Czech National Corpus as its data base; its entries for each verb are meant to contain all meanings and all possible surface realizations (as well as some other additional information).

#### 4.2 COMPETITION BETWEEN AN INNER PARTICIPANT AND A FREE MODIFICATION

*Competition* between two or more functors is understood to be a situation when a valency member occupies (meaning-wise) just one valency slot, but both (or more) functors apply (based on their “semantic” definitions). The current representation of the valency frame does not permit the labelling of one valency frame slot with more than one functor.

##### 4.2.1 COMPETITION BETWEEN AN ADDR AND A LOC/DIR3/DIR1

The obligatory functors LOC (location – answer to a question “where?”), DIR3 (direction to – answer to “where to?”) and DIR1 (direction from – answer to “from where?”) compete in the valency frame of some verbs with an ADDR (Addressee).

For instance:

**podat**

frame: ACT(.1) CPHR({přiznání ...}.4) ADDR(.3)  
*podat přiznání úřadu... to whom*  
(lit.: *to-file a-tax-return [to] the-office(Dat)*)  
frame: ACT(.1) CPHR({přiznání, ...}.4) DIR3()  
*podat přiznání na úřad... to where*  
(lit.: *to-file a-tax-return into the-office*)  
frame: ACT(.1) CPHR({přiznání, ...}.4) LOC()  
*podat přiznání na úřadě...where*  
(lit.: *to-file a-tax-return at the office*)

Other examples:

**ukrást** peníze bance.ADDR / z banky.DIR1  
(lit.: *to steal money the-bank(Dat).ADDR / from the bank.DIR1*)  
**odebrat** děti rodičům.ADDR / od rodičů.DIR1  
(lit.: *to-take-away the-kids the-parents(Dat).ADDR / from parents.DIR1*)  
**dát** listinu úřadu.ADDR / na úřad.DIR3  
(lit.: *to-give the-document the-office(Dat).ADDR. / at-the-office.DIR3*)

It is clear that there is only one valency slot (the valency members cannot occur simultaneously in the given sense in one clause) with different morphemic realizations. Because of the different morphemic realizations *and* due to the current definitions of functors ADDR, LOC, DIR3, DIR1 *and* because of the fact that one valency slot cannot be occupied by more than one functor, it is necessary to create three different valency frames.<sup>9</sup>

If the corresponding surface realization is omitted from the actual sentence, it is difficult for the annotator to make a decision as to which of the competing functors has to be assigned to the restored (obligatory) node. By convention, Addressee (as an inner participant) has priority, so a node labelled as the Addressee is added to the annotation in such a case.

#### 4.2.2 COMPETITION BETWEEN ADDR AND BEN

The dividing line between the inner participant ADDR (Addressee) and the free modification BEN (Benefactive) is not often clear. The situation is easy if the dative or the prepositional case “pro+4” (*for+Accusative case*) is present in the annotated clause. The dative is prototypically considered to be an Addressee, whereas the prepositional case “pro+4” is prototypically a Benefactive.

For instance:

*přinesl jí.ADDR pro tatínka.BEN dopis*  
(lit.: [he] brought her(PronPers.Dat).ADDR for [her] dad.BEN a-letter).

The situation is more complicated if only one morphemic realization from the previous two is present in the given clause. We have thus used the following criteria for distinguishing an Addressee and a Benefactive:

The dative is prototypically an Addressee; however, a dependent in the dative is labelled Benefactive if the dative construction can be substituted with a possessive pronoun.

For instance:

*barvit jí.BEN vlasy...její vlasy*  
(lit.: to-color her(PronPers.Dat) hair...her(PronPoss) hair)  
*amputovat mu.BEN nohu...jeho nohu*  
(lit.: to-amputate him the-leg...his leg)  
*líbat jí.BEN ruku...její ruku*  
(lit.: to-kiss her(PronPers.Dat) hand...her(PronPoss) hand)  
*vidět mu.BEN do duše...jeho duše*  
(lit.: to-see him into soul...his soul)

This rule, however, has further exceptions. For instance, substitution is possible in the following construction: *odebral nám tři body* (lit.: he took [from] us(Dat) three points), but the dative is still labelled as an Addressee here, namely because it is a valency member for the verb *odebral* (the valency relation always has precedence.) The possibly occurring Directional could only be a free modification here, e.g., *odebral nám.ADDR tři body z tabulky.DIR1* (lit.: he took [from] us(Dat).ADDR three points from the-chart.DIR1)

If there is an additional valency slot which is not an Addressee but a kind of Directional with the particular verb (see 4.2.1), then the dative construction is labelled Benefactive.

<sup>9</sup> One might consider using a special “group functor”, in this case for Addressee, Locative, and Directional, in order to create just one valency frame. For issues of semantic and syntactic coherence, see also (Levin, 2003) and (Kingsbury, Palmer, 2002).

For instance:

**odebrat** mu.BEN krev ze žíly.DIR1  
(lit.: [he] took-away him.BEN the-blood from the-vein.DIR1)  
– frame: ACT(.1) PAT(.4) DIR1()

Compare:

**odebrat** tělu.ADDR potřebné látky  
(lit.: to take-away the-body (Dat).ADDR the-necessary substances): an obligatory Addressee.  
**odebrat** z těla.DIR1 potřebné látky  
(lit.: to take-away from the-body.DIR1 the-necessary substances away): an obligatory Directional  
**odebrat** mu.BEN z těla.DIR1 potřebné látky  
(lit.: to take-away him(Dat).BEN from the-body.DIR1 the-necessary substances): an obligatory Directional, Benefactive in the dative case.

The prepositional form “pro+4”, while being prototypically a Benefactive, expresses an Addressee if this form can be substituted by a dative without a change of meaning.

For instance:

**přinášet** pro úřednici (=úřednici).ADDR dopis  
(lit.: [to] bring for a-clerk (=to) a-clerk(Dat)) a-letter)  
**přivést** pro maminku (=mamince).ADDR květiny  
(lit.: [to] bring for mum (=to) mum(Dat)) flowers)

The adequacy of this treatment of the prepositional form “pro+4” (i.e., the possibility to label it as an Addressee) is attested by examples of coordination of the two different forms (“pro+4” and the prepositionless dative) which should be annotated by the same functor.

For instance:

**poskytoval** mu bydlení a pro Alenu taky  
(lit.: [he] provided him(Dat) accommodation and for Alena too)  
**zajistil** nám pobyt a pro sebe taky  
(lit.: [he] booked us(Dat) a-stay and for himself too)  
**zaručil** nám i pro ně stejné podmínky  
(lit.: [he] guaranteed us(Dat) and for them the-same conditions)

The presence or absence of Benefactive and Addressee can also distinguish the meaning of the verb.

For instance:

**nosil** mu.BEN (kamarádovy) batohy (přenášet)  
(lit.: he carried him(Dat).BEN bags) ... (his bags to, e.g., save him work)  
vs.  
**nosil** mu.ADDR (kamarádovi) batohy (přinášet)  
(lit.: he was-bringing him(Dat).ADDR bags) ... (moving bags to his proximity)

#### 4.2.3 THE COMPETITION BETWEEN ORIG A DIR1

A similar situation arises between the inner participant Origin (ORIG) and the free modification Directional (DIR1), which expresses the direction “from where”. The problem

lies in the question “from where?” which can, in many cases, indicate not only the Origin but also the Directional.

If the valency slot has in its surface realization description the form “od+2” (from+Genitive case), which is quite typical of Origin, and, as another possibility also, the form “z+2” (“from inside”+Genitive), which is typical of Directional, we prefer to label such a slot as the (inner participant) Origin. We assume that both forms have the same semantics in such cases. This is displayed e.g. by verbs with the meaning “to gain something from somebody (= from somewhere)”.

For instance:

**čerpát** *od kolegy / z textu informace*  
(lit.: [to] gather from a-colleague / from a-text information)  
**obdržet** *od úřadu / z úřadu povolení*  
(lit.: [to] receive from an-office / from an-officer a-permit)  
**dostat** *od banky / z banky finanční podporu*  
(lit.: [to] get from a-bank / from[-inside] a-bank financial support)  
(please note the homonymy of the English preposition “from” in Czech, cf. also below)  
**půjčit si** *od banky / z banky peníze*  
(lit.: [to] borrow from a-bank / from[-inside] a-bank money)

Sometimes both prepositions (*od*, *z*) can appear in one clause in a text. One of the constructions will then be labelled as the free modification DIR1. It is up to the annotator to distinguish, usually on the basis of the context, the semantic difference between them.

For instance:

**půjčil si** *od tatínka. ORIG z účtu. DIR1 značnou sumu*  
(lit.: [he has] borrowed from [his] father. ORIG from the-account. DIR1 an-appreciable sum).

If the verbal complementation can only be realized as “z+2”, we assume this is the free modification Directional, not an Origin, and the Directional is not here a part of the valency frame. This rule applies e.g. for verbs with the meaning “to pay to somebody something from somewhere”.

For instance:

**financovat** *stavbu z rozpočtu*  
(lit.: [to] finance the-construction from[-inside] the-budget. DIR1)  
**hradit** *náklady z fondu oprav*  
(lit.: [to] cover costs from[-inside] the-resources. DIR1 of-repair)  
**dotovat** *výdaje ze státních rezerv*  
(lit.: [to] supplement expenses from[-inside] the-state reserves. DIR1)

However, if the complementation can be realized by the form “z+2” but there is also another valency member, namely Effect (mostly expressed by the prepositional forms “do+2” (to/into+Genitive), “v+4” (in/into+Accusative), “na+4” (on/to/onto/into+Accusative), we consider this to be a valency complementation and it is labelled Origin.

For instance:

**překládat** *z češtiny do němčiny*  
(lit.: [to] translate from Czech into German)

*změnit účes z kudrn na rovné vlasy*  
 (lit.: [to] change haircut from curler into straight hair)  
*klesnout z tisíce na pět set*  
 (lit.: [to] sink from [one] thousand to five hundred)

The situation is, however, more complicated in many other cases. The common meaning of “origin” (most often expressed by the prepositional constructions “z+2”, “od+2” as discussed above) can become split into more frames with different functors assigned to the slot with this surface realization.

For instance, the verb *pocházet* (lit. *come-from*) finished up with three different frames, with the slot in question labeled PAT, DIR1 and TFRWH (temporal “from when”), respectively:

*zboží pochází z Prahy*.PAT (shifted from ORIG)  
 (lit.: the goods come-from from[-inside] Prague)  
 (in the sense “from local (Praguian) producers”)  
*matka pocházela z Moravy*.DIR1  
 (lit.: [the] mother came-from from Moravia)  
*kniha pochází ze 12. století*.TFRWH  
 (lit.: [the] book comes-from from [the] 12<sup>th</sup> century)

By using the Origin or Directional functors in the valency frames, we often distinguish an abstract and a concrete meaning of a verb respectively (see also 4.4).

For instance:

*přecházet* (cross [over], change, switch)  
*přecházet z desetihodinového na osmihodinový provoz*.PAT  
 (lit.: [to] change from ten-hour-long to eight-hour-long shifts )  
 vs.  
*přecházet z jedné strany na druhou*.DIR1  
 (lit.: [to] cross [the street] from one side to the-other)  
*vymáčkout* (squeeze, press, get out)  
*vymáčkout z obyvatel/od obyvatel*.ORIG daně  
 (lit.: [to] get-out from/from[-inside] the-dwellers the-taxes(Acc))  
 vs.  
*vymáčkout z citrónu*.DIR1 šťávu  
 (lit.: [to] press from[-inside] the-lemon the-juice(Acc))

The competition of ADDR and BEN described earlier and the competition of ORIG and DIR1 confirms the assumption that the semantic classification cannot always readily correspond to formal indicators (i.e., to the surface realization by prepositions, morphemic cases, etc.). It is obvious from the PDT annotation that a solution to the problem of the competition of certain functors is a very difficult task and it is not yet satisfactorily solved, either formally (should we allow for more valency frames for a single meaning, or should we use groups of functors, etc.?) or in the practice of annotation. This problem and the current solution are considered open for further discussion and, undoubtedly, a more detailed examination is required.<sup>10</sup>

<sup>10</sup> For another account of the alternation of some types of valency complementations, see also (Benešová, 2002).

#### 4.3 OVERLAP OF MISCELLANEOUS TYPES OF FREE MODIFICATIONS

The specification of the functors denoting free modifications is based on their semantics. It is not easy to define the particular functor entirely unambiguously and “sharply”. The annotators have to help themselves, for the consistency of the annotation, by means of various criteria (often based on morphosyntactic rules) in the gray zone of an overlap of two (or sometimes even more) functors of miscellaneous types of free modifications.

##### 4.3.1 OVERLAP OF TWHEN AND LOC

Even though this overlap seems improbable, it occurs fairly frequently. To help us to resolve it, we have used a transformation of the construction in question into predication: if the most natural transformation into a complex (subordinate) sentence opens with the time conjunction „*když*“ (lit.: *when*), it indicates that a time functor should be assigned. The fact that the noun in the construction is an event noun can be used as a supportive criterion leading us to assign the temporal interpretation, too.

For instance:

<b>podlehli</b> v zápase.TWHEN (lit.: [they have] lost in [the] fight)	“..., when they fought”
v polemikách.TWHEN <b>likvidoval</b> soupeře (lit.: in the-argumentation [he] liquidated the-rivals)	“..., when he argued”
<b>oznámil</b> to v rozhovoru.TWHEN (lit.: [he] announced it in a-talk)	“..., when he talked”
akcie <b>patřily</b> v první vlně.LOC k nejatraktivnějším (lit.: the-shares belonged in the-first wave to the-most-attractive)	“where did they ...?”
v tomto příkladu.LOC <b>nejde</b> o jednoduchou úlohu (lit.: in this case is-not-the-matter about an-easy task)	“where it is not ...?”

##### 4.3.2 OVERLAP OF INTT AND LOC/DIR3/DIR1

The semantics of the governing verb (mostly a verb of motion) leads in some cases to an uncertainty as to whether or not Intention (INTT, a free modification) rather than an obligatory Direction (DIR1, DIR3) or Location (LOC) from the valency frame is concerned.

For instance:

<b>přišel</b> se-koupat (lit.: [he] came [to] swim)
<b>došel</b> nakoupit (lit.: [he] went [to] shop),
<b>vydat-se</b> na jahody (lit.: [to] set-out for strawberries),
<b>dorazil</b> překonat record (lit.: [he] arrived [to] beat the-record),
<b>odešel</b> se-rozcvičit (lit.: [he] went [to]warm-up)
<b>zůstal</b> na oběd (lit.: [he] stayed for lunch)

The current version of the PDT-VALLEX lexicon prefers to use the Direction (as an obligatory slot) in the above cases as well as in other similar cases, since Intent was not considered to be an obligatory slot in the valency frame but just a “pure” free modification (cf. Lopatková, Panevová,



this volume)<sup>11</sup>. On the other hand, some of its properties render its position on the inner participant↔free modification axis not quite clear. We leave the question open as to whether or not INTT should become an obligatory valency member in cases where the original verbal meaning, i.e. intentional movement to somewhere or from somewhere, fades away to such an extent that the spatial meaning is irrelevant. For instance, in such collocations as *jdu se oženit* (lit.: [I] am-going myself to-marry) (I want, I mean), *jdu jí napsat* (lit.: [I] am-going myself her to-write) (again: I want, I mean), *jde si zapnout kabelku* (lit.: [she] goes herself to-clip-up [her] bag) (she wants, she means), it is obvious that the voluntative modality of Intent (intent to do something) has a priority over the Direction. In this case the possibly occurring Direction (in the same clause) should be labelled as a free modification.

To summarize, INTT is currently always treated as optional (and because it is a free modification, it is never a member of a valency frame – cf. sect. 1 for definitions and principles) and, due to lack of agreement on a usable “semantic” definition, in almost all cases it is assigned on the basis of its morphemic realization (i.e., infinitive, “pro+4”, “na+4” or “k+3”) rather than on that of semantics.<sup>12,13</sup>

#### 4.4 ABSTRACT AND CONCRETE MEANING OF SURFACE DIRECTIONAL EXPRESSIONS

Abstract and concrete usages of verbs are often distinguished in the lexicon by using separate valency frames. The original examples of “general directionality” are split into several valency frames in the PDT annotation.

For instance:

##### **přijít**

<i>přijít ke stromu</i> .DIR3 (lit.: [to] come to the-tree)	– “přistoupit” (to move close to)
<i>přijít k penězům</i> .PAT (lit.: [to] come to the-money)	– “získat” (to get)
<i>přijít na řešení</i> .CPHR (lit.: [to] come onto a-solution)	– “napadnout”, “vyřešit” (to solve)

Other examples:

##### **ustoupit**

<i>ustoupit od zdi</i> .DIR1 (lit.: [to] go-away from a-wall)	– “vzdálit se” (to move away from)
<i>ustoupit od myšlenky</i> .PAT (lit.: [to] go-away from an-idea)	– “vzdát se” (to abandon [an idea])

<sup>11</sup> In the earlier works on verbal valency (esp. the works of Panevová quoted above), some free modifications were considered to have properties that it would not be counter-intuitive to designate them as obligatory, such as BEN or INTT.

<sup>12</sup> With the exception of *jít*, *chodit*, *vycházet za prací* (lit. go, be going, go-out for/to work), where the surface realization “za+7” (“for+ Instrumental case”) is used.

<sup>13</sup> Similarly, the decision as to whether to use INTT or AIM (Aim) depends solely on the form: while INTT only has been assigned in the cases just described, AIM has been used only if it was realized on the surface as a subordinate clause with the conjunction *aby* (lit.: to). This made the task easier for the annotators and the annotation is thus consistent, at the expense of hiding the semantic difference if it goes against the form. Lately, the specification of the difference between INTT and AIM was reconsidered to become less formal and more semantic (see Panevová, Lopatková, this volume), but such a treatment might only influence future versions of the PDT.

### **vycházet**

*vycházet z lesa*.DIR1

(lit.: [to] come-out from the-forest)

*vycházet z předpokladu*.PAT

(lit.: [to] start from an-assumption)

– “opustit a vzdálit se” (to move out of)

– “začít” (start with, from)

The consistency of annotation of this kind of problematic valency frames is low in the annotated data, since not all occurrences contain such clear-cut cases as the examples above. This group of valency frames is also considered open to further and more detailed examination. See also 4.2.3.

#### 4.5 CO-OCCURRENCE OF TIME AND LOCAL COMPLEMENTATIONS

Two local or time complementations such as *zítra k večeru* (lit.: *tomorrow towards evening*), *hluboko pod povrchem* (lit.: *deep under the-surface*) have their own specific character. It is difficult to treat them in the dependency syntax formalism because no clear dependency direction (and/or structure) can be established using the usual (omission-without-loss-of-grammaticality) criteria. Applying these theoretically-based criteria on a large amount of data during the annotation, we failed to consistently and uniquely determine the governor and the dependent: it was found empirically that one particular member can sometimes be omitted (without making the sentence ungrammatical, with the usual caution) but such a consideration did not generalize well, because in many instances neither the former or the latter part of such constructions can be omitted.

The “grammatical” omission can take place e.g. when the time complementation is in the Accusative (*Oblékla-se půl hodiny před začátkem představení*, lit.: [she] dressed-up half(Acc.) an-hour before the-start of-the-performance), where “*půl hodiny*” can be omitted, but it could also be the other way round (*Strávila tam dva měsíce před porodem*, lit.: [she] stayed there two months(Acc.) before the-delivery) – here, only “*před porodem*” can be omitted. An example of a case where neither part can be omitted is e.g. “*Leží to dva kilometry od řeky*” (lit.: [it] lies two kilometers away-from the-river).

The currently used solution for annotation in PDT 2.0 is as follows: the first part is always considered to be the governor; this means that the first part is always modified by the other local or time verbal complementation.

For instance:

**leží** to *hluboko pod povrchem*

(lit.: lies it deep under the-surface)

**pojedeme** na *západ od Prahy*

(lit.: [we will] go to the-west from Prague)

**dorazil** *pět minut po odjezdu vlaku*

(lit.: [he] arrived five minutes after the-departure of-the-train)

**vrátí se** *brzy po Vánocích*

(lit.: [he will] return himself soon after Christmas)

#### 4.6 THE FUNCTOR “STATE”

A question arose during the annotation as to whether a modification that is semantically different, but formally identical to the LOC (cf. 1.) or DIR3 (cf. 2.) functor should be distinguished in the valency frames.

For instance:

**ocitnout se**

*ocitla se v Praze.LOC*

(lit.: [she] found herself in Prague)

vs.

*ocitla se pod tlakem.???*

(lit.: [she] found herself under pressure)

**dostat se**

*dostala se do Brna.LOC*

(lit.: [she] got to Brno)

vs.

*dostala se do maléru.???*

(lit.: [she] got-involved herself in a-mishap)

Here we believe it is not appropriate to follow only the morpho-syntactic considerations (both complementations would then get the functor LOC). That was why we preliminarily set up a new functor which would label this type of dependency as “State”. So far, this functor has not been used but a special node attribute with a special value for State will serve this purpose, using the syntactically closest functor label. The annotator currently adds an alternative of “an undefined functor” (in such cases a star appears by the primary functor of a node in the annotated data). A special symbol “=” has been used in the valency lexicon so far.

For instance:

*dát věci **do souvislosti***

(lit.: [to] put things into perspective)

valency frame: ACT(.1) PAT(.4) DIR3(=)

*držel byt **v pořádku***

(lit.: [he] kept the-flat in order)

valency frame: ACT(.1) PAT(.4) LOC(=)

*hnát řešení **do krajností***

(lit.: [to] push the-solution into the-extremes)

valency frame: ACT(.1) PAT(.4) DIR3(=)

Other examples:

*jít **do likvidace***

(lit.: [to] go into liquidation)

*nechat sportovce **v klidu***

(lit.: [to] leave the-sportsman at rest)

*odsouvat osobnost **do zapomnění***

(lit.: [to] shift a-personality into oblivion)

It is important to say that the “functor” State requires further investigation from different points of view; various subtle semantic differences occur in such constructions and it is not yet clear how to describe them precisely and in sufficient detail. However, from the standpoint of further research, we consider even the mere separation of such constructions in the valency frames useful.

#### 4.7 VALENCY OF VERBS OF FOREIGN ORIGIN AND THEIR CZECH COUNTERPARTS

Valency frames of verbs of foreign origin are created having their “Czech” synonyms (if they exist) in mind. Thus, in most cases, the valency frame of the “foreign version” of a verb and its Czech counterpart is the same.

For instance:

<b>vystěhovat se</b> z venkova do města (lit.: [to] move from the-countryside to the-city)	valency frame: DIR1()
<b>emigrovat</b> z východu na západ (lit.: [to] emigrate from the-East to the-West)	valency frame: DIR1()
<b>zacházet s</b> penězi (lit.: [to] deal(handle) with money)	valency frame: PAT(s+7)
<b>disponovat se</b> zásobami (lit.: [to] deal(control, handle) with the-reserves)	valency frame: PAT(s+7)
<b>manipulovat s</b> mříží (lit.: [to] manipulate with the-grid)	valency frame: PAT(s+7)
<b>uvažovat</b> o životě (lit.: [to] think about life)	valency frame: PAT(o+6)
<b>meditovat</b> o zvycích (lit.: [to] meditate about traditions)	valency frame: PAT(o+6)

Other examples:

**dislokovat/umístit**  
(lit.: to dislocate/to lie down)  
**deportovat/vyhostit**  
(lit.: to deport/ to banish)  
**demontovat/rozebrat**  
(lit.: to dismantle/to strip down)  
**devalvovat/znehodnotit**  
(lit.: to devalue/to invalidate)  
**absolvovat/zakončit**  
(lit.: to go through(pass)/to finish)

#### 4.8 ONE OR TWO FRAMES?

Verbs with seemingly optional Patient form another class of uncertainty. Here, it is often unclear whether one verb has two meanings (and thus should be split into two different frames).<sup>14</sup> This problem concerns verbs such as:

- a) **podnikat, plavat, běhat** (to undertake, to swim, to run)
- b) **kousat, kouřit, kojit, zavěsit** (to bite, to smoke, to nurse, to hang up)
- c) **tančit, cvičit, trénovat** (to dance, to exercise, to practise)
- d) **mluvit, hovořit, číst, psát, zpívat** (to speak, to talk, to read, to write, to go, to sing).

If we decide to use two different frames (cf. group (a)), then the first frame does not include a Patient slot and the second one does (an obligatory one, of course). Otherwise, we stick with just one frame with either an obligatory Patient (cf. group (b)) or an optional Patient (cf. group (c)) or without a Patient slot altogether (cf. group (d)).

<sup>14</sup> Such verbs can have even more senses, which can be quite distinct, e.g. *komín kouří* (lit.: the chimney fumes).

The reason for treating these four groups differently is that they behave differently. The meaning of “doing or running an activity (without a specific object in mind)” has the right to have its own valency frame in case of group (a). These verbs thus get two frames. The first is simply a single-slot frame ACT (.1) – e.g., *Kamarád už dlouho podniká* (lit.: *a-friend already for-a-long-time has-a-business*), *Anna plave závodně* (lit.: *Ann swims professionally*). The second is a two-slot frame ACT (.1) PAT(.4) – e.g., *Plaval dvacet bazénů denně* (lit.: *he-had-swam twenty pools daily*), *Jirka podniká velké cesty* (lit.: *Jirka undertakes big journeys*). Notice also that the English translation of *podniká* is different in these two cases, a strong indication of two different meanings.

By contrast, the obligatory valency complementation is necessary for verbs in group (b). These verbs correlate very strongly with a specific Patient; therefore we consider a Patient to be always present here. The meaning of “doing or running an activity (without specific object in mind)” is just a sub-meaning of this valency frame. Thus, these verbs have the following two-slot frame: ACT (.1) PAT(.4).

In group (c), we decided to assign only one frame with an optional Patient: ACT(.1) ?PAT(.4). E.g., in the clause *Jirka denně cvičí a trénuje* (lit.: *Jirka daily exercises and practises*) the verbs *to exercise* and *to practise* have a special “abstract” semantic characteristic „doing or running an activity“, where it is unimportant *what* exactly he is exercising or practising. On the other hand, we can certainly express some particular activity (which is always going on “behind the scenes”) that Jirka can exercise or practise. Then, such an activity would be assigned the Patient functor.<sup>15</sup>

We believe that the last group of verbs, (d), has an additional semantics of the verb *umět* (lit.: *to know*) in one of its meanings, in sentences like *Pavel už mluví, ale ještě nečte, nepíše a nepočítá* (lit.: *Paul already talks, but [he] yet [does] not-read, not-write and not-count*), *Anna mluví hezky německy a už i zpívá* (lit.: *Ann speaks well German and already even sings*). These verbs have been assigned a valency frame without a Patient in this meaning (i.e., ACT (.1) only).

#### 4.9 RECIPROCITY IN VALENCY

The notion of reciprocity belongs to events where a two-way “direction” between its two participants can be observed, either happening simultaneously (*they met*) or mutually (“reciprocally”: *they were helping each other*). It is well known (Panevová, 2003) that reciprocity changes the surface realization of the verbal valency structure in a way non-reciprocal events do not.

In most cases, it is ACT and PAT that are in the relation of reciprocity. For example, ACT(.1) PAT(.4) is the valency frame for the verb *líbat* (*to kiss*) but the Patient is not expressed in such reciprocal clauses as *Sourozenci se líbali* (lit.: *Siblings each-other kissed*). We could have used the *siblings* here also as a Patient (by creating another node in the tectogrammatical representation and duplicating the *siblings* there), but that would make it indistinguishable from “*Siblings kissed siblings*” which means something different. This type of “missing” valency member thus gave us a reason for setting up a new value of a tectogrammatical lemma, namely “Rcp”: a new node with this value recorded in the tectogrammatical lemma attribute and a functor label that corresponds to the “missing” dependent is added if the

<sup>15</sup> One might argue that because some people are professional trainers, the verb *to train* should have been moved to group (a). We did not encounter any such example in the data, so we have assigned it to group (b) but of course its future reassignment cannot be excluded.

usage of the verb is reciprocal and the “second” member is reciprocally “included” in the “first” member which is expressed by a plural or as a coordination. Whereas the above clause is an example of the former expression (plural), *Honza a Marie se líbají* (lit.: *John and Mary [each-other] are-kissing*) is an example of the latter (coordination).

The “other” reciprocal participant is often seemingly expressed on the surface by forms of the morpheme (particle) “*se*” or “*si*” (Lit.: *him/herself, Acc. or Dat, or each other.*), sometimes in conjunction with the preposition “*s*” (*with*) used with the Patient; e.g., *Honza se líbá s Marií* (lit.: *John himself is-kissing with Mary*), *Tom s Pavlem si vyměňují známky* (lit.: *Tom with Paul themselves exchange(Pl.) stamps*). Since in these cases the Patient is expressed on the surface, the valency frames for these verbs have to account for it. The realization of this frame must thus also contain the preposition “*s*” with the instrumental: ACT(.1) PAT(.4;s+7). Naturally, the particle “*se*” or “*si*” is discarded in the annotation in all cases, since the reciprocal element can be accounted for by either the expressed participant or by the Rcp node. Other verbs with this kind of frame are e.g. *potkat* (*to meet*), *vítat* and *přivítat* (*to welcome*), etc. Other details about reciprocity (such as the surface realization possibilities of expression) are not marked in the verbal valency frame or in the annotated data because they can be handled by global “grammatical” rules. It is also true that almost every verb can be used reciprocally, at least in theory. Possible restrictions (if any) must be further studied in the future.

#### 4.10 SPECIFIC VALENCY FRAMES: DPHR AND CPHR

Some verbs can also have, on top of regular valency frames, specific frames when used in idiomatic expressions. The verb being assigned such a frame must be a part of an idiomatic construction. One of the frame slots is then labelled by a special functor DPHR (dependent part of a phraseme). The issue of idioms is very complicated and thus it is not easy to find the borderline between metaphorical and non-metaphorical meanings. We use the following principle: if the verb is used in an abstract meaning and has a metaphorical meaning in the given collocation, we mark the remaining part of the collocation as DPHR. Often, the surface realization of such a member is complex and the full power of the formal system of describing surface realizations must be used; e.g., the idiom “*dát něčemu zelenou*” (lit.: *[to] give to-something a-green-light*) has the following valency frame: ACT(.1) PAT(.3) DPHR(zelený.FS4).

If a verb is semantically “emptied” (i.e., the semantic content of the verb is reduced or generalized) and it also meets some other requirements (see Cinková, S., Kolářová, V. this volume), we mark one of its valency members as CPHR (compound phraseme), namely the one that gives the collocation the “real” meaning. The specificity of the verbal valency frames with the functor CPHR (called **support verb constructions**) is consistent, i.e., they do not undergo the process of shifting. For instance: *dostat od otce*.ORIG *příkaz*.CPHR (lit.: *[to] get from [his] father an-order*), *věnovat problému*.ADDR *pozornost*.CPHR (lit.: *[to] pay the-problem attention*).

## 5 CONCLUSION

The annotation of the verbal valency on the background of the Prague Dependency Treebank is a valuable contribution to Czech linguistics especially because a large list of verbs (more than 5300 verbs with 8200 valency frames) has been built on the basis of a corpus which has permitted verification and refining of the notion of valency as a substantial part of the Functional Generative Description theory. It was not necessary to make up valency complementation examples in order to fill out the theoretical schemes of valency frames because they were taken from real data. The results that have been achieved are considered



only as a first step, in this respect providing rich material for further linguistic and computational research. The annotation revealed a number of questions which we have tried to solve. However, many of them remain open for further research and discussion. These open questions, as well as the fact that some of the first decisions in the annotation of verbal valency were not correct and also that some rules of the annotation were changed during the annotation, can be considered as a positive rather than a negative result of our research. The PDT-VALLEX, which is actually a byproduct of the annotation, is an important source for further linguistic research as well as for computational processing of the Czech language. We also hope that it will be a useful resource for many different applications and further studies.

## REFERENCES

- BÉMOVÁ, A., BURÁŇOVÁ, E., HAJIČ, J., KÁRNÍK, J., PAJAS, P., PANEVOVÁ, J., ŠTĚPÁNEK, J. AND UŘEŠOVÁ, Z. (1997): Anotace na analytické rovině: návod pro anotátory. Technical Report ÚFAL TR-1997-03. Charles University. Prague.
- BENEŠOVÁ, V. (2002): Delimitace lexii českých sloves z hlediska jejich syntaktických vlastností. Diplomová práce, FF UK, Praha.
- CINKOVÁ, S., KOLÁŘOVÁ, V. (2006): Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. This volume.
- HAIJČ, J. (2006): The Prague Dependency Treebank, version 2.0. This volume.
- HAIJČ, J., PAJAS, P. (2004): Zápis valenčních rámců v PDT a jeho sémantika. <http://ufal.mff.cuni.cz/pdt>.
- HAIJČ, J., PANEVOVÁ, J., UŘEŠOVÁ, Z., BÉMOVÁ, A., KOLÁŘOVÁ, V. (2003): PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pp. 57-68. Vaxjo University Press.
- HAIJČ, J., UŘEŠOVÁ, Z. (2003): Linguistic Annotation: from Links to Cross-Layer Lexicons. In: Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pp. 69-80. Vaxjo University Press.
- HAIJČOVÁ, E. et al. (in prep.): The Guidelines for Tectogrammatical Annotation. Praha 2004. Part of CD-ROM: Prague Dependency Treebank v. 2.0, LDC, Univ. of Pennsylvania.
- HAIJČOVÁ, E., PANEVOVÁ, J., SGALL, P. (2002): K nové úrovni anotovaného korpusu, část 1, Slovo a Slovesnost, 63, 161-177.
- KINGSBURY, P., PALMER, M. (2002): The Proposition Bank: An annotated Corpus of Semantic Roles.
- LEVIN, B. (1993): English Verb Classes and Alternations. Chicago: University of Chicago Press, pp. 348.
- LOPATKOVÁ, M., PANEVOVÁ, J. (2006): Recent Development of the Theory of Valency in the Light of the Prague Dependency Treebank. This volume.
- LOPATKOVÁ, M., ŽABOKRTSKÝ, Z., SKWARSKA, K., BENEŠOVÁ, V. (2003): Valency Lexicon of Czech Verbs VALLEX 1.0. CKL/UFAL Technical Report TR-2003-18, 2003.
- PANEVOVÁ, J. (1974-75): On Verbal Frames in Functional generative Description. Part I, The Prague Bulletin of Mathematical Linguistics 22, pp.3-40, Part II, The Prague Bulletin of Mathematical Linguistics 23, pp. 17-52.
- PANEVOVÁ, J. (1980): Formy a funkce ve stavbě české věty. Academia, Praha.
- PANEVOVÁ, J. (1994): Valency Frames and the Meaning of the Sentence. In: The Prague School of Structural and Functional Linguistics (ed. Ph. L. Luelsdorff), Amsterdam-Philadelphia, John Benjamins, pp. 223-243.
- PANEVOVÁ, J. (1999): Valence a její univerzální a specifické projevy. In: Hladká, Z., Karlík, P., Čestina – Univerzálie a specifika. Brno, 29-37.
- PANEVOVÁ, J. (2003): Some Issues of Syntax and Semantics of Verbal Modifications. In: Proceedings MTT 2003, First International Conference on Meaning – Text Theory. Paris, Ecole Normale Supérieure, 139-146.
- SGALL, P. (1967): Generativní popis jazyka a česká deklinace. Academia, Praha.



- SGALL, P., HAJIČOVÁ, E., PANEVOVÁ, J. (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects* (ed. by J. Mey), Dordrecht: Reidel and Prague: Academia.
- STRAŇÁKOVÁ-LOPATKOVÁ, M., ŽABOKRTSKÝ, Z. (2002): *Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation*. In: *LREC2002, Proceedings, vol.III.* (eds. M. González Rodríguez, C. Paz Suárez Araujo), ELRA, pp. 949-956.
- ŘEZNÍČKOVÁ, V. (2003): *Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora*. In: *Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLAC 2003)*, pp. 88-97. UCREL, Lancaster University.

#### **ABSTRACT**

The core ingredient of the Prague Dependency Treebank (PDT; see Hajič, this volume) – “valency” – indicates the ability of lexical units to combine with other complementations. The PDT has adopted the concept of the valency theory of the Functional Generative Description (FGD) (see Sgall, 1967, Sgall et al., 1986). The valency theory of the FGD was first developed for verbs, then also for other parts of speech. We present a description of how we dealt with the valency of verbs during the annotation of the PDT and the way the verbal part of the valency lexicon (PDT-VALLEX) was built. We focus on some specific problems related to verbal valency (as well as some other verbal complementations) from the point of view of the PDT.

#### **ACKNOWLEDGEMENT**

The research reported in this paper was supported by the project of the Czech Ministry of Education No. MSM0021620838 and the grant of the Grant Agency of the Charles University No.375 /2005.



# Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank

SILVIE CINKOVÁ, VERONIKA KOLÁŘOVÁ

## 1 INTRODUCTION

Support Verb Constructions (SVCs) are combinations of a noun denoting an event or a state and a lexical verb. From the semantic point of view, the noun seems to be a part of a complex predicate rather than the object (or subject) of the verb, whatever the surface syntax may suggest. The meaning is concentrated in the noun component, whereas the semantic content of the verb is reduced or generalized.

In this article we deal with the question of how to treat SVCs in the Prague Dependency Treebank (PDT subsequently). In the second section we briefly describe what PDT is, what linguistic theory it is based on and what questions regarding the SVCs arose during the annotation. In the third section we explain how SVCs have been identified and inventoried in PDT. We also give a brief survey of how SVCs have been treated within other linguistic frameworks and, based on this knowledge, what conclusions were drawn for PDT. Of course, this survey does not claim to be exhaustive. The fourth section focuses on the semantic aspects of SVCs. The last section describes how the FGD-based valency theory has been implemented in the case of SVCs to provide both a consistent and a linguistically justified annotation in PDT.

## 2 SVCS AS A TYPE OF COMPLEX PREDICATES IN PDT

### 2.1 PDT

For the written Czech language, corpora of two types exist: (i) a databank of linear texts, i.e. the Czech National Corpus (CNC subsequently) at the Faculty of Arts, Charles University, Prague (this is a representative corpus of contemporary written Czech, a part of which, called SYN2000, contains about 100 million word-forms in its current version), and (ii) a dependency-based treebank, i.e. the Prague Dependency Treebank, which is a part of CNC annotated in several layers. The shallow-parsed shape of PDT, so-called analytical level contains approx. 90,000 sentences; the so-called tectogrammatical level of PDT captures the underlying syntactic structures of sentences and contains approx. 55,000 sentences. Both these corpora are annotated (by morphological tags in the full CNC, by morphological tags, syntactic functions, functors, co-reference, and TFA in PDT). The linear corpus is very useful for searching for morphemic and lexical phenomena, including information about their frequency, but the dependency treebank is invaluable whenever one investigates syntactic relations in the sentence. Due to the high degree of “free” word order in Czech, many modifications can occur as either preceding or following their governors. Thus, it is very difficult to formulate a query about syntactic relations in the linear corpus.

## 2.2 THEORY OF VALENCY APPLIED IN PDT

A verb occupies the central position in the sentence structure, so it is clear that one of the key syntactic relations is the valency of verbs, as well as the valency of deverbal nouns and adjectives.

Our approach to issues of valency is based on the theory of valency (especially valency of verbs) as developed in the framework of the Functional Generative Description (FGD subsequently, see Sgall, Hajičová and Panevová 1986; Panevová 1980).

In FGD, the valency frame of a verb, stored in the lexicon, can be described as present on the tectogrammatical level. The following complementations (i.e. the individual dependency relations) are included in the set as being able to fill individual slots of the valency frames of verbs:

(i) inner participants or arguments (they can be obligatory or optional): Actor (ACT), Patient (PAT), Addressee (ADDR), Effect (EFF), Origin (ORIG);

(ii) obligatory free modifications or adjuncts (especially those with the meaning of location (e.g. DIR3, LOC) and manner (MANN)).

Most of these complementations can be omitted on the surface layer of the sentence, but some of them must always be present (as PAT with the verb *potkat* 'to meet', MANN with the verb *chovat se* 'to behave', etc.), unless a textual deletion is concerned, in which case the presence of the complementation in the surface shape is not obligatory. E.g.: *Potkali jste ho?* 'Potkali'. (lit. 'Have you met him?' 'Met\_1stPlur', i.e. 'Have you met him?' 'Yes, we have'.)

In describing the valency frames of deverbal nouns and adjectives, we use the same set of complementations as with verbs. However, in comparison with the frames assigned to the source verb, the process of nominalization (condensation) may be accompanied by a reduction of the number of slots in the valency frames of derived nouns and adjectives at the underlying layer. Moreover, any complementation of a noun can be omitted on the surface layer.

## 2.3 RECORDING COMPLEX PREDICATES IN PDT

In the tectogrammatical annotation of the Prague Dependency Treebank, the need arose to mark complex predicates (subsequently CPs). A CP typically comprises a verb and a noun that make up both a syntactic and semantic unit (e.g. *věnovat pozornost*, lit. 'to pay attention'). It can appear as a nominalization as well (e.g. *věnování pozornosti*, lit. 'paying- of-attention', *pozornost věnovaná dětem*, lit. 'attention-paid to-children'). The PDT annotation also considers certain nouns and adjectives to be special kinds of CPs when appearing with the copula verb *to be*: *být schopen* 'to be able', *být ochoten* 'to be willing'; *Je povinností koalice nalézt řešení*, lit. 'It is an incumbency of the coalition to find a solution', i.e. 'The coalition is obliged to find a solution'. Their nominalizations are also considered to be CPs, such as *Petrova náchylnost k něčemu* 'Peter's predisposition to sth'. Nevertheless, this type of CPs, in which nominal components are mostly marked as PAT, will be omitted from this study. The CPs to be dealt with in this paper are solely those of the verb-noun type, such as *věnovat pozornost*, lit. 'to pay attention', *mít tendenci*, lit. 'to have tendency', *přijít s nápadem*, lit. 'to come up with idea', etc. In accordance with the rich literature in English, they will henceforth be referred to as support verb constructions (SVCs).

Lemmatizing CPs in PDT as multi-word units (MWU) was out of the question as there were already supporting valency frame lexicons for nouns and verbs, respectively. It would have been necessary to design another lexicon to capture the multi-word units. The multi-word units would of course have overlapped with the one-word lemmas whose frames had already been described by the existing lexicons, which would have led to confusions in

valency representations. Apart from the time and effort that a MWU lexicon would have cost, the selection of its lemmas would necessarily have been based on arbitrary decisions on the degree of a collocation's lexicalization. Another essential aspect would have been ignored: that is, that SVCs make up a productive mechanism in the language, allowing for well-formed *ad hoc* constructions (cf. Ekberg, 1989 and Dura, 1997, see below). They can hardly, therefore, be captured by a finite list. A MWU lexicon would have become more of a burden than a supporting tool. Therefore, CPs are not lemmatized as MWUs in PDT.

In SVCs, which is the CP type to be discussed here, the distinguishing feature is the marking of nouns / noun groups as CP components by a special functor CPHR ("Compound Phraseme"). A necessary condition for a noun to obtain a CPHR functor is for it to be an obligatory valency complementation of the verb in the given frame. This implies that an obligatory actant in a verb frame, e.g. PAT, is re-classified as a CPHR when the entire syntagm is considered a SVC (cf.: *to pay 30 dollars*.PAT × *to pay attention*.CPHR). By means of a different functor, we indicate that, from the semantic point of view, the noun within a SVC ceases to be the PAT of the verbal part of the SVC; the fact that it is not appropriate to provide the noun within a SVC with a semantic role is also confirmed by Macháčková (1983, p. 135). This further implies that PDT in fact lists SVCs as frames in the valency lexicon of verbs. Thus a list of SVCs can currently be obtained by searching PDT for frames containing CPHR.

### 3 CRITERIA FOR SVC IDENTIFICATION

#### 3.1 CPHR-CANDIDATE LIST

Before introducing the functor CPHR, a list of CPHR candidates had been collated by searching PDT for "a verb governing a noun governing a PAT-node". The given structure of the query was originally motivated by two aspects:

(i) The realization of SVC as two nodes has certain consequences for co-referential relations, also annotated in PDT (cf. Kučová – Kolářová – Žabokrtský – Pajas – Čulo, 2003). In particular, we wanted to capture the co-referential relations in those SVCs that correspond to synthetic predicates of control, e.g. *Petr se chystá přijít* × *Petr má plán přijít* ('Peter is getting ready to come' × 'Peter has the plan to come'; for more about predicates of control see Panevová – Řezníčková – Urešová, 2002).

(ii) Most nouns that appear in SVCs have their own argument structure, even if they never occur in predicates of control. They are regularly captured by the tectogrammatical tree structures without any problems. However, problems can arise when such nouns (i.e. those having their own dependent nodes) become part of a SVC. Due to certain types of TFA-contingent word order changes, the nodes governed by the SVC-noun node are sometimes located quite distant from their governing node, making the tree-structured diagram non-projective, which is generally to be avoided (cf. Hajičová et al., 2004, and Lopatková, 2003).

Possible CPHRs were separated from obvious trivial collocations (i.e. the type *to pay attention* from *to pay 30 dollars*, cf. Heid, 1998). To enhance the list and to determine the sorting criteria more exactly, both Czech and foreign literature on verb-noun structures was consulted.

#### 3.2 A CROSS-LINGUISTIC SURVEY OF SVC DESCRIPTIONS

Support verb constructions seem to be common in many European languages, as already noted by R. Jakobson (1932, see Jelínek, 2003, p. 50). In Czech, they had initially been believed to exemplify a negative influence from German (see Jelínek, 2003, pp. 45-46). Somewhat

ironically, it was solely in German that SVCs were first criticized from the stylistic point of view. As recently as in the 1970's, support verb constructions became a serious point of interest within German generative and transformational grammar (Rothkegel, 1973), (Persson, 1975) and in books on German as a foreign language (Helbig – Buscha, 1996; Günther – Pape, 1976). In German, SVCs have been thoroughly discussed and analyzed. In addition to that, German has affected Czech in many respects. Therefore, we took the literature on German SVCs as our point of departure, gradually extending the scope.

Helbig and Buscha, the classic German grammar for foreign learners, introduces support verbs as a special semantic class defined by the inability of the verbs to form a predicate alone (*Funktionsverben*). Support verbs have to make a cluster with a noun phrase which is then considered a part of the predicate. The noun phrase in a support verb construction is either formed by a noun in the accusative or by a prepositional phrase. The entire support verb construction (*Funktionsverbgefüge*, FVG) corresponds to a simplex lexical verb or to an adjective (with an auxiliary verb) having the same stem as the noun in the support verb construction. The nouns should be abstract noun derivations from verbs or adjectives, but never concrete nouns.

### 3.2.1 THE NOTION OF BASE AND COLLOCATE IN SVCs

Support verb constructions can also be looked upon as a collocation type. Malmgren (2002, p. 12)<sup>1</sup> describes a number of apparent support verb constructions calling them a kind of “prototypical collocations” that consist of a semantically impoverished verb and an abstract noun. The abstract noun keeps its meaning, hence it is the more stable member of the collocation – the collocational base. Its verbal collocate is generally unpredictable (Malmgren, 2002, p. 11, cf. Rothkegel, 1973, p. 39). Inspired by Melčuk's Meaning-Text-Theory (Kahane, 2003; Wanner, 1996), Malmgren finds and associates Swedish verbal collocates to the nouns by means of the lexical function Oper. Fontenelle (1992, p. 142) also claims that “support verbs roughly correspond to the type of lexical relation that can be encoded through the Oper lexical function used by Melčuk”. For examples of lexicons and lexical databases using Lexical Functions see e.g. Macleod (2002), Benson – Benson – Ilson (1997) and Polguère (2000).

The understanding of nouns as collocational bases in verb + abstract noun constructions is clearly shared by Čermák (e.g. 2003): “Abstract nouns seem to follow a few general patterns in their behaviour, which seem to be more structured, allowing for much less freedom than concrete nouns. The patterns the abstract nouns enter are determined by their function and meaning”<sup>2</sup>.

While Helbig and Buscha were struggling to identify a distinct class of “Funktionsverben”, and Baron and Herslund (1998), Rothkegel (1973) and Persson (1975, 1992) were trying to define support verb constructions by the semantic relation between the noun phrase and the verb, Fontenelle, Malmgren and Čermák focused on the noun, in full accord with the pregnantly formulated observation of Hanks (2001): “[...] it seems almost as if all the other parts of speech (verbs and function words) are little more than repetitive glue holding the names in place”.

<sup>1</sup> Malmgren's starting point is the system-oriented understanding of collocations coined especially by German linguists as Hausmann and Heid (1998, p. 302) rather than the original English contextualist approach to collocations (Malmgren, 2002, pp. 5-6).

<sup>2</sup> Though Čermák explicitly avoids the term “collocation”, using the expression “stable combinations” instead, among which “some are undoubtedly more frequent than others”.

Even in the cross-linguistic perspective, it is usually the noun that is the common denominator for the equivalent support verb constructions: “The verb [...], although often the only one that is correct and idiomatic, can seem totally arbitrary. In another language – mutatis mutandis – totally different verbs could often occur which would work as place holders; that is why prototypical collocations often cause translation problems” (Malmgren, 2002, p. 11).<sup>3</sup> Malmgren further notes that “sometimes, though far from every time, one can anticipate a sort of metaphors” in the choice of the verb. The eventual metaphors can be traced back and explained ex post, but they definitely do not prove to be predictable within one language, let alone cross-linguistically.

### 3.2.2 PRODUCTIVITY VS. LEXICALIZATION IN SVCs

Whereas traditional views emphasize that it is mostly the lexicalized units that tend to show specific syntax behaviour and, therefore, support verb constructions are to be considered as more or less lexicalized phrases, Ekberg (1989) and Dura (1997), as well as Persson (1992), concentrate on the apparent productivity of SVCs and the regular production patterns they form. Ekberg notes that many lexicalized phrases “have an almost completely or at least partly predictable meaning and new ones can be formed according to productive rules within the grammar” (Ekberg, 1989, p. 32), while Dura goes even further adding that “even the newly-formed phrases show the same syntactic restrictions as the lexicalized ones” and interpreting this phenomenon as evidence that “these restrictions rather indicate that something is meant as a lexicalization than that they are the result of lexicalization” (Dura, 1997, pp. 1-3). She considers article-less verb-noun combinations to be evidence that there is “a kind of word combination that is not controlled by the regular syntax but aims at lexical composition” and that it is thus “possible to form new phrases which can act as lexical units. The ordinary syntax is oriented at combining lexical units with obligatory grammatical categories, but there even seems to be another syntax, a syntax which allows language users to build larger conceptual units without involving the grammatical categories”.

### 3.2.3 COMMUNICATIONAL BENEFITS OF SVCs

While the first observations of support verb constructions were rather critical, Helbig and Buscha name many communicational advantages of support verb constructions, giving thereby an explanation of the extreme productivity of these constructions in the modern language. A significant feature of support verb constructions is their ability to indicate (or specify) the event structure (*Aktionsart*), (Helbig – Buscha, 1996, p. 78 and pp. 103-105). For more about event-structure modifications, see especially (Baron – Herslund, 1998 and Persson, 1975, 1992). Support verb constructions also help to fill in certain gaps in the vocabulary when no matching simplex verb exists. They make possible more general statements by means of an intransitive phrase matching a transitive simplex verb, they unify the argument structure in larger syntagms and they also make up an additional unergative form. Non fully lexicalized support verb constructions also allow for the insertion of multiple adjectival attributes and for compound noun formation, which makes them a good alternative in contexts where a simplex verb would be modified by too many adverbials. Jelínek (Jelínek, 2003, pp. 46, 48) mainly emphasizes the importance of SVCs in textual co-reference as well as in TFA.

Last but not least, Vlková (1990) studies the functional-stylistic aspects of SVCs.

<sup>3</sup> The quotations of Malmgren, Ekberg and Dura were translated from Swedish by S.C.



### 3.3 MODIFICATION OF THE CPHR-CANDIDATE LIST – RESULTING CRITERIA FOR CPHRs

As the above-mentioned literature on SVCs reveals, no universal criterion has yet been found to draw a line between CPHRs and non-CPHRs. The constraints concerning the surface structure of a SVC are obviously language-dependent and in addition they also result in scalar classifications. We agree with Persson (1992, pp. 156-157) that:

1) It is the semantic relation between the verb and the noun that makes a SVC, rather than the surface structure of the verb-noun group (see also Schroten, 2002, p. 93, and Boje, 1995, pp. 53, 145).

2) This relation could be looked upon as a kind of word formation rather than a syntactic process (see also Dura, 1997).

3) There are several types of semantic relation between verb and noun, which would result in different definitions for each type of SVC.

In order not to delay the annotation, we agreed upon a few relatively simple criteria to mark a noun as a CPHR, not all of which have to be met simultaneously. Basically, we allow for “typical” and “less typical” CPHRs. The features of CPHRs are as follows:

(i) Semantic features of the verbal and the nominal SVC component (cf. Section 4);

A support verb and a noun component make up a semantic unit, so it is usually possible to find an adequate synonymic synthetic predicate (or a copula + adjective predicate). For discussion on the effects of the choice between a synthetic predicate and a SVC on coreference relations, see Sections 5.3.1 and 5.4;

(ii) Valency features of the verbal and the nominal SVC component (cf. Section 5).

The absolute co-occurrence frequency was not considered as a criterion (cf. Malmgren, 2002, p. 14). Some kind of relative frequency information (mutual information score, log-likelihood ratio) could have been of some relevance but it was not looked at during the annotation.

## 4 SEMANTIC ASPECTS OF SVCs<sup>4</sup>

### 4.1 SUPPORT VERBS (VERBAL SVC COMPONENTS) IN PDT

#### 4.1.1 SEMANTIC BLEACHING – QUASI-MODALS AND QUASI-COPULAS

As already stated by many authors (e.g. Helbig – Buscha, 1996), support verbs are in fact lexical verbs that have to a large extent lost their lexical meaning, mainly providing the nouns with the morphological categories of verbs (which is the feature that makes them resemble a verb class, cf. Helbig – Buscha, 1996: *Funktionsverben*, and Jelínek, 2003: *operational verbs* (*operační slovesa*, p. 40)). Many students of this topic have observed that verbs, when occurring in a SVC, start to carry more abstract semantic features. Rothkegel (Rothkegel, 1973, p. 51) considers the semantic bleaching<sup>5</sup> of the verb the antipode of verbal polysemy. She shows that the meaning of a given lexical verb in SVCs neither matches any of its meanings outside SVCs, nor does it create new meanings when associated to the respective noun phrases. This, however, rather implies that the lexical verb acquires an additional, more abstract, meaning that is reserved for the verb's occurrence in SVCs, instead of just being

<sup>4</sup> Semantic classification of both the verbal and the noun component in Czech SVCs is described by Macháčková (1983, pp. 146-165).

<sup>5</sup> She quotes another author's terms, such as “das Verblassen der Merkmale bei den Verben”, “Bedeutungsentleerung”, “depletion of the designatum”.



deprived of a part of its original meaning. This observation indicates an ongoing grammaticalization process called *context-induced reinterpretation* (Heine – Claudi – Hünemeyer, 2001, p. 99) instead of speaking of mere semantic bleaching.

In PDT, SVCs which lack adequate synonymous synthetic predicates are often regarded as so-called quasi-modal verbs. As a rule, this concerns SVCs with ‘to have’: *mít právo* = *moci* (lit. ‘to have right’ = ‘can’), *mít povinnost* = *muset* (lit. ‘to have duty’ = ‘to have to’), *mít potřebu* = *chtít* (lit. ‘to have need’ = ‘to want’). Verbs of intention provide the same quality: *mít plán*, *mít tendenci* = *chtít* (lit. ‘to have plan, tendency’ = ‘to want’). A current-result copula feature<sup>6</sup> can often accompany the modality feature. If the *mít*-SVCs are regarded as duratives, the SVCs displayed below can be regarded as inchoatives and terminatives. (For more about event structure modifications, see especially Baron – Herslund, 1998; Persson, 1975, 1992, and Čermák, 1998). What is important is the fact that the support verbs often acquire quasi-copula features which were not present in their original meaning as lexical verbs. Due to the additional modification in the event structure, there is no need for any exactly matching synthetic predicate.

**Inchoative SVCs** (i.e. *začít mít*, *začít chovat*, lit. ‘to start having’):

*dát se do práce* (lit. ‘to give oneself into work’, i.e. ‘start working’), *dostat nápad* (lit. ‘to get idea’), *dostat se do styku* (‘to get in touch’), *najít odvahu* (‘to pluck up the courage’), *naskýtá se možnost udělat* (lit. ‘a possibility offers\_reflexive3thSing to do’, i.e. ‘There’s a possibility of doing’), *otevřít možnost* (lit. ‘to open a possibility’, i.e. ‘to give a possibility’), *pocítit potřebu* (i.e. ‘to get a need’), *pojmut podezření* (‘to get a suspicion’), *přistoupit k udělení cen* (lit. ‘to step to granting the awards’, i.e. ‘approach granting the awards’), *pustit se do práce* (‘set on working’), *sbírat odvahu* (‘summon up the courage’), *vzbudit touhu* (‘arouse desire’), *někomu vzniká povinnost udělat* (lit. ‘an obligation arises to sb’, i.e. ‘sb gets under obligation to do sth’).

**Terminative SVCs** (i.e. *přestat mít*, lit. ‘to stop having’):

*nenáležet* (*nenáleží mu už právo*, lit. ‘doesn’t belong him the right to do sth any more’, i.e. ‘he doesn’t have the right any more’), *nepřislušet* (*nepřisluší mu už oprávnění dělat* – lit. ‘doesn’t belong him the authorization to do sth any more’, i.e. ‘he has lost the authorization for sth’), *pozbyt odvahu* (‘to lose courage’), *přijít o možnost* (‘to forfeit the chance’), (*někomu*) *zaniká povinnost udělat* (lit. ‘sb\_DatSing expires the obligation to do sth’, i.e. ‘sb is no longer under an obligation to do sth’), *ztratit možnost* (‘to lose the possibility’), *ztratit chuť* (lit. ‘to lose the desire’, i.e. ‘not feel like doing sth any more’).

#### 4.1.2 VERBS WITH A CPHR-FRAME ONLY

In some approaches (cf. Feil, 1995), a distinction is made between the usual lexical verbs occurring in SVCs and semantically empty lexical verbs that can occur only in support verb constructions, such as *lave*, *foretage* and *gøre* (Danish, approx. ‘to make’, ‘to (under)take’ and ‘to do’). PDT has no problems with verbs that lack an “unmarked” frame, e.g. a PAT-frame, but only occur in a CPHR-frame, such as Czech *podniknout* ‘undertake’.

#### 4.2 SVC NOUN COMPONENT IN PDT

The noun phrase is generally considered the bearer of the semantic weight of the entire construction. The nouns are limited to abstract, often deverbal nouns: *rozhodnutí* ‘decision’, *otázka* ‘question’, but also non-deverbal ones (especially adjectival derivations, such as *možnost* ‘possibility’, *povinnost* ‘responsibility’, *schopnost* ‘ability’, *zodpovědnost* ‘incumbency’).

<sup>6</sup> In Czech: *fázová slovesa*, such as *začít*, *přestat*, *zůstat*, *stát se nějakým*.

and also some other types, such as *právo* 'right', *šance* 'chance', *příležitost* 'opportunity' (see also Macháčková, 1983, p. 128).

Noun components that share a support verb are often semantically related, e.g.:

- affections: *důvěra* 'trust', *něha* 'tenderness', *soucit* 'compassion', *soustrast* 'commiseration', *touha* 'desire'; professions: *funkce* 'appointment', *povolání* 'occupation', *praxe* 'practice', *profese* 'profession', *živnost* 'enterprise' (cf. also the cross-linguistic study by Schrotten, 2002);
- synonymic groups (often a Czech word matching a loanword): *kontakt* 'contact', *spojení* 'connection', *styk* 'touch', *vztah* 'relation'; *dohoda* 'agreement', *smlouva*, *kontrakt* 'contract'; *pokyn* 'instruction', *příkaz* 'command', *rozkaz* 'order'; *souhlas* 'consent', *svolení* 'approval'; *pokuta* 'fine', *sankce* 'sanction', *trest* 'penalty'; *iluze* 'illusion', *zdáání* 'impression';
- (rarely) antonymic groups: *milost* vs. *trest* ('mercy' vs. 'punishment'); *souhlas* vs. *zákaz* ('permission' vs. 'ban');
- one noun component can be associated with several support verbs that form aspect and event pairs, sometimes even synonymical groups: *dostat* – *mít* – *ztratit chuť* (lit. *to get* – *to have* – *to lose desire*, i.e. *to start feeling like* – *feel like* – *stop feeling like doing sth*).

When annotating the data, PDT annotators have to distinguish between abstract and concrete readings of nouns in context. Thus, the noun *nabídka* 'offer' in a clause like *V pondělí dostal nabídku*, lit. 'on Monday (he) got offer', will be either assigned a CPHR or a PAT: *V pondělí dostal nabídku*.CPHR = *v pondělí mu bylo něco nabídnuto* (lit. 'on Monday (he) was offered sth') vs. *V pondělí dostal nabídku*.PAT = *v pondělí obdržel dokument s nabídkou* (lit. 'on Monday (he) received document with offer').

## 5 VALENCY ASPECTS OF SVCs

Baron and Herslund (Baron – Herslund, 1998, p. 106-111) analyse the nominal structure of support verb constructions having a simplex verb match. In the traditional view, the argument structure of noun phrases is derived from the argument structure of the matching simplex verb. There are also opposite views claiming that support verb constructions inherit the argument structure of the given noun (e.g. Pedersen, 1990, p. 210).

Czech authors assume that both SVC components, the verbal and the noun component respectively, have their own valency properties (cf. esp. Macháčková, 1983, but in part also Čermák, 1974, and Jelínek, 2003). In PDT, we treat the phenomenon of valency within SVCs in the same way: both the verbal and the noun components have their own entry in the valency dictionary (in the so-called PDT-vallex, cf. Uřešová, this volume, and Hajič et al., 2003). Also, annotators of PDT have to decide if the respective complementation which has occurred in a sentence should be attached to the verb or to the noun.

The typical, transitional, but also some special (problematic) issues of the phenomenon of valency within SVCs are discussed in the following sections.

### 5.1 VALENCY OF THE VERBAL COMPONENT IN SVCs

As mentioned above, in PDT, two aspects especially were taken into account during the selection of constructions possibly treated as SVCs: (i) capturing co-referential relations in valency frames of both components of SVCs, especially those concerning grammatical co-reference (as in SVCs where the whole construction corresponds to the simplex verb representing the so-called verb of control), and (ii) SVCs in which the noun component has its original valency complementation causing, in some varieties of the word order, so-called non-projective constructions (cf. Hajičová et al., 2004, and Lopatková, 2003). Thus, the

present list of SVCs in PDT (i.e. the CPHR-Candidate List, see above, extended by some other verbs) is considerably limited by the two above-mentioned aspects. The fact that the list is not a complete register of SVCs is clearly documented by Macháčková (1983) and Čermák (1974) who present not only a richer material of abstract nouns but also a larger list of support verbs. While Čermák states that he found more than 430 verbs having the ability to function as support verbs (when the meaning does not change he regards the aspectual counterparts as the only verb, cf. Čermák, 1974, p. 299), the list of support verbs in PDT contains only about 150 items, including the aspectual counterparts of the particular verbs. The following overview of the types of valency of a verbal component in SVCs is not exhaustive (for a more detailed description, see Macháčková, 1983, p. 137ff.).

The following forms<sup>7</sup> of the noun component labelled by the functor CPHR were found in PDT:

- prepositionless accusative (these constructions represent the overwhelming majority of SVCs, e.g. *učinit rozhodnutí*, lit. ‘to make decision’);
- nominative (e.g. *zmocnilo se ho rozčilení*, lit. ‘overcame him rage’, i.e. ‘he was overcome with rage’);
- prepositionless instrumental (e.g. *hořet nenávistí*, lit. ‘to burn with hatred’);
- prepositional phrases (e.g. *přistoupit k hlasování*, lit. ‘to go up to voting’, i.e. ‘to proceed to voting’).

All studies dealing with the phenomenon of SVCs as well as data provided by PDT confirm that the prepositionless accusative is the most frequent form of the noun component in SVCs. In accordance with this observation, we will concentrate only on this type of SVCs in the following sections, calling them “SVCs with CPHR(4)”.

#### 5.1.1 FORMS AND SEMANTIC FUNCTIONS OF THE THIRD COMPLEMENTATION IN SVCs WITH CPHR(4)

When describing Czech SVCs with three complementations, Macháčková (1983, p. 139) considers the following distribution of semantic functions within SVCs: “There is mostly an agent in the subject position (resp. stimulus), the second valency position is occupied by an abstract noun, the third position is occupied by the addressee, recipient, sometimes also source, stimulus, aim of the event (action) / state.”<sup>8</sup> This description is general; it covers not only SVCs with CPHR(4) but also SVCs containing the forms mentioned above; therefore, it does not provide the forms of the possible valency slots, but rather their semantic functions.

In PDT, the third complementation in SVCs with CPHR(4) is expressed particularly by the following forms:

- prepositionless dative (e.g. *dát komu příkaz*, lit. ‘to give to-sb order’);
- prepositional phrases:
  - *od+2* ‘from+2’ (e.g. *dostat od koho příkaz*, lit. ‘to get from sb order’);
  - *z+2* ‘from+2’ (e.g. *nabýt z něčeho přesvědčení*, lit. ‘to gain conviction from sth’, i.e. ‘to come to believe that’);
  - *na+4* ‘on+4’ (e.g. *klást nároky na někoho*, lit. ‘to put demands on sb’);

<sup>7</sup> Macháčková mentions two more forms: prepositionless genitive, e.g. *dosáhnout úspěchu*, lit. ‘to reach of-success’, i.e. ‘to achieve success’, and prepositionless dative, e.g. *propadnout zoufalství*, lit. ‘to succumb to-despair’ (cf. Macháčková, 1983, p. 139).

<sup>8</sup> All quotations from Macháčková were translated from Czech by V. K.

– *v+6* ‘*in+6*’ or *u+2* ‘*at+2*’ (e.g. *budit obdiv v kom / u koho*, lit. ‘to raise admiration in sb / at sb’).

You can see that the forms extracted from PDT, at least so far as the SVCs with CPHR(4) are taken into account, are in agreement with Macháček’s description of the semantic functions of valency slots within SVCs containing three complementations:

The valency slot expressed by the prepositionless dative corresponds to the position of addressee or recipient (in PDT, it is mostly labelled by the functor ADDR (Addressee))<sup>9</sup>. The prepositional phrase *od+2* ‘*from+2*’ usually expresses the source and the prepositional phrase *z+2* ‘*from+2*’ is near the stimulus (in PDT, they are mostly labelled by the functor ORIG (Origin)). Support verbs having the third valency slot in the dative form and support verbs with the prepositional phrase *od+2* ‘*from+2*’ represent SVCs which allow for changes in voice. While SVCs with the valency slot in the dative render constructions in the active voice (*dát komu příkaz*, lit. ‘to give to-sb order’), SVCs with the valency complementation expressed by the prepositional phrase *od+2* ‘*from+2*’ can be regarded as constructions in the passive voice (e.g. *dostat od koho příkaz*, i.e. *bylo mu přikázáno*, lit. ‘to get from sb order’, i.e. ‘to be given an order (by sb)’). This understanding of the relation between the two above-mentioned types of SVCs is also shared by Macháček (1983, pp. 155–157). Various examples of SVCs allowing for changes in voice are given in Section 5.4.

The prepositional phrase *na+4* ‘*on+4*’ obviously serves several semantic functions. It is common with support verbs that, when used in a primary (i.e. non-figurative) sense, have obligatory free modification with the meaning of direction (in PDT mostly labelled by the functor DIR3), e.g. *klást něco někam*, lit. ‘to put sth somewhere’. The prepositional phrase *na+4* ‘*on+4*’ is one of the prototypical forms of the directional modification (e.g. *klást něco na něco / někoho*, lit. ‘to put sth on sth / sb’). When these verbs function as support verbs, it is not possible to express the third valency slot by an adverb, and only the prepositional phrase *na+4* ‘*on+4*’ remains (cf. *klást nároky na někoho*, lit. ‘to put (make) demands on sb’). Thus, we assign the prepositional phrase *na+4* ‘*on+4*’ within these SVCs the functor ADDR (Addressee). But there are also some other SVCs with the prepositional phrase *na+4* ‘*on+4*’. Considering the verb *obrátit* ‘to turn’, different forms of the third valency slot are possible even when the verb functions as a support verb (i.e. *obrátit pozornost na něco / k něčemu / někam*, lit. ‘to turn attention on sth / to sth / somewhere’). In this case, we assign the more general functor with the meaning of direction (DIR3) rather than the functor ADDR.

We also treat the support verbs *budit*, *vyvolat* ‘arouse, raise’ in a similar way, which in their non-figurative meaning are often accompanied by the free modification with the meaning of location. When they are used within the SVCs, the third valency slot can be expressed by several forms (*v+6* ‘*in+6*’, *u+2* ‘*at+2*’), including also an adverb (cf. *budit obdiv v kom / u koho / kde*, lit. ‘to raise admiration in sb / at sb / where’); thus we again decide on a more general functor, in this case the functor with the meaning of location (LOC).

SVCs with the third valency complementation rendering the semantic function of “aim of the event (action) / state” have not yet been found in PDT.<sup>10</sup>

<sup>9</sup> For the cases of the so-called shifting of participants, see Panevová (1980), but also Urešová (this volume).

<sup>10</sup> Moreover, the complementation with the meaning of aim (labelled by the functor AIM) would probably not have been considered to be a member of a valency frame at all.

### 5.1.2 CHANGES IN VALENCY FRAMES OF SVCS WITH CPHR(4) IN COMPARISON WITH THE NON-FIGURATIVE SENSE OF THE VERBS

The valency properties of the verb component within SVCs can not only be investigated as far as the semantic functions of particular valency slots are concerned but we can also compare the valency frames of SVCs with CPHR(4) with the non-figurative sense of these (support) verbs. Taking the latter aspect into account, the following types of valency can be differentiated:

(a) The valency behaviour of a support verb is the same as in its non-figurative sense (i.e. the support verb has the same number of valency slots and the forms of particular complementations are also the same; this type can be illustrated by the examples of SVCs quoted in the previous section).

(b) The support verb acquires a new valency slot which is not present in the valency frame of the respective verb in its non-figurative sense (the emergence of the new slot can mostly be explained by means of an analogy with the valency properties of a corresponding simplex verb or another support verb, cf. below).

(c) Sometimes a support verb can even lose a valency slot typical of the respective verb in its non-figurative sense (see below).<sup>11</sup>

*Acquiring new valency slots by means of analogy.* Some support verbs that have almost lost their lexical meaning can acquire a third valency slot, although there is no reason for this complementation in the valency frame of the respective verb in its non-figurative sense. This especially concerns verbs such as *dělat*, *udělat*, *činit*, *učinit* ('to make') and also the verbs *tvořit*, *vytvořit* ('to create / form / raise'). Thus the following SVCs exist: *(u)dělat / učinit na někoho dojem*, lit. 'to make impression on sb' (probably by analogy with the corresponding simplex verb *zapůsobit na někoho*, lit. 'to impress on sb'), *položít / klást někomu otázku*, lit. 'to put to-sb question', i.e. 'to ask sb a question' (probably by analogy with another support verb, e.g. *dát někomu otázku*, lit. 'to give to-sb question', because the respective valency slot in the valency frame of the source verb of the noun *otázka* 'question' is expressed by the genitive, cf. *otázka se koho*, i.e. 'to ask sb', so it cannot be an analogy to this simplex verb). We can also explain by means of an analogy the third valency slot of the verb *vyjádřit* 'to express' in SVCs such as *vyjádřit někomu úctu* (it is probably an analogy with the support verb *projevit* 'to express / to show', cf. *projevit někomu úctu*, lit. 'to express / show to-sb respect', because the respective valency slot in the valency frame of the source verb of the noun *úcta* 'respect' is expressed by the accusative, cf. *uctívat koho* 'to respect sb' or 'to worship sb').

*Losing valency slots.* The reverse process, i.e. losing a valency slot typical of a verb in its non-figurative sense, can also be observed, e.g. *podat výkon*, lit. 'to pass performance', i.e. 'to perform' (the valency slot with the meaning of Addressee is missing here), *dostat chuť dělat něco*, lit. 'to get liking to do sth', i.e. 'to feel like doing sth' (the valency slot with the meaning of Origin is missing here).

### 5.1.3 RECORDING OF SVCS IN PDT-VALLEX

From the above, it follows that one support verb can have several different valency frames according to the type of SVCs the concrete support verb is involved in. The following valency frames of particular support verbs illustrate the manner of recording SVCs in the PDT-vallex

<sup>11</sup> The facts described in points (b) and (c) represent one more reason why such verbs with their valency slots should be considered to be SVCs.



(the list of abstract nouns within braces provides the set of nouns found in the data depending on the particular support verb, jointly representing a SVC; the realization of forms of valency slots in the PDT-vallex is discussed by Uřešová in this volume).

The verb *dostat* ‘to get’:

ACT(.1) ORIG(*od*+2 ‘from+2’) CPHR({*rozkaz* ‘order’, *úkol* ‘task’,...}.4)

ACT(.1) CPHR({*chuť* ‘liking’, *nápad* ‘idea’,...}.4)

The verb *klást* ‘to put’:

ACT(.1) ADDR(.3) CPHR({*dotaz*, *otázka* ‘question’,...}.4)

ACT(.1) ADDR(*na*+4 ‘on+4’) CPHR({*nárok* ‘demand’, *požadavek* ‘requirement’,...}.4)

The verb *vyjádřit* ‘to express’:

ACT(.1) CPHR({*přesvědčení* ‘conviction’, *údiv* ‘surprise’, *spokojenost* ‘satisfaction’,...}.4)

ACT(.1) ADDR(.3) CPHR({*důvěra* ‘trust’, *úcta* ‘respect’, *podpora* ‘support’,...}.4)

The verb *vyvolat* ‘to cause / rouse / raise’:

ACT(.1) CPHR({*diskuse* ‘discussion’, *jednání* ‘action’, *potíž* ‘trouble’,...}.4)

ACT(.1) CPHR({*dojem* ‘impression’, *důvěra* ‘trust’, *pochybnost* ‘doubt’,...}.4) LOC(*v*+6 ‘in+6’; *u*+2 ‘at+2’;\*)

## 5.2 COMPETITION BETWEEN THE VALENCY RELATION TO THE NOUN AND TO THE SUPPORT VERB

The origin of the third valency slot within SVCs formed by verbs that have almost lost their lexical meaning (this again especially concerns verbs such as *dělat*, *udělat*, *činit*, *učinit* ‘to make’ and also the verbs *tvořit*, *vytvořit* ‘to create / form / raise’) can be explained not as an analogy with the respective simplex verb or with another support verb, but rather as a valency slot of the noun component of the SVC. The competing valency slot is interpreted as a member of the valency frame of the noun component whenever the valency relation to the noun is stronger than the relation to the support verb.

When a “competing” valency slot is expressed by a prepositional phrase, it is relatively clear that a complementation of the noun component of the SVC is concerned (e.g. *mít zájem o něco*, lit. ‘to have interest about sth’, i.e. ‘to be interested in sth’; for more examples, see Section 5.3.1).

In cases of the third valency slots expressed by prepositionless cases, deciding whether the valency slot of the verb or the valency slot of the noun is concerned is more complicated. A prepositionless genitive is always the valency slot of the noun component (e.g. *dělat rekonstrukci bytu*, lit. ‘to do reconstruction of the flat’). With a prepositionless instrumental we decide on the basis of the context and the word order, cf. the following examples: while in the construction *udělat pohyb rukou*, lit. ‘to make motion by-hand’, i.e. ‘to gesture’ we interpret the word ‘hand’ as the valency slot of the noun ‘motion’, in the construction *udělat tou rukou pohyb*, lit. ‘to make by-that hand motion’, i.e. ‘to gesture’ we expound the word ‘hand’ as the free modification of the verb ‘to make’, i.e. ‘to make by means of hand’. The clear example with the prepositionless instrumental can be exemplified by the construction *vyjádřit pohrdání něčím*, lit. ‘to express contempt with-sth’, i.e. ‘to express contempt for sth’ where ‘sth’ is the valency slot of the noun component. The most disputable examples are represented by SVCs with the third valency slot expressed by a prepositionless dative. Nevertheless, in considering

SVCs formed by nouns derived from verbs with one valency slot expressed by the dative, the valency relation to the noun is stronger than that to the support verb in these constructions, so we interpret the respective valency slot expressed by the dative as a valency complementation of the noun component, cf. the following examples:

SVC *tvořit / vytvářet / stavět překážku / bariéru něčemu*, lit. ‘to create / to form / to raise obstacle / barrier to-sth’ (it may be analogous with the valency of the simplex verb *bránit* ‘to prevent’, cf. *bránit čemu*, lit. ‘to prevent to-sth’, but also the source verb of the noun *překážka* ‘obstacle’, i.e. *překážet* ‘to hinder’, has the valency slot expressed by the dative, cf. *překážet komu*, i.e. ‘to be in sb’s way’)

(1) ... *zdražuje dopravu a vytváří překážky mezinárodnímu obchodu* (CNC, reduced)

(1’) lit. ‘(it) increases prices (of traffic) and creates obstacles to-international trade’.

SVC *dělat / činit návrh / nabídku někomu*, lit. ‘to make suggestion / offer to-sb’ (it may be analogous with the valency of the support verb *dát* ‘to give’, cf. the SVC *dát někomu návrh*, lit. ‘to give to-sb offer’, but also the source verbs of nouns *návrh / nabídka* ‘suggestion / offer’ have the valency slot expressed by the dative, cf. *navrhnout / nabídnout někomu něco*, lit. ‘to suggest / to offer to-sb sth’)

(2) ... *činí její vyhlášovatel návrh konkrétně neurčeným osobám, aby...* (CNC, reduced)

(2’) lit. ‘makes her announcer suggestion to-concretely unspecified persons to...’.

SVC *dělat / činit / učinit někomu ústupky*, lit. ‘to make concessions to-sb’ (only the source verb of the noun *ústupek* ‘concession’, i.e. *ustupovat* ‘to make-way’, exists, having its own valency slot expressed by the dative, cf. *ustupovat někomu*, lit. ‘to make-way to-sb’, i.e. ‘to compromise with sb’)

(3) *Bylo nutné učinit větší ústupky lidovcům* (CNC, reduced)

(3’) lit. ‘(It) was necessary to make bigger concessions to-members of KDU-ČSL’.

### 5.3 VALENCY OF THE NOUN COMPONENT IN SVCs

In this section, we will concentrate especially on the valency properties of the noun component within SVCs (see Section 5.3.1). Extending the scope to nominalizations of SVCs, our investigations will also take into concern the valency behaviour of nouns which “leave” their SVC and occur alone in the text (see esp. Section 5.3.2). As mentioned above, both deverbal nouns as well as non-deverbal ones can serve as the noun component of SVCs.

#### 5.3.1 ORIGINAL VALENCY COMPLEMENTATIONS OF THE NOUN COMPONENT WITHIN SVCs

According to the occurrences of SVCs found in PDT, it seems that **deverbal nouns** have their original valency complementations in the vast majority of SVCs (common differences from the valency behaviour of verbs these nouns are derived from are described in Jirsová, 1966, and Novotný, 1980, concerning especially cases when the form of a complementation changes to a prepositional phrase, e.g. *nenávidět někoho* ‘to hate sb’ vs. *nenávist k / vůči někomu* ‘hatred for sb’).

A valency complementation of the noun component within SVCs can be expressed by:

(i) a prepositionless case, e.g. *provést opravu něčeho*, lit. ‘to make repair of-sth’, *budit pocit něčeho*, lit. ‘to raise feeling of-sth’, *vyjádřit pohrdání něčím*, lit. ‘to express contempt with-sth’,



i.e. 'to express contempt for sth', *vydat pokyn někomu*, lit. 'to issue instruction to-sb', *vyhlásit rozkaz někomu*, lit. 'to pronounce order to-sb', *dělat ústupky někomu*, lit. 'to make concessions to-sb';

(ii) a prepositional phrase, e.g. *mít rozhovor s někým*, lit. 'to have conversation with sb', *vést debatu o něčem*, lit. 'to hold discussion about sth', *vznést námitku vůči někomu*, lit. 'to raise objection to sb', *podniknout krok k čemu*, lit. 'to take step to sth', *vyvést soud nad někým*, lit. 'to pronounce judgement on sb', *vytvářet tlak na někoho*, lit. 'to exert pressure on sb', *mít obavu o někoho*, lit. 'to have fear for sb', *mít vztah k někomu*, lit. 'to have relation to sb', *projevit souhlas s někým*, lit. 'to express agreement with sb', *provést útok na někoho*, lit. 'to make attack on sb', *dát se do práce na něčem*, lit. 'to set to work on sth';

(iii) an infinitive or a subordinated clause, e.g. *vydat pokyn + inf.*, lit. 'to issue instruction to + inf.', *učinit rozhodnutí, že...*, lit. 'to make decision that...';

Also some **non-verbal nouns** have, within SVCs, their original valency complementations, often acquired from the words these nouns are derived from (esp. deverbal adjectives), e.g. *věrnost někomu* 'faithfulness to-sb', *oddanost někomu* 'devotion to-sb', *zodpovědnost za něco* 'responsibility for sth', *přednost před něčím* 'preference to sth', *impuls, možnost, příležitost, šance inf. / k čemu* 'stimulus, possibility, opportunity, chance inf. / to sth'; *právo inf. / na něco* 'right inf. / to sth' (for complementations expressed by an infinitive, see below).

Macháčková (1983, p. 136) observes an interesting influence of the valency properties of the verb component on the noun component within SVCs: "When a noun serves as a noun component within a SVC, it can keep the form of its valency complementations (*mít, chovat úctu ke komu* 'to have respect to sb'), or – if the support verb has its own valency complementations – it "conforms" with the support verb. This concerns support verbs with three participants like *dát* 'to give', *poskytovat* 'to provide', *vzdát* 'to give / render', *věnovat* 'to devote', *projevit* 'to show / display', *vyslovit* 'to express'. Thus there is *důvěra ke komu*, lit. 'confidence to sb', i.e. 'confidence in sb', but *projevit, vyslovit důvěru komu*, lit. 'to express, pronounce confidence to-sb', similarly there is *péče o Jana* 'care for / of John', but *poskytnout péči Janovi*, lit. 'to provide care to-John', i.e. 'to take care of John' because the verb *poskytnout* 'to provide' has the valency *komu co* 'to-whom what'. So the expression and alignment of participants is determined especially by the support verb; if the verb has no complementation other than an abstract noun (beside its subject; translator's note), then the realization of other participants is determined by the valency properties of the noun component: *mít zalíbení v kom, čem*, lit. 'to have fancy in sb, sth'.

Sometimes both possibilities still compete as with verbs *budit, vyvolat* 'to arouse, raise'. In one case, the valency behaviour is determined by the noun component; it's a matter of a congruent and a non-congruent attribute: *budit obdiv všech*, lit. 'to raise admiration of all (people)', *budit Janův obdiv*, lit. 'to raise John's admiration'; when an original adverbial of location is concerned, then the valency behaviour is determined by the verb component: *budit v kom (u koho) obdiv*, lit. 'to raise in sb (at sb) admiration'.

The above examples of SVCs demonstrate, among other things, that in some cases an original valency complementation of the noun component cannot be expressed at all (e.g. \**Projevil Petrovi důvěru k němu / k Petrovi*, lit. 'He expressed to-Peter trust to him / to Peter', \**Poskytl Janovi péči o něj / o Jana*, lit. 'He provided to-John care of him / of John'). The question arises whether we have to consider this valency complementation to be present at least at the underlying (so-called tectogrammatical) layer of sentences in PDT, understanding it to be an

obligatory complementation of the given noun.<sup>12</sup> To maintain the consistency of the valency lexicon and data, we decided on the following solution: to restore the node for the original valency complementation of the given noun in the tectogrammatical tree (therefore the valency structure of the noun corresponds to its valency frame stored in the PDT-vallex) but to label the restored node by the special tectogrammatical lemma QCor, i.e. Quasi-Control<sup>13</sup>. In the tree, the node with the lemma QCor is connected to the respective valency complementation of the support verb by an arrow, representing in a graphic way the referential identity of the two given nodes and, therefore, the co-referential relation between them.

#### 5.3.1.1 ACTOR OF A NOUN COMPONENT OF A SVC

Also, a valency complementation with the meaning of Actor is an example of an original valency complementation of the noun component within a SVC. It is frequently the case that the subject of a support verb can be understood to be identical to the non-expressed Actor of a noun component in a SVC (for various possibilities of identity of particular valency complementations of a verbal component and a noun component, see Section 5.4). However, we will see that, even in these cases, the impossibility of expression of an Actor of a noun within a SVC deserves further discussion. Macháčková (1983, p. 135) presents constructions in which we really cannot add an Actor to the noun component of the SVC, e.g. *Jan dostal strach*, lit. 'John got fear', i.e. 'John became scared', but \**Jan<sub>i</sub> dostal Janův<sub>i</sub> strach*, lit. 'John got John's fear'. According to the method described above, in these cases we restore the new node for the Actor of the given noun in the tectogrammatical tree and we label it by the special tectogrammatical lemma QCor. However, sometimes at least, it is possible to express the Actor of the noun component by a possessive pronoun as illustrated in (4). In addition, an expression of the Actor of a noun is possible in cases when this complementation and the subject of the support verb are not identical. The subject of the verb and another valency complementation can be identical depending on the particular SVC, cf. Macháčková's example *budit Janův obdiv* 'to raise John's admiration' vs. (5).

(4) *Petr znovu položil Janovi svoji.ACT otázku.*

(4') lit. 'Peter again put to-John his.ACT question', i.e. 'Peter again asked John (his) question';

(5) *Chci obrátit vaši.ACT pozornost na osudy oněch lidí.* (CNC, reduced)

(5') lit. '(I) want to turn your.ACT attention to life-stories of-those people'.

Our further observations demonstrate that the Actor can also be expressed by a possessive pronoun within SVCs which consist of a support verb and a non-deverbal noun. We especially want to quote those non-deverbal nouns which are not usually considered to have an Actor but only their original valency complementation, expressed mostly by a prepositionless

<sup>12</sup> In the cases of deletion in the surface shape of the sentence, nodes are introduced into the tectogrammatical tree to "recover" a deleted word.

<sup>13</sup> The name of the tectogrammatical lemma QCor indicates similarity to the subject of an infinitive modifying a verb of control which is labelled by the tectogrammatical lemma Cor. The connection between the two types of deletion mentioned consists in the impossibility of an overt expression of the deleted node in the surface shape of the sentence, and both types of deletion also represent constructions with grammatical co-reference. For more details on capturing co-referential relations in PDT see Kučová – Kolářová – Žabokrtský – Pajas – Čulo (2003), for the treatment of constructions with verbs of control in PDT see Panevová – Řezníčková – Uřešová (2002).

genitive<sup>14</sup>, e.g. *příklad čeho*.PAT ‘an example of-sth’, *verze čeho*.PAT ‘a version of-sth’, *alternativa čeho / čemu / k čemu*.PAT ‘an alternative of-sth / to-sth / to sth’, *varianta čeho*.PAT ‘a variant of-sth’, *cesta k řešení*.PAT, lit. ‘the road to solution’. It seems that when such a noun serves as a noun component within a SVC, it acquires its Actor simply by means of the connection with a support verb, cf. (6), (7), and (8).

(6) *Petr má nějakou (svoji.ACT) alternativu k vašemu řešení*, lit. ‘Peter has some (his.ACT) alternative to your solution’;

(7) *Petr má nějakou (svoji.ACT) verzi řešení toho problému*, lit. ‘Peter has some (his.ACT) version of-solution of-that problem’;

(8) *Petr má (svoji.ACT) zvláštní strategii*, lit. ‘Peter has (his.ACT) strange strategy’.

The valency complementation with the meaning of Actor can be recognized clearly in constructions with those non-deverbal nouns which “leave” their SVC and occur alone in the text<sup>15</sup>, see (9), (10), and (11).

(9) *Petrova.ACT alternativa řešení.PAT problému je jistě výhodnější*, lit. ‘Peter’s.ACT alternative of-solution of-problem is surely more-favourable’;

(10) *Petrova.ACT verze řešení.PAT problému je lepší*, lit. ‘Peter’s.ACT version of-solution of-problem is better’;

(11) *Petrova.ACT strategie je opravdu zvláštní*, lit. ‘Peter’s.ACT strategy is really strange’.

Some SVCs consisting of a non-deverbal noun and the support verb *mít* ‘to have’ correspond to simplex modal verbs. Then the non-deverbal nouns have the Actor and they combine with an infinitive which is typical of modal verbs. It is interesting also that the valency complementation of the noun component expressed by an infinitive can, within the “modal SVCs”, be substituted by a complementation expressed by a prepositional phrase, e.g. *na+4* or *k+3* (‘to sth’), which is not admissible with modal verbs.

(12) *Petr má šanci vyhrát*.PAT, lit. ‘Peter has chance to win.PAT’;

(13) *Petr má šanci na výhru*.PAT, lit. ‘Peter has chance for victory.PAT’;

(14) *Petr má právo volit*.PAT, lit. ‘Peter has right to vote.PAT’, i.e. ‘Peter is entitled to vote’;

(15) *Petr má právo na vlastní volbu*.PAT, lit. ‘Peter has right for his-own choice.PAT’;

(16) *Každý má své.ACT nezadatelné právo volit*.PAT, lit. ‘Everyone has his.ACT inalienable right to vote.PAT’;

(17) *Petr má příležitost se zamyslet*.PAT nad novou situací, lit. ‘Peter has opportunity to think.PAT about new situation’;

(18) *Petr má příležitost k zamyslení*.PAT, lit. ‘Peter has opportunity to thinking.PAT’;

(19) *Petr má úkol připravit*.PAT občerstvení, lit. ‘Peter has task to prepare refreshment.PAT’;

(20) *Každý má nějaký svůj.ACT úkol*, lit. ‘Everyone has some his.ACT task’.

The valency complementation with the meaning of Actor is again very common in constructions with the non-deverbal nouns which occur in the text without their “modal SVC”, cf. (21), (22), (23), and (24).<sup>16</sup>

(21) *Petrova.ACT šance najít*.PAT zaměstnání tím výrazně vzrostla, lit. ‘Peter’s.ACT chance to find.PAT job rapidly increased’;

<sup>14</sup> In PDT, this valency complementation is labelled in most cases by the functor PAT (Patient).

<sup>15</sup> “Becoming independent” can be understood also as a type of nominalization of the given SVC.

<sup>16</sup> It is, of course, also possible to express the Actor by the genitive form in these constructions.

- (22) *Petrovou*.ACT *povinností je přijít včas*, lit. ‘Peter’s.ACT duty is to come in time’;  
 (23) *Petrovo*.ACT *právo se odvolat*.PAT *mu nikdo nemůže upřít*, lit. ‘Peter’s.ACT right to appeal.PAT him nobody can deny’;  
 (24) *Petrův*.ACT *úkol připravit*.PAT *občerstvení se zdál být snadný*, lit. ‘Peter’s.ACT task to prepare.PAT refreshment seemed to-be easy’.

In connection with the issues of the Actor of nouns within SVCs, Macháčková (1983, p. 135) mentions also the “ability” of this valency complementation to even become the subject of the sentence: “But only connection of the deverbal noun with the support verb in a finite form allows the Actor to become the subject of the sentence in a similar vein as the Actor of the source verb: *Zemědělci osévají půdu.*, lit. ‘Farmers sow ground’ – *Zemědělci provádějí osev půdy.*, lit. ‘Farmers carry out sowing of ground’.” Fillmore, Johnson and Petruck (2003, p. 244) also highlight this phenomenon in English SVCs and describe how they treat it within the framework of the project called FrameNet: “Certain semantically neutral verbs can turn an event noun or a state noun into a verb phrase-like predicate and allow for the expression of a frame element as their subjects. We call such verbs support verbs. For example, both sentences in (13) report on the same event, that of deciding something and (13)(b) is not about an event of making. We want to record the fact that the noun phrase *the committee* instantiates the same frame element in both sentences, and recognizing the role of the support verb *make* allows us to do so.

- (13) a. *The committee decided to convene again next month.*  
 b. *The committee made a decision to convene again next month.*”

Atkins, Fillmore and Johnson (2003, p. 270) consider the subject of the sentence formed by a SVC to be both the grammatical subject of the support verb and the “logical” subject of the noun component. These authors also differentiate between an internal and an external realization of the frame element (i.e. the valency complementation) of the noun component; while the Actor expressed by a possessive pronoun or adjective is regarded as the internal realization of the valency complementation of the noun, its external realization is represented by the grammatical subject of the support verb (cf. Atkins – Fillmore – Johnson, 2003, p. 275).

In PDT, the subject of the sentence is recorded as depending on the support verb. However, in order to indicate the fact that the subject of the verb and the Actor of the noun are identical, we use the method described above; that is, we restore the node for the Actor of the noun (with the tectogrammatical lemma QCor) and then we capture the respective co-referential relation between the two nodes by the arrow.

In addition, Macháčková points out that the valency complementation of the noun component which is not the Actor (so it is usually the Patient) is often deleted within SVCs, e.g. *údržbář opravil vodovod* ‘the service engineer repaired water main’, but *údržbář provedl opravu* ‘service engineer made repair’. This type of deletion is not usually possible within constructions with the simplex verb, cf. *\*údržbář opravil* ‘service engineer repaired’ (see Macháčková, 1983, p. 135). In PDT, the node for the Patient of the noun is restored in such SVCs. Nevertheless, due to the fact that this node is not identical with any participant of the support verb and can only be identified from the context, the node is not labelled by the tectogrammatical lemma QCor, but by another lemma corresponding to the respective co-referential relation (for more information about capturing co-referential relations in PDT, see Kučová – Kolářová – Žabokrtský – Pajas – Čulo, 2003).

### 5.3.2 INHERITANCE OF VALENCY COMPLEMENTATIONS FROM THE VERBAL COMPONENT OF A SVC

Baron and Herslund (1998) suggest that it is the support verb constructions that provide the noun phrases with an argument structure which the noun phrases then inherit when they occur alone. Baron and Herslund “regard such nominals, noun phrases as compounds, as reduced clauses which exhibit the same argument structure as a clause” (1998, p. 106) and support verb constructions as “transitional forms between clauses with simplex verbs and complex nominals”. They argue by means of a transformation test that nominal constructions have both semantic and syntactic properties in common with the support verb construction which they do not share with the simplex verb (1998, p. 107).

In Czech, there also exist constructions formed by an original noun component of a SVC which occurs in the text without its support verb but which inherits some of its valency complementation<sup>17</sup>. With deverbal nouns, the respective valency position is present in its valency frame. However, the form of the valency complementation does not correspond to the form of the respective valency slot of the verb the noun is derived from, but to the form of the respective valency slot of the support verb. Non-deverbal nouns inherit from support verbs not only the form of the valency complementation but the whole valency position. The valency complementation inherited from the verbal component of a SVC concerns the third valency complementation of support verbs described in Section 5.1.1, and is rendered esp. by one of the two following forms:

- (i) prepositionless dative;
- (ii) prepositional phrase *od+2* ‘from+2’.

#### 5.3.2.1 NOMINAL CONSTRUCTIONS WITH THE VALENCY COMPLEMENTATION IN PREPOSITIONLESS DATIVE

*Deverbal nouns.* The influence of the participant of a support verb (expressed by a prepositionless dative) on the valency behaviour of deverbal-noun components is most transparent in constructions with nouns derived from verbs with a participant expressed by a prepositionless accusative (e.g. *podpora* ‘support’, *pochvala* ‘praise’, *informace* ‘information’, *podnět* ‘impulse / impulsion’, *uznání* ‘appreciation’, *zpráva* ‘message’) or genitive (e.g. *otázka* / *dotaz* ‘question’). Typical changes of surface expressions of valency complementations of verbs within the process of nominalization are described by Karlík and Nübler (1998). According to them, in valency frames of nouns denoting action, the original accusative form changes to genitive, and the genitive does not change. It seems that at least some of the nouns mentioned above do not allow for the expression of the valency complementation by the genitive form at all, e.g. \**informace někoho.ADDR* ‘information of-sb’, \**dotaz někoho.ADDR* ‘question of-sb’, but examples of nouns with the complementation in the prepositionless dative occur. Some of the nouns allow for both the genitive and dative forms of the complementation (e.g. *pochvala někoho / něčeho.PAT* ‘praise of-sb / sth’ or *pochvala někomu.PAT* ‘praise to-sb’; *podpora někoho / něčeho.PAT* ‘support of-sb / sth’ or *podpora někomu / něčemu.PAT* ‘support to-sb / sth’)<sup>18</sup>. The question arose as to the origin of the dative form, and the influence of the third participant of a support verb provides one of

<sup>17</sup> As mentioned above, “becoming independent” can be understood also as a type of nominalization of the given SVC.

<sup>18</sup> Possibility / impossibility of expression of the complementation by a form of prepositionless genitive with other mentioned nouns is discussed in Kolářová (in prep.).



the possible explanations. The following examples illustrate the notional process of inheriting the dative form from the support verb: e.g. *pochválit někoho* 'to praise sb', but *udělit někomu pochvalu*, lit. 'to award to-sb a praise' → *pochvala někomu* 'a praise to-sb'; *informovat někoho* 'to inform sb', but *dát / poskytnout někomu informaci*, lit. 'to give / provide to-sb information' → *informace někomu* 'information to-sb'; *podnítit někoho* 'to stimulate sb', but *dát někomu podnět*, lit. 'to give to-sb impulse' → *podnět někomu* 'impulse to-sb'; *uznávat někoho* 'to appreciate sb', but *vyjádřit někomu uznání*, lit. 'to express to-sb appreciation' → *uznání někomu* 'appreciation to-sb'; *podporovat někoho* 'to support sb', but *vyjádřit někomu podporu*, lit. 'to express to-sb support' → *podpora někomu* 'support to-sb'; *otázát / dotázát se někoho* 'to ask sb', but *dát / položit někomu otázku / dotaz*, lit. 'to give to-sb a question' → *otázka / dotaz někomu* 'a question to-sb'. Constructions with nouns modified by the valency complementation in the dative can be documented by examples from CNC, cf. (25), (26), (27) and (28); information about their absolute and relative frequency in CNC and PDT is also given in Kolářová (in prep.). Although Macháčková (1983) does not deal with the valency behaviour of nouns leaving their SVC, her insights support the above idea: "While simplex verbs are modified by an Addressee expressed by the prepositionless dative (*příkázat komu* co, lit. 'to order to-sb sth', i.e. 'to order sb to do sth') but sometimes also by the prepositionless genitive (*ptát se koho* 'to ask sb'), prepositionless accusative (*informovat koho* 'to inform sb') or an attribute (*souhlasím s tvou cestou* 'I agree to your journey'), within SVCs the Addressee is expressed first of all by the prepositionless dative: *Dal jim rady, svolení, informace, otázky*, lit. 'He gave them suggestions, permission, information, questions'" (p. 153). However, there are some nouns derived from verbs with a participant expressed by a prepositionless accusative that allow for the expression of the same participant by the prepositionless dative (e.g. *prosba* 'request' / *výzva* 'appeal' / *varování* 'warning' *někomu* 'to-sb', cf. also (29)), although this form is not influenced by any participant of the support verb (e.g. *dělat propagaci*, lit. 'to make promotion', i.e. 'to promote', *mít prosbu*, lit. 'to have a request', *učinit výzvu*, lit. 'to make an appeal', *vyslovit varování*, lit. 'to express / pronounce warning'). This phenomenon yields untypical changes of surface expressions of the valency complementation of the verbs within the process of nominalization, i.e. Acc → Dat or Gen → Dat<sup>19</sup>. Again, some of the nouns also allow for the expression of the complementation by the prepositionless genitive form, e.g. *systém varování obyvatelstva*.ADDR *v okolí Jaderné elektrárny Dukovany* 'the system of-warning of-population.ADDR in the neighbourhood of the Nuclear power station Dukovany'.

To date, we have found about thirty nouns which allow for complementation in the dative corresponding to verbal valency complementation in the accusative or genitive.

(25) *Psychologicky vhodná byla jeho.ACT závěrečná otázka panu Ježkovi.ADDR* (CNC, reduced)

(25') lit. 'Psychologically suitable was his concluding question to-Mr Ježek.ADDR', i.e. 'His concluding question to Mr Ježek was psychologically suitable';

(26) *Operativní informace uživatelům.ADDR knihovny o mimořádných situacích.PAT v knihovně svědčí o...* (CNC, reduced)

(26') lit. 'Operative information to-users.ADDR of-library about extraordinary situations. PAT in library manifests about...';

<sup>19</sup> Examples of deverbal nouns with the complementation in the dative which is not inherited from the verbs they are derived from nor influenced by participants of a support verb are more precisely illustrated in Kolářová (in prep.).

(27) *Situace byla podnětem Raiffeisenovi.ADDR k založení.PAT místních družstev.* (CNC, reduced)

(27') lit. 'Situation was impulsion to-Raiffeisen.ADDR to establishment.PAT of-local associations';

(28) *Dva dny nato neoficiální posel Wendell Wilkie přijel do Anglie s osobní zprávou Winstonu Churchillovi.ADDR od prezidenta.ACT Roosevelta.* (CNC, reduced)

(28') lit. 'Two days after-that unofficial envoy Wendell Wilkie came to England with personal message to-Winston Churchill.ADDR from president Roosevelt.ACT';

(29) *Jako poznámku uvádíme prosbu autorům.ADDR píšícím na počítači, aby pečlivě dbali.* PAT na rozlišování písmene O od čísla 0. (CNC)

(29') lit. 'As note we present request to-authors.ADDR writing on PC to carefully mind.PAT distinguishing letter O from numeral 0.'

*Non-deverbal nouns.* Some non-deverbal nouns can also occur with their valency complementation in the dative form influenced by the third participant of the support verb, esp. the support verbs *dát* 'to give' or *udělit* 'to award' (e.g. *důtka* 'admonishment', *políček* 'slap', *pohlavek* 'slap', *pokuta* 'fine / penalty', *ultimátum* 'ultimatum'). Constructions with the nouns modified by the complementation in the dative are documented by examples found in PDT or CNC, cf. (30), (31) and (32).

(30) *Odložení jeho ratifikace si hierarchie vykládá jako políček polskému papeži.PAT od polského parlamentu.ACT* (PDT)

(30') lit. 'Postponement of-his ratification hierarchy interprets as slap to-Polish pope.PAT from Polish Parliament.ACT';

(31) *Vedoucí má podepsat návrh na pokutu Zemědělskému družstvu.PAT Kosova Hora za znečištění.CAUS vody v Sedlčanech.* (CNC, reduced)

(31') lit. 'Boss has-to sign draft of fine to-Collective farm.PAT Kosova Hora for contamination. CAUS of-water in Sedlčany';

(32) *Následovalo ultimátum vládě.ADDR, aby zajistila.PAT návrat země k plně sekulárnímu státu.* (CNC, reduced)

(32') lit. '(There) followed ultimatum to-government.ADDR to arrange.PAT regress of-country to fully secular state.'

### 5.3.2.2 NOMINAL CONSTRUCTIONS WITH THE VALENCY COMPLEMENTATION EXPRESSED BY THE PREPOSITIONAL PHRASE OD+2 'FROM+2'

Inheriting the valency complementation expressed by the prepositional phrase *od+2* 'from+2' is very frequent, although there are not many support verbs with the third valency slot expressed by this form (e.g. *dostat* 'to get', *získat* 'to obtain'). While the valency complementation of the support verb is labelled by the functor ORIG (Origin; e.g. *somebody.ACT got from secretary.ORIG affirmation.CPHR*), it gets the meaning of Actor with deverbal nouns (e.g. *affirmation from secretary.ACT*, i.e. *secretary.ACT assured*). In a similar vein, we also label it by the functor ACT with non-deverbal nouns. Constructions with the nouns modified by the valency complementation in the form *od+2* 'from+2' inherited from the support verb are documented by examples from PDT or CNC (for deverbal nouns cf. (33), (34), and (35), for non-deverbal nouns cf. (36), (37), (38)).

(33) *Nedávno jsme zde slyšeli velice pozitivní ujištění od ministra.ACT zahraničních věcí USA...* (CNC, reduced)

(33') lit. 'Recently (we) here heard very positive affirmation from secretary.ACT of state of-USA...';



- (34) *Takový byl alespoň slib od okresní nemocnice*.ACT (CNC)  
 (34') *'That was at least the promise from the regional hospital*.ACT';  
 (35) *Mezitím se z vysílačky ve voze ozývají rozkazy od dispečera*.ACT pro všechny řidiče.  
 (CNC)  
 (35') lit. *'In-meantime from walkie-talkie in carriage are-heard orders from dispatcher*.ACT  
*for all drivers*;  
 (36) *Theresa Weldová zařadila do svého programu salchow, což jí vyneslo důtku od rozhodčích*.  
 ACT (CNC, reduced)  
 (36') lit. *'Theresa Weld included to her program salchow which her earned reprehension from*  
*referees*.ACT';  
 (37) *Měl by si s sebou vzít dostatek peněz na pokuty od dopravní policie*.ACT (CNC,  
 reduced)  
 (37') lit. *'(He) had with him to take enough money for fines from traffic police*.ACT';  
 (38) *Petice byla ultimátem od rodičů*.ACT (PDT, reduced)  
 (38') *'The petition was the ultimatum from the parents*.ACT'.

Nevertheless, not all valency complementations expressed by the form *od+2* 'from+2' modifying nouns denoting action can be interpreted as a result of inheriting the valency complementation from a support verb. Sometimes there is no support for this form even in the valency frame of the verb the noun is derived from, e.g. *\*odprodat od+2* 'to sell from+2', but *odprodej od+2* 'sale from+2', cf. (39).

(39) *M. Zeman navrhl možnost jeho*.PAT *odprodeje* *od státu*.ACT *židovským obcím*.ADDR  
 (CNC)

(39') *'M. Zeman suggested possibility of its*.PAT *sale* *from state*.ACT *to Jewish communities*.  
 ADDR'

It follows from our observations that non-deverbal nouns which serve as a noun component within SVCs can, in addition to their original valency complementations, also have the valency position with the meaning of Actor as well as another valency slot inherited from the support verb. In this sense, they can be considered to be equal to deverbal nouns<sup>20</sup> and, moreover, they should be treated in the valency dictionary in a similar way to deverbal nouns.

#### 5.4 SHARING OF VALENCY COMPLEMENTATIONS OF THE VERBAL AND THE NOUN COMPONENT OF SVCs

As stated above, some valency complementations of the verbal as well as the noun component of a SVC can be referentially identical. In other words, the verbal and the noun component share some valency complementation. The form of the complementation is equal (e.g. prepositionless dative, cf. *poskytnout pomoc Petrovi*, lit. 'to provide help to-Peter', and also *pomoc Petrovi*, lit. 'help to-John') or different (e.g. *Janův obdiv* 'John's admiration' vs. *budit obdiv v Janovi* 'to raise admiration in John', *péče o Jana* 'care of John' vs. *poskytnout péči Janovi*, lit. 'to provide care to-John', i.e. 'to take care of John'). The semantic function of the complementation may also be the same or different. Actors of both components are shared in most SVCs, but

<sup>20</sup> Even new verbs can be derived from the non-deverbal nouns, e.g. *dát pokutu*, lit. 'to give fine' → *pokutovat* 'to fine', *dát pohlavek*, lit. 'to give slap' → *zpohlavkovat* 'to slap', cf. also Čermák, 1974, p. 299.

other complementations can also be concerned, e.g. Addressee. In PDT, the shared valency complementation which is not present in the surface shape of the sentence is restored (it concerns esp. a valency complementation of the noun component) and labelled by the tectogrammatical lemma QCor. Then the node is connected with the shared valency complementation of the support verb by an arrow representing graphically the referential identity of the two given nodes and therefore the co-referential relation between them.

The following types of sharing valency complementations can be distinguished:

**(a) SVCs corresponding to constructions with the respective simplex verb in the active voice**

**(ai)** SVCs in which the ACT of the noun component and the ACT of the verbal component are identical.

This group contains the overwhelming majority of SVSc which can be represented esp. by the so-called quasi-modal verbs (e.g. *mít právo*, lit. 'to have right', *mít povinnost*, lit. 'to have duty', *mít potřebu*, lit. 'to have need'), verbs of intention (e.g. *mít plán* 'to have plan', *mít tendenci* 'to have tendency'), inchoative SVCs (e.g. *dát se do práce*, lit. 'to give oneself into work', *najít odvahu*, lit. 'to find courage', *pojmut podezření*, lit. 'to entertain suspicion'), terminative SVCs (e.g. *pozbyt odvahu*, lit. 'to lose courage', *přijít o možnost*, lit. 'to forfeit chance') and lot of other SVCs such as e.g. *učinit rozhodnutí*, lit. 'to make decision', *věnovat pozornost*, lit. 'to devote attention', *projevit zájem*, lit. 'to express interest', *provést omezení*, lit. 'to make restriction'.

**(aia)** SVCs in which the ACT of the noun component and the ADDR (or another valency complementation of the verbal component which is not the ACT) are identical, e.g.: *dát možnost*, lit. 'to give possibility', *ukládat povinnost*, lit. 'to give duty', *vzbudit (v někom) dojem*, lit. 'to raise (in sb) impression'.

**(aiii)** SVCs in which the ACT of the noun component and the ACT of the verbal component as well as the ADDR (or another valency complementation) of the noun component and the ADDR of the verbal component are identical, e.g.: *dát příkaz*, lit. 'to give order', *dát radu*, lit. 'to give advice', *klást otázku*, lit. 'to put question', *udělit pochvalu*, lit. 'to award praise', *poskytnout pomoc*, lit. 'to provide help'.

**(b) SVCs corresponding to constructions with the respective simplex verb in the passive voice**

**(bi)** SVCs in which the ACT of the noun component and the ACT of the verbal component are identical, e.g.: *Petr dostal možnost přijít*, lit. 'Peter got possibility to come' = *Petrovi bylo umožněno přijít* 'Peter was allowed to come'; *Petr získal možnost pracovat*, lit. 'Peter obtained possibility to work' = *Petrovi bylo umožněno pracovat* 'Peter was allowed to work'.

**(bii)** SVCs in which the ACT of the noun component and the ORIG of the verbal component as well as the ADDR (or another valency complementation) of the noun component and the ACT of the verbal component are identical, e.g.: *Petr dostal (od šéfa.ORIG) příkaz přijít*, lit. 'Peter got (from boss) order to come' = *Petrovi bylo (šéfem.ACT) přikázáno přijít*, lit. 'Peter was (by boss) ordered to come'; *Petr dostal (od šéfa.ORIG) pochvalu*, lit. 'Peter got (from boss) praise' = *Petr byl pochválen (šéfem.ACT)*, lit. 'Peter was praised (by boss)'.

## 6 CONCLUDING REMARKS

There is no doubt that SVCs represent a very complicated, complex linguistic phenomenon and an investigation of this problem involves many particular aspects. We have only touched

on the two of them dealing with the semantic and valency properties of SVCs. Such issues as word order and TFA within SVCs have been left out and merit further discussion. We have outlined the basic principles of the annotation of SVCs in the tectogrammatical tree structure of PDT and presented the method of their recording in the PDT-vallex. Real examples from CNC and PDT illustrate the fact that the noun component of a SVC, the non-deverbal as well as the deverbal one, can, to a large degree, be influenced by the valency properties of the verbal component. More inquiries into the issues of the process of nominalization of SVCs, including also the valency behaviour of adjectives derived from support verbs, would probably yield further interesting observations.

## REFERENCES

- ATKINS, S. – FILLMORE, CH. J. – JOHNSON, CH. R. (2003): Lexicographic Relevance: Selecting Information from Corpus Evidence. In: *FrameNet and Frame Semantics. International Journal of Lexicography* (Special Issue, Guest Editor: T. Fontenelle), volume 16, 2003. pp. 251-280.
- BARON, I. – HERSLUND, M. (1998): Support Verb Constructions as Predicate Formation. In: *The Structure of the Lexicon in Functional Grammar*. Eds. H. Olbertz, K. Hengeveld, J. S. García. Amsterdam/Philadelphia. pp. 99-116.
- BENSON, M. – BENSON, E. – ILSON, R. (1997): *The BBI Dictionary of English Word Combinations*. Amsterdam-Philadelphia: John Benjamins.
- BOJE, F. (1995): Hvor finder man 'finde anvendelse'? In: *Nordiske Studier i Leksikografi. Rapport fra Konferanse om leksikografi i Norden*, Reykjavík 7.-10. juni 1995. Eds. Ásta Svavarsdóttir, Guðrún Kvaran, Jón Hilmar Jónsson. Reykjavík. Skrifter utgitt av Nordiske forening for leksikografi, Skrift nr. 3. pp. 51-68.
- BRAASCH, A. – OLSEN, S. (2000): Formalised Representation of Collocations in a Danish Computational Lexicon. In *The Ninth EURALEX International Congress, Proceedings, Vol. II*, Stuttgart. pp. 475-488.
- BRAASCH, A. – OLSEN, S. (2000): Towards a Strategy for a Representation of Collocations – Extending the Danish PAROLE Lexicon. In *Second International Conference on Language Resources and Evaluation, Proceedings, Vol. II*, Athens. pp. 1009-1064.
- ČERMÁK, F. (2003): Abstract Nouns Collocations: Their Nature in a Parallel English-Czech Corpus. In: *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Birmingham: University of Birmingham Press.
- ČERMÁK, F. (1998): Linguistic Units and Text Entities: Theory and Practice. In: *Actes EURALEX'98 Proceedings*. Eds. Th. Fontenelle, Ph. Hilgsmann, A. Michiels, A. Moulin, S. Theissen. Université de Liège, Liège. pp. 281-290.
- ČERMÁK, F. (1974): Víceslovná pojmenování typu verbum – substantivum v češtině (Příspěvek k syntagmatice tzv. abstrakt). In: *Slovo a slovesnost*, 4, 35. pp. 287-306.
- DURA, E. (1997): Substantiv och stödverb. Göteborg: Göteborgs universitet. Meddelanden från Institutionen för Svenska Språket 18.
- EKBERG, L. (1987): Gå till anfall och falla i sömn. En strukturell och funktionell beskrivning av abstrakta övergångsfaser. Lund: Lund University Press. Lundastudier i nordisk språkvetenskap A 43.
- FEIL, R. (1995): Funktionsverber i det danske sprog. In: *Nordiske Studier i Leksikografi. Rapport fra Konferanse om leksikografi i Norden*, Reykjavík 7.-10. juni 1995. Eds. Ásta Svavarsdóttir, Guðrún Kvaran, Jón Hilmar Jónsson. Reykjavík. Skrifter utgitt av Nordiske forening for leksikografi, Skrift nr. 3. pp. 137-148.
- FILLMORE, CH. J. – JOHNSON, CH. R. – PETRUCK, M. R. L. (2003): Background to FrameNet. In: *FrameNet and Frame Semantics. International Journal of Lexicography* (Special Issue, Guest Editor: T. Fontenelle), volume 16, 2003. pp. 235-250.

- FONTENELLE, T. (1993): Using a bilingual computerized dictionary to retrieve support verbs and combinatorial information. In: *Acta Linguistica Hungarica*, 41 (1-4). pp. 109-121.
- FONTENELLE, T. (1992): Co-occurrence Knowledge, Support verbs and Machine Readable Dictionaries. In: *Papers in Computational Lexicography, COMPLEX'92*, Budapest. pp. 137-145.
- Günther, H. – Pape, S. (1976): Funktionsverbgefüge als Problem der Beschreibung komplexer Verben in der Valenztheorie. In: *Untersuchungen zur Verbvalenz: eine Dokumentation über die Arbeit an einem deutschen Valenzlexikon*. Ed. Helmut Schumacher. Tübingen: Narr. Forschungsberichte/Institut für deutsche Sprache Mannheim. pp. 92-128.
- HAJIČ, J. – PANEVOVÁ, J. – UŘEŠOVÁ, Z. – BÉMOVÁ, A. – KOLÁŘOVÁ, V. – PAJAS, P. (2003): PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*. Växjö, Sweden, November 14 – 15, 2003. Eds. J. Nivre, E. Hinrichs. Växjö University Press. pp. 57-68.
- HAJIČOVÁ, E. – HAVELKA, J. – SGALL, P. – VESELÁ, K. – ZEMAN, D. (2004): Issues of Projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, 81. pp. 5-22.
- HANKS, P. (2001): The Probable and the Possible: Lexicography in the Age of Internet. In: *Asialex 2001 Proceedings*, Seoul, Korea. Ed. Lee Sangsup.
- HEID, U. (1998): Towards a corpus-based dictionary of German noun-verb Collocations. In: *Actes EURALEX'98 Proceedings*. Eds. Thierry Fontenelle, Philippe Hilligsmann, Archibald Michiels, André Moulin, Siegfried Theissen. Liège: Université de Liège, Départements d'anglais et de néerlandais. Vol. I. pp. 301-312.
- HEINE, B. – CLAUDI, U. – HÜNNEMEYER, F. (2001): *Grammaticalization. A conceptual framework*. Chicago: University of Chicago Press.
- HELBIG, G. – BUSCHA, J. (1996): *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Leipzig: Verlag Enzyklopädie.
- JELÍNEK, M. (2003): O verbonominálních spojení ve spisovné češtině. In: *Přednášky a besedy z XXXVI. běhu LŠSS*. MU Brno. pp. 37-51.
- JIRSOVÁ, A. (1966): Vazby u dějových podstatných jmen označujících duševní projevy. In: *Naše řeč*, 49. pp. 73-81.
- KAHANE, S. (2003): The Meaning-Text Theory. In *Dependency and Valency. An International Handbook on Contemporary Research*. Berlin: De Gruyter.
- KARLÍK, P. – NÜBLER, N. (1998): Poznámky k nominalizaci v češtině. In: *Slovo a slovesnost*, 59. pp. 105-112.
- KOLÁŘOVÁ, V. (in prep.): Valence deverbativních substantiv v češtině. ÚFAL MFF UK. Manuscript of PhD thesis. Supervised by Jarmila Panevová.
- KUČOVÁ, L. – KOLÁŘOVÁ, V. – ŽABOKRTSKÝ, Z. – PAJAS, P. – ČULO, O. (2003): Anotování koreference v Pražském závislostním korpusu. MFF UK, Prague, TR-2003-19.
- LOPÁTKOVÁ, M. (2003): O homonymii předložkových skupin v češtině (Co umí počítač?). Karolinum, Praha.
- MACLEOD, C. (2002): Lexical Annotation for Multi-word Entries Containing Nominalizations. In *Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002)*; Las Palmas, Canary Islands, Spain. pp. 943-948.
- MACHÁČKOVÁ, E. (1983): Analytické predikáty. Substantivní názvy dějů a statických situací ve spojení s funkčními slovesy. *Jazykovědné aktuality*, 10, 1983, volumes 3 and 4. pp. 122-176.
- MALMGREN, S.-G. (2002): *Begå eller ta självmord? Om svenska kollokationer och deras förändringsbenägenhet 1800-2000*. Göteborg: Göteborgs universitet. Institutionen för svenska språket. Rapporten från ORDAT.
- NOVOTNÝ, J. (1980): Valence dějových substantiv v češtině. In: *Sb. pedagogické fakulty v Ústí nad Labem*. SPN, Praha.
- PANEVOVÁ, J. (1980): *Formy a funkce ve stavbě české věty*. Praha, Academia.
- PANEVOVÁ, J. – ŘEZNÍČKOVÁ, V. – UŘEŠOVÁ, Z. (2002): The theory of control applied to the Prague Dependency Treebank (PDT). In: *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks*. Università di Venezia, 2002. pp. 175-180.

- PERSON, I. (1992): Das kausative Funktionsverbgefüge (FVG) und dessen Darstellung in der Grammatik und im Wörterbuch. Deutsche Sprache 20. pp. 153-171.
- PERSON, I. (1975): Das System der kausativen Funktionsverbgefüge. Eine semantisch-syntaktische Analyse einiger verwandter Konstruktionen. Inaugural Dissertation. Malmö: Liber.
- POUGUÈRE, A. (2000): Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In: Proceedings of EURALEX 2000. pp. 517-527.
- ROTHKEGEL, A. (1973): Feste Syntagmen. Grundlagen, Strukturbeschreibung und automatische Analyse. Tübingen: Niemeyer. Linguistische Arbeiten.
- SCHROTEN, J. (2002): Light Verb Constructions in bilingual dictionaries. In: From Lexicology to Lexicography. Eds. Francine Melka & Celeste Augusto. Utrecht: University Utrecht, Utrecht Institute of Linguistics OTS. pp. 83-94.
- SGALL, P. – HAJČOVÁ, E. – PANEVOVÁ, J. (1986): The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht: Reidel; Prague: Academia.
- VLKOVÁ, V. (1990): Příspěvek k analýze multiverbálních spojení typu provádět rekonstrukci. In: Slovo a slovesnost, 51, volume 1. pp. 1-15.
- WANNER, L. (1996): (Ed.) Lexical Functions in Lexicography and Natural Language Processing. Studies in Language Companion Series (SLCS), Vol. 31. Amsterdam-Philadelphia: John Benjamins.

## RESUME

Složenými predikáty (SP) rozumíme zejména konstrukce složené z významově vyprázdněného slovesa a nějakého abstraktního substantiva (často označujícího děj nebo stav). Toto substantivum může být jak deverbativní (např. *učinit rozhodnutí*, *provést údržbu*), tak nedevertativní (např. *dát pohlavek*). Zmíněný typ predikátů je občas označován také za predikáty verbonominální, příp. analytické. V rámci anotací PDT volíme pro tyto predikáty název „složené predikáty“, a to z toho důvodu, že termín *analytický* je vyhrazen pro tzv. analytickou rovinu PDT a atributy s ní spojené, termín *verbonominální* pak zpravidla používáme pro označení jednoho z podtypů složených predikátů, a to predikátů tvořených sponovým slovesem *být*. V anglickém textu používáme pro označení všech typů složených predikátů termín *complex predicates*; pro složené predikáty, které nejsou tvořeny sponovým slovesem *být*, pak volíme obecně užívaný termín *support verb constructions*. V tomto příspěvku se věnujeme složeným predikátům bez sponového slovesa *být*.

V rámci jednotlivých složených predikátů rozlišujeme slovesnou část (SČ) a jmennou část (JČ). Slovesná část SP se může nominalizovat, pak jde o spojení dvou substantiv (např. *věnování pozornosti*), případně jde o konstrukci s deverbativním adjektivem (např. *pozornost věnovaná dětem*).

V PDT jsou zachycovány kromě jiného i koreferenční vztahy mezi některými uzly tektogramatické stromové struktury a anotuje se rovněž aktuální členění věty. Při rozhodování, které konstrukce máme při anotacích označit za složený predikát, jsme vybíraly zejména konstrukce s gramatickou koreferencí (tzv. složené predikáty kontroly) a dále taková spojení slovesa a substantiva, jejichž jmenná část má nějaké vlastní valenční doplnění, jehož slovosledné postavení může v tektogramatické stromové struktuře způsobovat neprojektivní konstrukce. Současný seznam sloves, která mohou vstupovat do složených predikátů, byl tedy těmito hledisky značně limitován a určitě není vyčerpávající (lze ho získat prohledáváním dat nebo valenčního slovníku (tzv. PDT-vallexu), čítá zhruba 150 položek).

Po prostudování české i zahraniční odborné literatury jsme došli k závěru, že pro složené predikáty jsou charakteristické zejména následující skutečnosti:

(i) Sémantické vlastnosti slovesné a jmenné části SP

Slovesná část SP je často významově vyprázdněná, význam celého SP je dán významem jmenné části. Většinu sloves vstupujících do složených predikátů je možné přiřadit ke slovesům, která nazýváme kvazifázová, protože spolu se jmennou částí vyjadřují jednu z fází průběhu děje. Substantiva mající společnou slovesnou část jsou často významově propojena, někdy dokonce tvoří určité sémantické třídy.



Dané spojení slovesa a substantiva je možné ze sémantického hlediska považovat za jednu (víceslovnou) lexikální jednotku, k tomuto spojení je tedy většinou možné nalézt adekvátní synonymní vyjádření pomocí syntetického predikátu. Vzhledem ke konvenci zavedené v PDT-vallexu však víceslovné lexikální jednotky nezachycujeme pomocí jednoho uzlu, nýbrž každá z částí má svůj vlastní uzel. K zachycení složeného predikátu jako jedné lexikální jednotky využíváme následující anotační prostředky:

*Funktory pro jmennou část SP.* Vzhledem k tomu, že jmenná část SP tvoří se slovesem jednu lexikální jednotku, není adekvátní ji považovat za jeden z aktantů příslušného slovesa. Jmenné části SP tedy přiřazujeme speciální funktor CPHR („compound phraseme“, zkráceno z „část složeného predikátu“).

*Tektogramatické lemma QCor.* Skutečnost, že je možné nějaký SP považovat ze sémantického hlediska za jednu lexikální jednotku, má za důsledek referenční totožnost určitých valenčních členů substantiva a slovesa tvořících daný SP. Relevantní referenčně totožné valenční doplnění jmenné části SP je zpravidla povrchově vypuštěno. Domníváme se však, že v hloubkové struktuře věty je toto valenční doplnění přítomno, proto ho na tektogramatické rovině PDT doplňujeme a přiřazujeme mu speciální tektogramatické lemma QCor (tj. Quasi-Control). Od uzlu s lemmatem QCor pak vede koreferenční šipka k tomu valenčnímu doplnění slovesa, s nímž je uzel s lemmatem QCor referenčně totožný. Existuje více různých typů totožnosti (sdílení) valenčních členů.

(ii) Valenční vlastnosti slovesné a jmenné části SP

V PDT počítáme s tím, že jak slovesná část, tak jmenná část SP může mít svoji vlastní valenci: v PDT-vallexu bude mít jak sloveso, tak substantivum své vlastní heslo a svůj vlastní valenční rámec.

*Valence slovesné části složených predikátů.* Jmenná část s funktorem CPHR může být vyjádřena formou bezpředložkových i předložkových pádů. Nejčastější formou JČ v rámci SP je bezpředložkový akuzativ. V PDT se vyskytují zejména následující formy třetího valenčního doplnění SČ v rámci SP s bezpředložkovým akuzativem: dativní doplnění (funktor ADDR; *dát někomu možnost*); *od*+2 (funktor ORIG; *dostat od někoho úkol*); *z*+2 (funktor ORIG; *nabýt z něčeho dojem*); *na*+4 (funktor ADDR/DIR3; *klást na někoho nároky*); *v*+6 nebo *u*+2 (funktor LOC; *budit v někom nepříjemný pocit, vzbuzovat u někoho pochybnosti*).

Hodnotíme-li valenci slovesné části SP z toho pohledu, zda je určité valenční doplnění typické i pro bezpříznakové užití slovesa, můžeme vymezit následující typy: (a) valence, kterou má dané sloveso i mimo užití ve složeném predikátu; (b) valence, kterou dané sloveso získává až při zapojení do složeného predikátu (většinou jde asi o analogii k valenci jednoslovného synonymního slovesa); (c) někdy může dokonce sloveso svoji valenci typickou pro bezpříznakové užití ztrácet. Slovesa, která vstupují do SP, mohou mít tedy i v rámci složených predikátů různé valenční rámce.

*Valence jmenné části složených predikátů.* Valenci jmenné části SP zkoumáme jak v případě, kdy je JČ součástí SP, tak v případě, kdy se daná JČ od SČ osamostatní a vystupuje v textu sama o sobě.

Z dokladů složených predikátů v PDT se zdá, že *deverbativní substantiva* mají v rámci SP svou vlastní valenci v naprosté většině případů. Může jít o valenční doplnění vyjádřené prostými pády (např. *provést opravu něčeho*), předložkovou vazbou (např. *vést debatu o něčem*), i formou infinitivu nebo vedlejší věty (např. *vydat pokyn + inf*). Také některá z *nedeverbativních substantiv* mají v rámci SP svou vlastní valenci, často získanou od slov, od kterých byla odvozena (zejména od deverbativních adjektiv, např. *zodpovědnost za něco*).

Pro valenční chování deverbativních substantiv v rámci SP je typické, že ta valenční doplnění, která jsou referenčně totožná s nějakým valenčním doplněním SČ, jsou v povrchové realizaci věty vypuštěna. V naprosté většině případů takové valenční doplnění nelze vůbec doplnit (např. *\*Poskytl Janovi péči o něj/o Jana*), v případě totožnosti konatelů je výjimečně možné Aktora jmenné části vyjádřit pomocí přivlastňovacího zájmena *svůj* (např. *Petr Karlovi znovu položil svoji.ACT otázku*). Některá z nedeverbativních substantiv v rámci SP získávají valenční doplnění, o kterých se u nich běžně neuvažuje (zejména Aktor vyjádřený přivlastňovacím zájmenem, u substantiv tvořících součást SP, která jsou synonymním vyjádřením modálních sloves, pak i infinitivní vazba a její varianty, např. *Petrova.ACT šance najít.PAT zaměstnání*).

V případě, že se substantivum osamostatní od slovesné části svého SP a vystupuje v textu samo, může přebrat formu třetího valenčního doplnění slovesné části svého SP. Jedná se zejména o dativní vazbu a o valenční doplnění vyjádřené formou *od+2*. K přebírání formy valenčního doplnění může dojít jak u deverbativních, tak u nedeverbativních substantiv.

Vliv dativní formy třetího valenčního doplnění slovesné části SP na valenční chování jmenné části daného SP je nejzřetelnější v konstrukcích se substantivy odvozenými od sloves s příslušným valenčním doplněním vyjádřeným akuzativem (např. *podpořit někoho*, ale *vyjádřit někomu podporu* → *podpora někomu*) nebo genitivem (např. *otázet se někoho*, ale *dát / položit někomu otázku* → *otázka někomu*). U jiných substantiv však oporu pro dativní vazbu u slovesné části SP nemáme (např. *vyzvat někoho*, *učinit výzvu*, ale *výzva někomu*). Dochází tak ke specifickým formálním změnám Ak → Dat a Gen → Dat. Některá ze substantiv s dativní vazbou odpovídající u slovesa akuzativu si uchovávají i možnost vyjádření příslušného valenčního doplnění pomocí genitivu (např. *podpora někoho.ADDR*, *varování někoho.ADDR*).

K převzetí formy třetího valenčního doplnění slovesné části SP dochází často i v případě vazby *od+2* (např. *získat od někoho.ORIG* *slib* → *slib od někoho.ACT*).

Složené predikáty bezesporu představují komplexní jazykový jev, jehož zkoumání zasahuje do mnoha různých oblastí. Dotkly jsme se pouze dvou z nich; krátce jsme se zabývaly sémantickými vlastnostmi složených predikátů, největší pozornost jsme pak věnovaly jejich valenčním vlastnostem. Popsaly jsme také základní pravidla anotace složených predikátů v PDT a jejich zachycení v PDT-vallexu.

#### ACKNOWLEDGEMENT

The research reported in this paper was supported by the grant of the Czech Ministry of Education LN00A063 and the grant of the Grant Agency of Charles University GA-UK 489/2004.





## On the Delimitation of Analytic Verbal Forms

MARKUS GIGER

0. The question of the border between free syntagms (FS) on the one hand and analytic verbal forms (AVF) on the other is of major interest in linguistic investigation: Practically all models of language count on a split between lexicon and grammar. However, the question of where to draw the frontier between them is not at all easy to answer. Especially in the case of potential verbal forms, there is often little agreement between linguists as to whether a combination of a semantically rather “empty” inflected verb with an infinite form of a verb with full lexical meaning has to be interpreted as AVF (the inflected verb therefore as an auxiliary) or as FS (the inflected verb would then be autosemantic).

In the course of my work on resultative constructions in Czech, (Giger 2003a) I paid attention to the interpretation of these constructions in previous literature. Resultatives are built in Czech mainly by a combination of the verbs *být* ‘be’ and *mít* ‘have’ with an (original past) participle in *-n-* or *-t-*: *Okno je rozbité* ‘The window is broken’, *Otec má polévku uvařenou* ‘The soup for father is cooked, Father has a cooked soup’ (lit. ‘Father has cooked the soup’). When speaking about the present perfect in English, the so-called perfect in German, the ‘passé composé’ in French or the ‘passato prossimo’ in Italian (all of which have a practically identical formal structure) there is little doubt that they have to be counted as part of the verbal paradigm, as AVF, but there is no such evidence for resultatives in Czech (or any other West-Slavic language). Nevertheless, previous literature most often implies an answer to the question, even by not explicitly asking it: When Vilém Mathesius, author of the first work on Czech possessive resultatives (Mathesius 1925), calls his study “*Slovesné časy typu perfektního v hovorové češtině*” (‘Verbal tenses of perfect type in colloquial Czech’), this seems necessarily to imply an interpretation of the constructions in question as grammeme(s) of the category of tense and therefore as AVF. Similarly, Karel Hausenblas gives his work the title *Slovesná kategorie výsledného stavu v dnešní češtině* (‘The verbal category of resultative state in modern Czech’): although he does not imply an interpretation as verbal tense any more, his using the term ‘category’ leads us to think of a grammatical (inflectional) category, based on the opposition between action and state. Members of a grammatical category are grammemes, and grammemes are expressed by inflected forms so, in the specific case, these would be analytic. In the same way, the presentation of *mít* + participle as a special case of aspect in the works of Jarmila Panevová and other members of the Prague Institute for formal and applied linguistics (Panevová et al. 1971: 35, Panevová/Sgall 1971, 1972) leads to the analogous implication, although this time through the grammatical category of aspect. On the other hand, the collective *Mluvnice češtiny* (‘Grammar of Czech’; MČ 2: 174) declares on behalf of the same constructions: “Tyto další konstrukce s participiem trpným v žádném případě nemohou být považovány za tvary sloves” (‘These further constructions with passive participle [i.e. those apart from *být* + participle] cannot be considered verbal forms in any way’).

However, even here we are not told on what theoretical basis this statement was made. A special solution to this point is proposed by Krupa (1960) on behalf of the possessive resultatives in Slovak: using a set of tests, he considers a part of the combinations *mať* + *n/-t-* participle to be perfects and a part to be FS.<sup>1</sup>

Another construction in Czech (and also in Slovak and Sorbian) for which the present question is relevant is the closely related *dostat* 'get' + *n/-t-* participle: *Dostal jsem prémie přidělení* 'I was allocated a premium'. Daneš (1968, 1976) calls this construction 'recipient passive' and so connects it with the category of voice. Nevertheless, Daneš clearly recognizes the problem discussed here when he says (1968: 289): "Konstrukce se slovesem *dostat/dostávat* (...) můžeme pokládat za zvláštní druh dějových pasivních konstrukcí, které sice nemají plně charakter tvaroslovné kategorie, ale jsou více než příležitostným opisem." ('The constructions with the verb *dostat/dostávat* we can consider a special kind of actional passive constructions which do not have the full character of an inflectional category but are more than a casual paraphrase').

A third, somewhat more distant example is that of the Slovak construction *ísť* 'go' + infinitive. We do not find *ísť* among the auxiliary verbs of Slovak, either in Ondrus (1964: 103) or in Pauliny (1965: 97). However, the analysis of Orlovský (1964) shows that he clearly recognized the semantic shift of the construction, because he has, on the one hand, *Šiel pozrieť, kto tam* 'He went to see who is there', *Šiel si sa po večeri prejsť* 'You went on a walk after supper' among sentences with the meaning of purpose (o. c.: 231) and, on the other hand, *Líca sa jej šli plameňom chytiť* 'She was going to get her cheeks on fire', *Pamäť ma ide nechať* 'I am going to lose my memory', *Ide byť predstavenie* 'There will be a performance immediately' (o. c.: 233) among sentences with the meaning of factual content. Finally, we do find *ísť* among the auxiliaries<sup>2</sup> of Slovak in MSJ (1966: 365f.): Here *ísť* in sentences like *Srdce jej ide puknúť od žiaľu* 'Her heart is going to break with grief' or *Teraz ide hovoriť veliteľ* 'The commander is going to speak now' is called a 'limitné pomocné sloveso' ('auxiliary of limit') which expresses 'the final state of preparation before the realization of the action' or 'realization of the action in the near future', respectively. It is said that *ísť* in this function has 'a certain relation with the category of tense'. Among the periphrastic verbal forms of Slovak, or within the category of tense, *ísť* + infinitive is not claimed (cf. MSJ 1966: 430f., 462-464, 478f.). So we can conclude that, for the

<sup>1</sup> In Giger (1997) I tried to show that the factor responsible for the differences between the two types is the opposition reversibility/irreversibility of the resultative state. It remains questionable, whether this should be the deciding factor in delimiting an AVF from a FS. – For the discussions on resultatives in Slovak cf. also Ondrus (1964: 123-126).

<sup>2</sup> However, it is necessary to say, that the term 'auxiliary' in MSJ does not mean, that the syntagm built by the auxiliary has to be interpreted as an AVF. 'Auxiliaries' are called all modal verbs, verbs specifying phases, and copulae (cf. o. c.: 362n.). The verb *byť* 'be' used in analytic future and actional passive is called 'formálne, opisné sloveso' ('formal, periphrastic verb') (MSJ 1966: 463). – It is worth noticing, by the way, that the authors of MSJ draw a border between auxiliary (in their sense) and periphrastic verb among the constructions *byť* + participle: while the actional passive is alleged among the AVF (cf. MSJ 1966: 473-481), the copula verb *byť* forming resultatives is to be found among the auxiliaries (o. c.: 371). There seems to be a semantic criterion thus, but it is not explicated. The authors of MČ, on the other hand, point out, that habituality in Czech passive can be expressed only by the verb *býť* 'be' and not by the participle of the full verb (*býval chválen* 'he used to be praised' like *býval zlý* 'he used to be bad', not *\*býl chváliván*), so actional passive behaves like the combination of copula verb + adjective and should, therefore, not necessarily be considered an AVF either (MČ 2: 172). This is a formal criterion, which is, however, isolated; it remains unclear, why even this criterion has to be used and not other.

authors of MSJ, *ist* + infinitive is not a completely FS any more but neither is it an AVF. There are, however, no explicit criteria for the different handling of *Teraz ide hovoriť veliteľ* and *Teraz bude hovoriť veliteľ*.

1. An answer to the question of the delimitation of AVF can be looked for in two sets of frameworks: either in a synchronic formal model of language, which tries to define grammatical (inflectional) categories and where AVF can be called those syntagms that express grammatical categories, or in a diachronically oriented framework, which describes the rise of grammatical forms out of the lexicon. In the central part of this study, I will try to show how the two types of frameworks operate and where they meet.

First I will take, as an example for a synchronic formal model of language, the ‘Meaning  $\Leftrightarrow$  Text Theory’ founded by Igor A. Meščuk, more precisely a part of Meščuk’s ‘Cours de Morphologie générale.’ This system tries to modulate human language, originally with the goal of an application in machine translation.<sup>3</sup> In the first part of his ‘Cours,’ Meščuk distinguishes grammatical and lexical meanings, which differ mutually by polar properties:

Significations lexicales	Significations grammaticales
1. Sont universelles et toujours majoritaires.	1’. Ne sont pas universelles et sont toujours minoritaires.
2. Forment un ensemble ouvert.	2’. Forment un ensemble fermé.
3. Tendent à être directement liées à la réalité extralinguistique.	3’. Tendent à n’être liées à la réalité extralinguistique qu’indirectement.
4. Ne sont pas très bien (ou même pas du tout) structurées.	4’. Sont très bien structurées.

Meščuk (1993a: 257)

Grammatical meanings are more ‘linguistic’ and they characterize the individual system of a natural language. Among the grammatical meanings, Meščuk distinguishes inflectional meanings and derivational meanings. Inflectional meanings are prototypical grammatical meanings, and they are obligatory in the sense that their use is obligatory when using a certain word class. A category Meščuk defines as follows: “Nous appelons catégorie un ensemble maximum de significations qui s’excluent mutuellement dans la même position (sémantique ou logique » (Meščuk 1993a: 261). As examples he uses colours (*red, blue, yellow* etc.), vehicles (*car, truck, bus* etc.), tense in French (‘présent’, ‘imparfait’, ‘passé simple’) or gender in French (masculine, feminine). Crucial to the present question are of course the last, the inflectional categories (‘catégories flexionelles’), which are defined as follows:

<sup>3</sup> On the ‘Meaning  $\Leftrightarrow$  Text Theory’ cf. Weiss (1999) and the introductory literature quoted on the homepage of the ‘Observatoire de linguistique Sens-Texte’ at the Université de Montréal (<http://www.olst.umontreal.ca/textdownloadeng.html>).

### Définition I.30: catégorie flexionnelle

Soit une catégorie **C** comprenant les significations 's<sub>i</sub>' :

$$C = \{s_1, s_2, \dots, s_n \mid n \geq 2\}$$

La catégorie **C** est appelée *catégorie flexionnelle d'une classe K* = {K<sub>j</sub>} de signes en  $\mathcal{L}$  si et seulement si les deux conditions suivantes sont simultanément vérifiées:

1. (a) Auprès de tout signe K<sub>j</sub>, exactement une (et une seule) 's<sub>i</sub>' est obligatoirement exprimée  
et  
(b) toute signification 's<sub>i</sub>' est exprimée obligatoirement auprès d' au moins un signe K<sub>j</sub>.
2. Les 's<sub>i</sub>' sont exprimées régulièrement, c' est-à-dire que:
  - (a) une 's<sub>i</sub>' est strictement compositionnelle (le résultat de l'union  $\oplus$  d'une 's<sub>i</sub>' à une 'K' peut toujours être calculé par une règle relativement générale);
  - (b) si la classe **K** est numériquement large, alors pour toute 's<sub>i</sub>', le nombre de signes qui l' expriment est relativement petit et ces signes sont distribués selon des règles relativement générales;
  - (c) la plupart des 's<sub>i</sub>' sont exprimées auprès de (presque) tous les signes de la classe **K**.

(Melčuk 1993a: 263)

The main definition (1.) refers to the category **C** of a language  $\mathcal{L}$ , which occurs with class **K** (a word class) and has at least two members. Of these members, with every element of class **K** has to occur exactly one, and every element of **C** ('s<sub>1</sub>', 's<sub>2</sub>', 's<sub>3</sub>' etc.) must occur with at least one element of **K**. The three secondary definitions (2.) refer to three dimensions of the linguistic sign (regarding the members of **C**): on the one hand the meaning – it has to be such that it combines according to a relatively general rule with the meaning of the particular members of **K** (2a); on the other hand, the external form – it has to be such that there are relatively few different forms of every member of **C**, and they are distributed according to relatively general rules (2b). And finally (2c) refers to syntagmatics of the elements of **C**: The majority of them are expressed with every or almost every member of **K**. In other words, the inflectional category has to be obligatory and regular in expression.

This definition leads to the definition of grammeme: A grammeme is an inflectional meaning that belongs to an inflectional category (Melčuk 1993a: 264). Melčuk, however, applies some limitations or tries to specify some postulates more precisely: **C** must occur with every member of **K**, but not with every one of its appearances (e.g., infinitives often do not express the category of tense). Not every element of **K** has necessarily to combine with every element of **C**: there are defective paradigms in languages, pluralia tantum, singularia tantum, perfectiva tantum, imperfectiva tantum and so on, and there are 'partial grammemes' such as the partitivein -u in Russian, which occurs only with some masculine nouns in the singular (Melčuk 1993a: 269).

When we try to apply this definition to our examples from West Slavic, the situations of course are different: in the case of the resultative we would have to establish a new inflectional category if we wanted to accept Czech or Slovak resultatives as a grammeme of an inflectional category, because semantically they cannot be a grammeme of either the category of tense, aspect or voice.<sup>4</sup>

<sup>4</sup> It is not possible to go into details here; for detailed argumentation cf. Giger (1997, 2000, 2003a: 101-108, 299-348).

This would mean the establishment of a new binary grammatical category (which could be called ‘dynamicity’ and whose grammemes would be ‘eventive’ – all actional forms of the verb – and resultative). Such an inflectional category *C* would refer to a class of words *K* of a language *L* (the verb in Czech or Slovak). It would be obligatory in the sense that every verbal form of Czech or Slovak could only be either eventive or resultative, and every verbal form would have to be either eventive or resultative (part 1a of Meščuk’s definition). Eventive and resultative would be expressed obligatorily with at least one Czech or Slovak verb (1b). The opposition would be compositional, which means that the semantics of the forms could easily be decomposed as semantics of the verb + actionality or semantics of the verb + resultativity (2a). The number of signs expressing resultativity would be relatively small (the auxiliary verbs plus the participles 2b; however, there would be no special linguistic sign for the eventive, whose coding would be only negative<sup>5</sup>). Most problematical would be part (2c) of Meščuk’s definition: in a binary opposition, it is not possible to speak of the majority of the grammemes, and the distribution of the two grammemes would be rather unbalanced: while all verbs build eventive forms, the building of resultative forms is restricted in many ways (cf. Giger 2003a: 174–226). On the other hand, this pertains also for the category of voice, so it seems clear that Meščuk’s definition of the inflectional category does not prevent us from establishing a category ‘dynamicity’ with a grammeme ‘resultative’, while, of course, not imposing it, either. The decisive point of the problem remains elsewhere: it resides in the question of the complementary distribution of the two potential grammemes. It is not difficult to call a binary opposition *a* vs. *b* obligatory in the sense that every form has to be either *a* or *b*. The question, however, is in how many contexts the speaker must compulsorily use *a* and in how many contexts *b*; even more, whether there is one of the members of the binary opposition for which such contexts exist at all. As is shown in Giger (2003a: 377–380), for Czech resultatives these contexts are rather marginal, and while explicit expression of the resultative state requires a resultative construction, non-use of the resultative construction does not necessarily mean non-existence of the resultative state. That seems to be what Bisang formulated at a grammaticalization workshop in February 2001 at Konstanz University, when speaking about South East Asian languages: “Weitgehendes Fehlen obligatorischer grammatischer Kategorien (Indeterminiertheit). Indeterminiertheit bedeutet, dass der Sprecher lediglich ein Konzept zu nennen braucht, wenn er die Information zu grammatischen Kategorien als beim Hörer nicht-aktiviert betrachtet. Aus der Nicht-Erwähnung einer Kategorie lässt sich nicht schliessen, dass der Sprecher X nicht meint”. Similarly, the speaker of Czech or Slovak can express the concept of resultativity explicitly, but most often he does not have to. On the other hand, he cannot, in the same sense, use singular instead of plural or present instead of past.

There is, however, one more important point to emphasize at the end of this paragraph: the notion of inflectional category is graduated, even with the strict structuralist, Meščuk. Each of the secondary definitions (2a–c) is graduated, and so is the notion of inflectional category as such; something can be an inflectional category more or less. The opposition between eventive and resultative in Czech and Slovak we can interpret as a peripheral inflectional category or, in Meščuk’s terms, a quasi-grammeme, which is his term for an expression that is regular, that builds “forms of the same word”, but is not obligatory (Meščuk 1993a: 303). ‘Analytic forms’, according to Meščuk, are those syntagmatic groups that express grammemes or quasi-grammemes (1993a: 354). This would allow us to so nominate the resultatives in Czech and Slovak AVF.

<sup>5</sup> Meščuk (1993a: 352) admonishes, that in the case of a binary opposition synthetic – analytic form this entails the necessity to postulate a zero sign as opposed to the explicit auxiliary.

2. With our other two examples, the situation is, to a certain extent, different: at least the recipient passive would not request a new inflectional category, but just another grammeme in the already recognised inflectional category of voice.<sup>6</sup> That is how Faßke and Michalk resolved the question of the formally identical recipient passive in Upper Sorbian, calling it ‘indirect passive’ and ranging it among the grammemes of the Upper Sorbian category of voice (Faßke/Michalk 1981: 221-224). On the other hand, the recipient passive in Czech (and more so in Slovak) is much rarer than the possessive resultative, as is shown in Giger (2003b, 2004), which is concerned with fulfilling part (2c) of Meščuks’s foregoing definition. Non-obligatoriness pertains also for the recipient passive in the sense that it can be replaced by the direct passive. Otherwise, what was said about the possessive resultative pertains for the recipient passive.

As for the Slovak construction *íšť* + infinitive, its inherent meaning seems to be prospectivity (cf. Dik 1987: 61, 82; Dik quotes Comrie’s definition of ‘prospective’: “A state is related to some subsequent situation, such that the seeds of that subsequent situation are already present in the earlier state”). This is a certain kind of aspectual meaning, which holds true equally for the type *Ide ho rozhodiť od hnevu* ‘He is going to burst with rage’ (where we know that he will not burst at the end) as for the type *Idem sa ženiť* ‘I am going to marry’. Only in the second case, however, does the construction *íšť* + infinitive enter into concurrence with the future tense. So we are confronted with a certain aspectual meaning but, for formal reasons, (the aspectual system of Slovak is built on completely different formal means) we would hardly accept it as a third aspectual grammeme on the same levels as imperfectivity and perfectivity. The fact that it combines with these grammemes excludes such a solution.<sup>7</sup> On the other hand, the type *Idem sa ženiť*, *Ide hovoriť veliteľ* has clear properties of an immediate future tense, so the potential inflectional category, in which it could be integrated, could be tense as well. It remains, however, not obligatory (*Ide hovoriť veliteľ* implies immediate future, but *Bude hovoriť veliteľ* does not exclude immediate future) and restricted for stylistic reasons, and we should bear in mind that the construction combines with tense too.

3. On the basis of the above, the question arises as to how to modulate the continuum between FS and AVF (or, in Meščukian terms, the broad field of quasi-grammemes). Here, the results of intensive investigations into grammaticalization in the last two decades can be useful. Also the theory of grammaticalization proceeds from a split between grammar and lexicon in language. Grammaticalization means the transition from lexicon to grammar, the rise of grammatical means out of lexical entities (Lehmann 1995: 1; on grammaticalization of auxiliaries cf. especially Heine 1993). The transition is continuous and there is no strict border between lexicon and grammar, but an evolution from less grammatical to more grammatical. Lehmann describes this process by three parameters on the paradigmatic and syntagmatic axes, respectively, from which arises a system of six criteria:

<sup>6</sup> For voice as grammatical category in the framework of Meščuk cf. Meščuk (1993b).

<sup>7</sup> The statement of MSJ (1966: 366) that the expression of near future by the construction *íšť* + infinitive is limited to the use of the imperfective aspect with the main verb is not correct: It is easy to find counterexamples as *Haiderova Strana Slobodných hovorí, že ide urobiť reformy sociálneho štátu a privatizáciu* ‘Haider’s Freedom party says, they are going to reform the social state and realise privatization’, *Fakultnú nemocnicu idú presťahovať z Mickiewiczovej do Petržalky* ‘They are going to shift the Faculty Hospital from the Mickiewicz street to Petržalka’ (SNK).



	paradigmatic	syntagmatic
weight	integrity	structural scope
cohesion	paradigmaticity	bondedness
variability	paradigmatic variability	syntagmatic variability

Lehmann (1995: 123)

'Integrity' refers to the "substantial size of a sign, both on the semantic and phonological sides", 'structural scope' to "the extent of the construction which it enters or helps to form", 'paradigmaticity' to "the degree to which it enters a paradigm, is integrated into it and dependent on it", 'bondedness' to the "cohesion of a sign with other signs in a syntagm, the degree to which it depends on, or attaches to such other signs", 'paradigmatic variability' to "the possibility of using other signs in its stead or of omitting it altogether", and 'syntagmatic variability' to "the possibility of shifting it around in its construction" (Lehmann l.c.). These criteria are not independent on each other, but they correlate, at least to a certain degree.

An example for the change of integrity in the sense of the phonological size is the development of the Proto Slavic perfect auxiliary \**jesi* 'you are' into -s in Czech *Řekl<sub>s</sub>* to 'You said it', *Tys to řekl* 'It was you who said it'. There are no such phenomena in the constructions considered here, the verbs *mít/mať*, *dostat/dostať* a *íst'* remain unchanged, and there is no 'split' between the forms used in the sense of the verb with full meaning and the auxiliary (cf. in contrast to this, the Common Czech opposition *Řekl<sub>s</sub>* to vs. *Seš<sub>s</sub>* *dobřej* 'You are great', Moravian *Řekl<sub>sem</sub>* to 'I said it' vs. *Su<sub>s</sub>* *rád* 'I am glad'). There is, however, loss of semantic size in all examples: in Czech, we find possessive resultatives such as *Je mi jako bych tě měla již ztracenou* 'I feel as if I had already lost you', *Měl obě nohy amputované* 'He had both legs amputated', where the meaning of the main verb is in direct opposition to the meaning of *mít* used as a full verb. Among the Czech examples for the recipient passive, we find *Dostali potvrzeno, že to je chřipka* 'They got confirmation that it is an influenza' or *Spartané dostali nakopáno* 'The Spartans got a kick'; in Slovak *Ten, kto sa zaujíma o moje piesne, ich dostáva pripravené naozaj kvalitne* 'Who is interested in my songs, gets them prepared really qualitatively', where the act of giving is already interpreted in a broad sense, but there are so far no examples such as the Upper Sorbian *Su to přeč wzate krynyli* lit. 'They got it taken away'. Among the examples for the Slovak prospective construction, we found *Ide byť predstavenie* or *Ide hovoriť veliteľ*, where there is no movement, and we can construct a sentence like *A vy ty idete zostať?* 'And you are going to stay here?' where the original meaning of *íst'* as a full verb and the meaning of *zostať* are in direct opposition. This kind of loss of semantic substance is usually called 'bleaching' in grammaticalization literature.<sup>8</sup> As far as paradigmaticity is concerned, the opposition of the auxiliaries *být/byť*, *mít/mať*, and *dostat/dostať* forms a certain paradigm of participle constructions in Czech and Slovak but, nevertheless, the borders of the whole paradigm are not absolutely clear (there is e.g. another verb which combines with the *n-/t*-participle, *zůstat/zostať* 'remain' – will this be a further quasi-grammeme?), and

<sup>8</sup> Another important symptom of 'bleaching' is the building of the grammaticalizing construction from the full verb, from which the new auxiliary originated, cf. English *I have had*, *I am going to go*. While a recipient passive from *dostat/dostať* is semantically not possible, a possessive participle construction from *mít/mať* could arise only after a semantic shift to a perfect (cf. Macedonian *imam imano* in Giger 2003: 489). However, examples for the prospective construction from Slovak *íst'* can be found, e. g. *Nezdalo sa jej, že naozaj ide ísť* 'It didn't seem to her that she was really going to go' ([http://papuch.dobre-jedlo.sk/papuch\\_odvoz.htm](http://papuch.dobre-jedlo.sk/papuch_odvoz.htm)).

there is a clear difference in comparison with such categories as number, case or tense in Czech and Slovak. As for the prospective construction, the paradigmatic integration seems even less evidential, as *ísť* is quite isolated in the function of an auxiliary. Paradigmatic variability concerns the degree of obligatoriness, as already discussed above. The degree of change of structural scope can be shown with all three constructions examined here: *mít/mať* as well as *dostat/dostať* are transitive verbs; when used as full verbs, they combine with an object, but not with a participle: *Dostal som knihu, Mám knihu*. They function thus at the level of a sentence. When the verbs combine with participles without binding an accusative object at the same time, their structural scope is reduced to a verbal phrase: *Dostal som vynadané* 'I was scolded', *Majú otvorené* 'They are open'. A formal effect of this state could be the loss of congruence between an eventual object and the participle. This is not the rule in Czech or Slovak possessive resultatives so far (either in use or norm), but examples have been found in either language (cf. Krupa 1960: 54, Giger 2003a: 394-407).<sup>9</sup> Somewhat different but analogous is the situation with the verb *ísť*: as a full verb it combines with a subject and eventually with an infinitive (*Peter už ide* 'Peter is leaving already', *Peter ide spať* 'Peter is going to bed'). In its function as an auxiliary *ísť* does not need a subject any more, but it combines necessarily with an infinitive (*Ide pršať* 'It's going to rain'). Bondedness and syntagmatic variability do not show any changes in the case of our three constructions as far as I can see: the auxiliaries are morphologically no closer bound to the main verb than are the original full verbs, when used in a syntactically analogous environment, and their freedom "to be shifted around in the construction", as Lehmann says, is not limited either.<sup>10</sup>

4. However, there remains one problem which has not been addressed so far. If we want to utilize the criteria mentioned in the foregoing paragraph in an operational way, then we need to quantify them. There are aspects which are not too difficult to quantify e.g. the phonological integrity. Lehmann postulates that it is possible to quantify semantic size as well (1995: 161). Paradigmatic variability is more difficult to quantify, as it is dependent on context. Structural scope can be related to levels of linguistic description such as sentence – phrase – word form – stem, so there is a certain hierarchy. Bondedness can be connected with concepts such as free morpheme, agglutinative affix, flexional affix, infix (Lehmann 1995: 162). Syntagmatic variability Lehmann would quantify by the number of positions that the element in question may assume in a syntagm. However, if we assume these procedures of quantifying to work, they are not yet made operational in our question of the delimitation of AVF against FS. Here, we would need to know where to draw the borders in the continuum. At the same time, the method itself tells us that borders are arbitrary because the process of grammaticalization is continuous. If we want to draw borders, we should at least try to find some appropriate points in this continuum. I will try to propose some:<sup>11</sup>

<sup>9</sup> Another, quite common and normatively not banned effect is the binding with objects that represent clearly valencies of the main verb, with which *mít/mať*, *dostat/dostať* would not combine as a full verb (cf. Giger 2000, 2003a: 271-273, 283f.).

<sup>10</sup> For a certain exception to this rule with Czech possessive resultatives cf. Giger (2003a: 384-393).

<sup>11</sup> These points have only a general and provisory character, they are well known in parts and have to be specified for each concrete construction, that is being examined. The question of paradigmaticity is not sufficiently taken into consideration because of the difficulties mentioned above with constructions only moderately grammaticalized, which tend to stand outside of clear cut paradigms and tend to form binary oppositions, which are problematic, when AVF are involved (cf. footnote 5).

1. Is the auxiliary etymologically isolated against the full verbs of the language (in other terms, is there no etymologically identical full verb)?
2. Are there formal differences between auxiliary and full verb and/or between the original FS and the actual construction in question?
3. Did the auxiliary change its valency behaviour in comparison with the etymologically identical full verb?
4. Does the auxiliary combine with main verbs whose semantics are opposed to the semantics of the full verb from which the auxiliary is derived?
5. Does the auxiliary combine with the etymologically identical full verb?
6. Are there contexts in which the construction in question is obligatory? Which are these contexts; how many are there; how are they defined?
7. Does the non-use of the construction lead to the exclusion of its meaning?
8. Are there word order restrictions in the construction in question that are not typical of comparable FS?
9. Which are the systemic restrictions for combinations of the auxiliary with full verbs?

Some of these questions are interrelated, of course: a positive answer to 1. excludes questions 2. and 5., and questions 4. and 5. are only specific cases of the general question 9. A positive answer to questions 1.-8. is a point in favour of the interpretation of a certain syntagm as AVF; the more positive answers there are, the clearer is this interpretation. For the second part of 6 and for 9, it pertains that the more contexts with obligatory use and the fewer restrictions there are, the more we will interpret the construction in question as AVF again.

While question 9 implies the problem of frequency in the system of language, there remains a further, tenth, point: the question of frequency in text. Stronger grammaticalization will entail higher frequency in a representative language corpus. This can be significant for the comparison of such grammemes as past or actional passive with the quasi-grammeme resultative (cf. Giger 2003a: 45f., 412f.) or for the comparison of the recipient passive with the resultative (cf. Giger 2003b: 89), all the while realizing that the recipient passive can hardly ever have a similar frequency as the resultative, because the number of verbs able to build it will always be lower. Finally, it is possible – with caution and in the case of sufficiently similar structures (and corpuses!) to compare the frequency of a practically identical quasi-grammeme in two languages (cf. Giger 2004 on the recipient passive in Slovak and Czech).<sup>12</sup>

<sup>12</sup> As for the Slovak prospective construction, a short survey in the SNK in June 2003 (in the 30 millions corpus Nitra at that time), gave 286 responses to the query *id.\*.\*t* (lemmatization was not available then). This means, of course, that the survey contains only such cases, where a conjugated form of *ísť* stands directly before the infinitive without another word between. From these 286 sentences, ten do not contain the structure *ísť* + infinitive at all and do not have to be considered further (*Paradoxne, <ide opäť> o známu lokalitu s koncentráciou rómskeho obyvateľstva*). From among the remaining 276 sentences there are several, for which we can exclude the interpretation as prospective construction due to the context (*Často si <idem zacvičiť>, dosť bicyklujem, behám, čo mi, samozrejme, pomáha aj pri tenise*). In other cases, the context imposes this interpretation, e.g. *Iste, ale si predstavte, že sa <ide prerokúvať> zákon o štátnom podniku* 'Certainly, but imagine, that a law about a public enterprise is going to be debated on'. In many cases, at least without broader context, the decision can be taken only intuitively. These constraints accepted there are about 220 examples for the prospective construction in the survey, while the recipient passive in Slovak could be found – under the same conditions – only 56 times. The frequency of the prospective construction is considerable also when we consider its colloquial stylistic shape against the background of the corpus of mostly written texts.

5. By answering the ten questions above (and subsequent ones that may be relevant to a concrete construction in a certain language) we can structure the transient area between FS and AVS. We can distinguish, for example, between a rather strongly grammaticalized quasi-grammeme – as the resultative – and a more weakly grammaticalized one as the recipient passive, especially in Slovak (taking into consideration the differences in combining the auxiliary with verbs contradicting its original meaning, the less clear changes of valency behaviour, the missing obligatoriness contexts and the much lower frequency in system and text). The Slovak prospective, on the other hand, is again grammaticalized quite strongly, as I have tried to show above. Even an exhaustive description of the above ten points will not give a clear-cut answer to the question of how to delimit AVF in a language against FS. It will, however, allow the formulation of reasons for the solution chosen. At the same time, it will provide a complex description of the verbal construction in question and this description should form a part of the grammar of the respective language.

## REFERENCES

- BISANG, W. 2001. Sprachtypologie und Grammatikalisierung – die Markierung von Tempus und Aspekt in den Sprachen Ost- und Südasiens zwischen Lexikon, Pragmatik und Grammatik (paper presented during workshop Grammatikalisierung vs. Lexikalisierung at Konstans University 1. 2. 2001).
- DANEŠ, F. 1968. Dostal jsem přidáno a podobné pasívní konstrukce. In: *Naše řeč* 51, 269-290.
- DANEŠ, F. 1976. Semantische Struktur des Verbs und das indirekte Passiv im Tschechischen und Deutschen. In: Löttsch, R., Růžicka, R. (Hrsg.): *Satzstruktur und Genus verbi*. Berlin, 113-124. (*Studia grammatica* 13).
- DIK, S. 1987. Copula Auxiliatization: How and Why? In: Harris, M., Ramat, P. (eds.): *Historical Development of Auxiliaries*. Berlin etc., 53-84. (*Trends in Linguistics. Studies and Monographs* 35).
- FASSEKE, H., MICHALK, S. 1981. *Grammatik der obersorbischen Schriftsprache der Gegenwart*. Bautzen.
- GIGER, M. 1997. Ireverzibilita výsledných stavů jako faktor při interpretaci slovenských časových souvětí. In: Nábělková, M. (zost.): *Varia VI*. Bratislava, 106-118.
- GIGER, M. 2000. Syntaktické modelovanie slovenských posesívnych rezultatívnych konštrukcií v rámci dependenčnej gramatiky. In: *Jazykovedný časopis* 51, 17-26.
- GIGER, M. 2003a. Resultativkonstruktionen im modernen Tschechischen (unter Berücksichtigung der Sprachgeschichte und der übrigen slavischen Sprachen). Bern etc. (*Slavica Helvetica* 69).
- GIGER, M. 2003b. Die Grammatikalisierung des Rezipientenpassivs im Tschechischen, Slovakischen und Sorbischen. In: Sériot, P. (éd.): *Contributions suisses au XIIIe congrès mondial des slavistes à Ljubljana, août 2003*. Bern etc. 2003, 79-102. (*Slavica Helvetica* 70).
- GIGER, M. 2004. Recipientné pasívum v slovenčine. In: *Slovenská řeč* 69 (2004), 37-43.
- HAUSENBLAS, K. 1963. Slovesná kategorie výsledného stavu v dnešní češtině. In: *Naše řeč* 46, 13-28.
- HEINE, B. 1993. *Auxiliaries. Cognitive Forces and Grammaticalization*. New York-Oxford.
- KRUPA, V. 1960. Stavové perfektum v slovenčine. In: *Sborník Filozofickej fakulty Univerzity Komenského. Rad III Philologica*. 11-12, 47-56.
- LEHMANN, C. 1995. [1982]. *Thoughts on Grammaticalization*. München-Newcastle. (*LINCOM Studies in Theoretical Linguistics* 01).
- MATHEIUS, V. 1925. Slovesné časy typu perfektího v hovorové češtině. In: *Naše řeč* 9, 200-202.
- MČ: *Mluvnice češtiny*. 1. Fonetika. Fonologie. Morfonologie a morfemika. Tvoření slov. 2. Tvarosloví. 1986. 3. Skladba. 1987. Praha.
- MEEČUK, I. 1993-2000. *Cours de Morphologie générale*. 1: Introduction et première partie: le mot (1993a). 2: Significations morphologiques. 3: Moyens morphologiques, syntactiques morphologiques.

- 4: Signes morphologiques. 5: Modèles morphologiques, principes de la description morphologique. Montréal. [Russian translation: 1997-2001. Kurs obščej morfologii. 1: Vvedenie. Slovo. 2: Morfoložičeskie značenija. 3: Morfoložičeskie sredstva, morfoložičeskie sintaktiki. 4: Morfoložičeskie znaki. Moskva-Vena (Wiener slawistischer Almanach. Sonderband)].
- MEEČUK, I. A. 1993b. The inflectional category of voice: towards a more rigorous definition. In: Comrie, B., Polinsky, M. (eds.): Causatives and transitivity. Amsterdam-Philadelphia, 1-46. (Studies in Language Companion Series 23).
- MSJ: Morfológia slovenského jazyka. Ved. red. J. Ružička. 1966. Bratislava.
- ONDRUS, P. 1964. Morfológia spisovnej slovenčiny. Bratislava.
- ORLOVSKÝ, J. 1965. Slovenská syntax. Bratislava.
- PANEVOVÁ, J., BENEŠOVÁ, E., SGALL, P. 1971. Čas a modalita v češtině. Praha. (AUC. Philologica. Monographia 34).
- PANEVOVÁ, J., SGALL, P. 1971. Relativní čas. In: Slovo a slovesnost 32, 140-148.
- PANEVOVÁ, J., SGALL, P. 1972. Slovesný vid v explicitním popisu jazyka. In: Slovo a slovesnost 33, 294-303.
- PAULINY, E. 1965. Krátka gramatika slovenčiny. Bratislava.
- SNK: Slovenský Národný Korpus. <http://korpus.juls.savba.sk>
- WEISS, D. 1999. Sowjetische Sprachmodelle. In: Jachnow, H. (Hrsg.): Handbuch der sprachwissenschaftlichen Russistik und ihrer Grenzdisziplinen. Wiesbaden, 873-909. (Slavistische Studienbücher. Neue Folge 8).



## Slovak National Corpus – History and Current Situation

MÁRIA ŠIMKOVÁ

Since the second half of the 20th century we have witnessed the rapid development of the following disciplines, many of them being arbitrarily defined: sociolinguistics; psycholinguistics; pragmatic linguistics; text linguistics; cognitive linguistics. Particular positions are occupied by those disciplines that combine linguistic and mathematical methods, which began to develop with the introduction of cybernetics and with the interest in artificial intelligence, machine translations, etc. The increased performance of computer technology ushered in (and continues to bring about) new options for processing a large volume of data when processing natural language automatically. In the 1990s, large text corpora were being emphasised to such an extent that those years are titled the *corpus linguistics* decade. Besides the quantitative increase in the number of corpus workplaces and general national and specialised corpora (probably mainly in Eastern and Central Europe in the above decade), the early 90s were also characterised by a qualitative change in the attitudes of linguistics and other branches, and of interested experts, towards the corpus. A marked shift took place, from the question “*why corpus?*” to pragmatic considerations as to the best utilization of corpora, not only for improving the quality (increasing exactitude) or speeding up linguistic researches and their wider inter-disciplinarity, but also for the utilization of corpora as a reference source for information for various areas of science and research, as a tool for research and development of linguistic technologies and other application of artificial intelligence. The initial difficulties that characterised the introduction of corpora in the 1960s (inadequate output of computers, incompleteness of mathematical formalised descriptions of natural language and rejection by linguists, who were used to traditional theories that were based on small volumes of material that were often highly abstracted) were manifested in various forms in Slovakia thirty years later.

The establishment and operation of the Mathematical Linguistics and Phonetics Department of the Slovak Language Institute of the Slovak Academy of Sciences (today renamed as the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences) was the first promising development project in the area of mathematical and computer linguistics in Slovakia (1962–70). Ján Horecký, who was its initiator and head, programmatically strove to develop the principles and methods of mathematical (algebraic) linguistics on the basis of the material of the Slovak language. Nevertheless, the lexicon of morphemes, which the department was preparing, was never finished.

In the next period, the mainly quantitative analysis of texts developed in the area of mathematical linguistics in Slovakia. J. Mistrik’s frequency lexicons are well known, as well as the partial studies of some researchers who focused their attention on the statistics of linguistic phenomena.

Slovakia only subscribed to the worldwide trend of the development of computer and linguistic technologies as late as 1989, when the topical area “Computer processing of lexis” was included in the programme of the symposium “Methods of research and description of lexis of the Slavonic



languages”, which was held within the framework of the 7<sup>th</sup> Meeting of the Lexicology-Lexicographic Commission of the International Slavists Committee. It consisted of three Slovak (J. Horecký, J. Furdík, P. Žigo) and two foreign prepared contributions; 1 foreign and 1 domestic (J. Horecký) contribution to the discussion (cf. the proceedings of the homonymous symposium, 1990). V. Blanár glossed the topical area as follows in his closing speech: “The idea is being confirmed that the capacity of the human brain is not sufficient to master the continuing growth of information. Humans can meet many information and encyclopaedic challenges only with support from automatic data processing... Moreover, automatic data processing stimulates linguistic research... An important aspect is that such an approach requires looking at many linguistic phenomena from new points of view” (Blanár, 1990, p. 292).

More time passed between the verbalization and the implementation. It was characterised mainly by a lack of technologies and prepared experts in the area, but also by the steps that were directed systematically towards the establishment of the new discipline in Slovakia. After discussions on the options of co-operation of the L. Štúr Institute of Linguistics of the Slovak Academy of Sciences and the Information Centre of the Slovak Academy of Sciences, a new computer linguistics working group was established in 1990. The working group was headed by J. Horecký. They began to work on an integral concept of the future corpus of the Slovak language and lexical database (Jarošová, 1993). Work on a theoretical computer model of the Slovak language (Páleš, 1994) was an important element in this preparatory phase, but, the main work was a practical collection of texts in electronic form, and their first linguistic analyses (Benko, 1993; Šimková, 1993). The collection of the texts was extremely laborious due to the lack of technical and personal background; it was verbatim, word by word, without any tendency to representativeness or at least balance. An opportunistic approach was adopted, i.e., those texts were included in the corpus which were easily obtained and processed. No annotations were made (except for the basic bibliographic information) and the software equipment was also minimal (WordCruncher, later WordSmith; MicroConcord used to be used for preparation of concordances in the MS DOS mode).

The corpus of texts of the Slovak language was gradually made available up to 2002 for internal use within the framework of the L. Štúr Linguistic Institute of the Slovak Academy of Sciences. In its final phase, the 30-million corpus included mainly journalistic texts, some texts of professional proceedings and journals, and a small quantity of belles lettres. A specific part of the corpus consisted of electronic versions of the following lexicographic productions of the L. Štúr Institute of Linguistics of the Slovak Academy of Sciences. Short Lexicon of the Slovak Language (issues 2 and 3); Rules of Slovak Orthography (1998); Synonymic Lexicon of the Slovak Language (issue 1); Academy Lexicon of Foreign Words; Lexicon of the Slovak Language (5 volumes).

One fact should be underlined, i.e., that even the minimal body of information available was in very active use from the beginning, for linguistic purposes (mainly lexicographic ones), and for the purposes of maintaining contacts with foreign corpus workplaces and projects. Most studies presented data processing technologies, selected statistical indicators, or foreign context and theory and practice of lexicographic utilisation of corpora, but there were also more lexical-grammar and comparison studies. The documentary material requested used to be individually prepared and provided to the authors of the above studies. Nevertheless, the existing corpus of texts of the Slovak language and the lexical database were the most widely used in the lexicographic team, which was preparing the concept of a big new monolingual dictionary of the Slovak language (its first volume is just about complete), as well as when preparing the 3<sup>rd</sup> and 4<sup>th</sup> issues of the Short Lexicon of the Slovak Language and the Rules of Slovak Orthography (issues 1998 and 2000). The knowledge and experience gained were honed in international

events abroad and at home. The international seminar “Text Corpora and Multilingual Lexicography” was organised by Ľ. Štúr’s Institute of Linguistics of the Slovak Academy of Sciences and Pedagogic Faculty of Comenius University in Bratislava in 1999. The event was organised within the framework of the international project TELRI II which took place within the framework of the European Commission programme INCO-COPERNICUS. The international seminar “Czech and Slovak Languages in Computer Processing” was organised by the same organisers in Bratislava in 2001 (the event with homonymous proceedings, 2001, resulted from participation in the above project).

This ad-hoc method for building and operating the corpus of texts of the Slovak language gradually showed itself to be impracticable in the long-term horizon. The most important aspect was that it was not comparable with the situation in the neighbouring countries. Moreover, demand for publicly accessible linguistic information began to increase in the late 1990s from the current users. The demands of the lexicographers increased in the context of the volume and the structure of corpus texts, and the efficiency of their utilisation in conceptual work. More demands emerged within the context of Slovakia’s accession to the European Union. After consideration was given to the optimal place and method for the systematic building of a new corpus with internationally comparable parameters, the current project was developed. The project assumed the establishment of a new specialized workplace with adequate technical and personnel background. Preparatory works were launched after the project was approved by the Government of the Slovak Republic on 13.2.2002. The works consisted of building and equipping workrooms in the loft of the building of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences, and the purchase and installation of hardware and software. A working team of the Slovak National Corpus Department of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences was established at the end of 2002. The team comprises seven members.

Despite the fact that the corpus of texts of the Slovak language and lexical database had been built up in the Institute from ca 1993 to 2002, the Slovak National Corpus had no texts available, while contracts with providers of the existing texts either were not completed, or did not contain any clause that would enable incorporation of the texts into the corpus that would be accessible via Internet. Similarly, the technology for processing them did not comply with current standards. The corpus of Slovak language texts had been indexed (without any lemmatization and without any annotations, except for basic bibliographical data) and was operated under MS DOS by WordCruncher, which manifested marked capacity limits even at the level of 200,000 individual occurrences of words and at the overall capacity of 20 million words. The actual work on the building of the Slovak National Corpus (essentially from the beginning of 2003) was launched by the preparation of a licence agreement on other uses of the author’s works according to the Authors Act, by the preparation of a concept of the structure of data in the corpus, and methods for their primary processing, i.e., conversion, tokenization, bibliographic and style-genre annotation (cf. Garabík, 2004; <http://korpus.juls.savba.sk>). In keeping with the tradition of the preceding corpus of texts of the Slovak language and in the context of other current projects, the Slovak National Corpus continues in part in its primary orientation to its user – the lexicographer. In addition, its scope was extended to the wider public (laymen interested in language, students, teachers, editors, and other persons who work with words and/or texts) and experts in the area of grammar research and in the area of NLP.

Our preparation of the project of the Slovak National Corpus was based on the following background: experience in the preparation of existing corpus projects, mainly in Czech; the

requirements of potential users of the electronic database of Slovak texts; the real potential of the working group that is of a minimal size (a staff of seven persons), where persons from many different areas met, but which lacks graduates in computer or corpus linguistics, as no university has such disciplines on their curricula. The following basic objectives were listed in the concept of the Slovak National Corpus for 2003 – 2006 (Šimková, 2003, 2004):

1. Building a general monolingual corpus of written texts of the contemporary Slovak language (1955 – 2005) and making its representative part (200 mill. words) accessible via Internet; lemmatizing and morphologically annotating the accessible part; syntactically annotating a selected specimen.
2. Making the whole file of collected texts, which were electronically processed but bear no linguistic information, available to the staff of the L. Štúr Institute of Linguistics of the Slovak Academy of Sciences, as well as to their external partners on the premises of the Institute, for the purposes of science and research, mainly for lexicographical purposes (the scope is dependent on our technical background and on the willingness of our text providers).
3. Building specific corpora / databases
  - terminology database (in collaboration with the Ministry of Justice of the Slovak Republic and branch terminology committees);
  - database of lexicographical works (making available the lexicographical production of the L. Štúr Institute of Linguistics of the Slovak Academy of Sciences in electronic form via Internet, possibly on CD);
  - corpus of diachronic texts and corpus of dialect texts (on the basis of the needs of the researchers in the respective branches and according to technical background; mainly OCR of ancient prints or manuscripts and transcriptions of spoken language will be demanded);
  - parallel corpus/corpora (mainly for the so-called small languages, where such corpora are good tools for translators and interpreters, but also a good tool for making the language visible and accepted worldwide);
  - Corpus of spoken expressions (the technical and time demands for their transcription will require separate financial and personnel resources).
4. Creation of appropriate software tools (archiving texts; evidence database; conversions and filtrations of texts; lemmatizer; morphological annotator), use and adaptation of existing software tools (parser, corpus manager).

Our data collection was governed by the rule “as many texts as possible, as manifold as possible”. Our approximation towards a representative sample of written texts in the current Slovak language was only very rough: one third consists of journalistic texts, another of fiction texts and the final third of specialized and non-fiction texts. Translations were prominent in the two latter groups, as they have a special position in small national and language societies (such as the Slovak one). Moreover, they were very poorly represented in the previous lexicographical manuals of the Slovak language. Approximately one third of translated fiction, specialized, or non-fiction texts were suggested for the Slovak National Corpus. Translations also occur in the category of journalistic texts, but their identification is substantially more problematic, sometimes even impossible. For instance, translations of agency news provide no indication that the text has been translated. Such information cannot be collected automatically.

Due to the acute need of materials for the team of lexicographers who were preparing the new monolingual dictionary of the current Slovak language, and conditioned by the accessibility and readiness of the provider of the texts, we agreed to accept any text in the first phase which could be gained without excessive effort (acquiring texts from approaching the provider through

explaining the objective, the content, and the non-commercial character of the project, to the execution of the respective contract on the use of the work for scientific and research purposes in accordance with the law on copyright requires, on average, one or two months). In the next phase, we focused our attention on authors or publishers of specific texts which were missing in our representation of genres, or were not adequately represented (e.g. children's literature, the majority of specialized texts in the areas of natural and technical sciences).

When we had succeeded in concluding the starting number of contracts for the inclusion of texts into the corpus, we summarised the methodology of segmentation (tokenization) of Slovak text and its external, bibliographic and style-genre annotations. Concurrently, we initiated the preparation of the morphological tagset itself, as well as of the annotation tools (Forróová – Horák, 2004; Forróová – Garabík – Gianitsová – Horák – Šimková, <http://korpus.juls.savba.sk>). The texts gained were continuously processed and made available for use via the Internet. This approach could be demanding on users trying to become informed on the scope and structure of the texts that were effective at that moment. Nevertheless, the most important achievement was that they were able to work with Slovak texts. The first version prim 0.1 (primary, general corpus), made available in August 2003, contained 26 million tokens. The second version prim 0.2, made available in December 2003, contained 166 million tokens. The third version prim 1 with new tokenization and revised style-genre annotation, made available in July 2004, contained 192 million tokens. In the previous tokenization version, the final scope included paragraphs, titles, tags etc. As a result, version prim 0.2 actually contained fewer than 150 million tokens. Therefore, the increase between versions prim 0.2 and prim 1 was ca 50 million tokens. Moreover, version prim 1 was made available with lemmatization and also, internally, with morphological annotation that was implemented using the tagger and disambiguator produced by the Mathematical and Physical Faculty of Charles University in Prague (authors J. Hajič and J. Hric). The current version, prim 2, was made available at the beginning of November 2005. It provides via the Internet to interested parties a corpus of 246 million tokens from almost 250 providers. In the context of licence agreements, the staff of the L. Štúr Institute of Linguistics of the Slovak Academy of Sciences have ca 10 million more tokens available (some providers of texts do not agree to the inclusion of their texts in the corpus on the Internet, but they agree to their availability for internal use within the L. Štúr Institute of Linguistics of the Slovak Academy of Sciences, for instance, in the context of the preparation of the new monolingual dictionary). Besides the preceding automated morphological annotation, the latest version is also automatically tagged on the basis of our own Slovak tag set (internal lexicographical annotation will be detailed in the next text).

The data structure of the Slovak National Corpus in the version prim 1 (that was the first version to provide a reasonable option of paying attention to style and genre classification) represented almost 182 million tokens (95%) from journalistic texts, 7 million from (3.5%) artistic texts, and 3 million (1.5%) from specialized and non-fiction texts. The disproportion in favour of journalistic texts was very marked. When presenting our corpus, we used to state that it was extremely unbalanced. Nevertheless, the share of non-journalistic texts was sufficiently relevant for us to create the first version of a balanced corpus *primvyn 1*. Within the framework of the basic structure with 60% journalistic texts, 20% fiction, and 20% specialized literature, it contained ca 12 million tokens. Balancing the entire range of corpus texts is also essential for the needs of morpho-syntactical research into the Slovak language in the corpus material (grant project Vega in collaboration with the Philosophical Faculty of Prešov University in Prešov). The project also investigates the distribution of language phenomena in specific types of texts. Representative selection of texts from the linguistics

point of view can be influenced as a result, as well as for the purposes of other grammar researches on the corpus material. The frequency of tokens found in *primvyv 1* manifested the standard distribution not only of the most frequent prepositions, conjunctions, pronouns and particles, but also of the most frequent lexical words, as known from preceding researches and from analogical frequency findings, e.g. in the related Czech language, which were carried out on the representative corpus SYN2000 (Šimková, 2004).

The targeted collection of specific types and kinds of texts was clearly manifest in the new internal structuring of the current version *prim 2.0* as follows: 73% journalistic texts; 13% fiction; 4% specialized literature and non-fiction; 10% texts without the necessary annotation due to various reasons (work on its completion is ongoing). The proportion of translations into the Slovak language makes up 70% in fiction texts (more than 23 million out of 33 million tokens) and 46% in specialized texts (more than 5 million out of 11 million tokens). Our opinion is that this composition reflects relatively realistically the situation in the production and reception of the respective kinds of texts among Slovak readers, and it underlines the old querying of the orientation of the preceding excerpt (prior to 1990) exclusively to top domestic production. The language of the translated texts is Slovak also, but is enriched by lexical and grammatical tools that also name other, unfamiliar facts and enrich the language system in this way. Due to the scope of the specialized texts (all of which were also included in the new balanced corpus, in such a way that their share is 20%, while some of the fiction texts, selected at random, were added so that their share makes up 20% and the remainder of journalistic texts, with the 60% share), the balanced corpus *prim 2.0-vyv* could be offered to the users of the Slovak National Corpus, with a volume of almost 56 million tokens.

Another important result of the new version of the corpus was an increased volume of texts dating from before 1990, resp. 1995, when no text existed in electronic form, or was not archived anywhere. Their share in the version *prim 2.0* is 17.5 million tokens. This could be attained only via intensive scanning and OCRing texts (almost 60,000 pages were processed per man-year in 2005) and their re-construction, which was performed in various volumes by ca 40 collaborators, mainly students. In the context of the goal of the project (i.e., to cover the thesaurus of the current Slovak language since 1955 and prepare material in this way, mainly for the purposes of conceptual works on the new monolingual dictionary of the current Slovak language), the investment is well substantiated. Nevertheless, there continues to be a marked lack of texts of specialized literature. Their representation in the corpus is necessary either in the context of the preparation of the above dictionary, or their use is planned in the context of the creation of a Slovak Terminological Database. The collection of texts (mainly those in the areas of technical and natural sciences) is obviously determined by the following factors: a) new scientific production in specific domains is more frequently written in foreign languages than in Slovak b) older scientific works are often considered obsolete and not relevant even from the point of view of terminology, and their authors are not disposed to make them available for any purposes.

After repeatedly mentioning the availability of the Slovak National Corpus for scientific and research and other non-commercial use via Internet, we should briefly mention the ways and options of working with it. First, searching in the Slovak National Corpus was implemented via a simple web interface (basic search without any support of regular expressions and without displaying external annotation). The corpus manager Manatee with the client Bonito (which was produced by the Faculty of Information Technology of Masaryk University in Brno, author P. Rychlý) could be used once contractual terms and conditions were agreed. The more recent versions of the Slovak National Corpus can be searched using our own



corpus manager Korman, which was developed in the Slovak National Corpus Department of the L. Štúr Institute of Linguistics of the Slovak Academy of Sciences. The corpus manager facilitates the basic search including displaying bibliographic and style-genre annotation, as well as context extensibility. The corpus manager is available virtually for free: the searched string can be entered immediately on clicking agreement to the non-commercial use of the corpus on the introductory page. A specific form must be signed as the basis for using Manatee and Bonito. Then, the user gets his or her own password and has more statistic and frequency data available when searching the corpus as a whole or the studied expressions or forms. Average daily attendance on the corpus' web site is ca 200 entries. Ca 200 new users are registered annually. The individual password needs to be renewed by users at the beginning of each calendar year. This is a way to keep the database of users up-to-date, and discourage idle users. Foreign users are mostly from the neighbouring Czech Republic, but also those from Australia, Canada, Japan, Singapore, etc. can be found.

As previously mentioned, our work on the Slovak National Corpus up to the present also includes the share of the linguistics component. Nevertheless, due to its character, it is being built at a substantially slower pace, the first relevant results being obtained as late as in 2005. The rules of morphological annotation were in development from the beginning of 2003 (Forróová – Horák, 2004; Forróová – Garabík – Gianitsová – Horák – Šimková, <http://korpus.juls.savba.sk>). They formed the subject of a discussion at the end of 2003 and, after minor adjustments, they were accepted as a basis of our own Slovak annotation. More differences emerged in automated morphological annotation when testing the tool that has been developed by the Faculty of Mathematics and Physics of Charles University in Prague: either in the approach of its authors (the so-called *engineering* approach, without any separation of some categories that are relevant for the Czech and Slovak languages, such as verbal aspect and incompleteness and the high error rate of the glossary of Slovak lemmas and forms), or in the language systems of the Czech and Slovak languages and in the theoretical assessment of some categories (e.g., adverbs, particles, secondary prepositions). The first phase of manual morphological annotation was launched at the beginning of 2004, using the co-operation of students of philological departments of other universities in Bratislava, Prešov, and Ružomberok. The tag set was gradually modified again (on the basis of our experience with the first annotations) and the annotation was also adjusted. G. Orwell's novel 1984 was annotated twice before the end of 2004, in the quantity of 102,000 tokens, and its corrections launched. In 2005 the corpus of the texts with manual morphological annotations was extended by the following selected texts with double annotations: the daily Sme and the Internet journal InZine (ca 50,000 tokens); non-fiction Internet encyclopaedia Wikipedia (ca 50,000 tokens). The version *prim 2.0* was automatically tagged on the basis of the first version of the corpus with 130,000 tokens (manually annotated and corrected). Nevertheless, its error rate approaches ca 10%. Therefore, the phase of corrections of the results of the automatic morphological annotation was launched, in such a way that after manual annotation and disambiguation the corpus would have at least 1 million tokens and was adequate for training purposes for our own Slovak annotation tool. Developments also led to the launch of our own morphological analyser and generator of forms. The Slovak Dependency Treebank could be used for the purposes of improvement of the speed and efficiency of the corrections of morphological annotations. Work on the above corpus was launched in summer 2005 within the framework of the Slovak National Corpus, using technical tools and the linguistics and technical manual of the Faculty of Mathematics and Physics of Charles University in Prague. The first phase includes a double syntactic annotation of the texts that underwent manual morphological annotation. The next phase will include the option of



linking manual morphological annotation on the analytical level and of automated morphological annotation.

In its current form, the Slovak National Corpus provides the basic research material for all categories of users and anybody who is interested in the Slovak language. Nevertheless, it is not a substitute for orthographic or grammar manuals. It is only a basis for their creation, a basis that is readily accessible via Internet and essentially provides wider potential within the framework of the automated processing of large numbers of realistic texts. After completing its first big phase in 2006, its results should be made available on CDs/DVDs also. The next phase will include either a continuation of the work of building and balancing the primary national corpus and linguistic annotation of selected texts, or the work of the team and its partners will be oriented more towards the creation of the Slovak Terminological Database and building parallel corpora.

#### BIBLIOGRAPHY

- BENKO, VLADIMÍR: Počítačové korpusy a analýza textu. In: Text a kontext. Red. F. Ruščák. Prešov: Pedagogická fakulta v Prešove Univerzity P. J. Šafárika v Košiciach 1993, s. 43 – 50.
- BLANÁR, VINCENT: Na záver sympózia o metódach výskumu a opisu lexiky slovanských jazykov. In: Metódy výskumu a opisu lexiky slovanských jazykov. Zost. V. Blanár. Bratislava: Jazykovedný ústav L. Štúra SAV 1990, s. 289 – 292.
- FORRÓOVÁ, MARTINA – HORÁK, ALEXANDER: Morfológická anotácia korpusu. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 174 – 183.
- FORRÓOVÁ, MARTINA – GARABÍK, RADOVAN – GIANITSOVÁ, LUCIA – HORÁK, ALEXANDER – ŠIMKOVÁ, MÁRIA: Návrh morfológického tagsetu SNK. <http://korpus.juls.savba.sk/publications>
- GARABÍK, R.: Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 164 – 173.
- GARABÍK, RADOVAN – GIANITSOVÁ, LUCIA – HORÁK, ALEXANDER – ŠIMKOVÁ, MÁRIA: Tokenizácia, lematizácia a morfológická anotácia Slovenského národného korpusu. In: <http://korpus.juls.savba.sk/publications>
- JAROŠOVÁ, ALEXANDRA: Korpus textov slovenského jazyka. In: Slovenská reč, 1993, roč. 58, č. 2, s. 89 – 95.
- PÁLEŠ, EMIL: SAPFO. Parafrázovač slovenčiny. Bratislava: Veda 1994. 305 s.
- Slovenčina a čeština v počítačovom spracovaní. Ed. A. Jarošová. Bratislava: Veda 2001. 196 s.
- ŠIMKOVÁ, MÁRIA: Možnosti využitia programu WordCruncher pri analýze textu (na báze Sládkovičovej a Kraskovej poézie a ľudových rozprávok). In: Text a kontext. Red. F. Ruščák. Prešov: Pedagogická fakulta v Prešove Univerzity P. J. Šafárika v Košiciach 1993, s. 51 – 58.
- ŠIMKOVÁ, MÁRIA: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. In: Počítačová podpora prekladu. Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2003, s. 15 – 19.
- ŠIMKOVÁ, MÁRIA: Slovenský národný korpus – východiská a plány. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 150 – 158.
- <http://korpus.juls.savba.sk>

## ABSTRACT

História budovania Slovenského národného korpusu v Jazykovednom ústave L. Štúra SAV nadväzuje na úsilia profesora Jána Horeckého zo 60. a začiatku 90. rokov 20. storočia. Súčasná práca oddelenia SNK JÚTŠ SAV sa začala rozvíjať na základe uznesenia vlády Slovenskej republiky č. 137 z 13. 2. 2002, ktorým bol schválený pilotný projekt do r. 2006. Základnou úlohou je budovanie Slovenského národného korpusu v celej šírke štýlov, žánrov a vecných oblastí, ale aj regiónov, vydavateľstiev, generácií a pod., v prvej fáze ohraničené na písané texty z obdobia rokov 1955 – 2006. Okrem písaných textov sa plánuje aj korpus hovorenej slovenčiny, tvorba paralelných korpusov a pod.

Slovenský národný korpus ako elektronický súbor jazykových dát s výkonnými nástrojmi na vyhľadávanie a triedenie skúmaných jazykových prostriedkov je od r. 2003 prístupný verejnosti na adrese <http://korpus.juls.savba.sk>. Postupne sa dávajú k dispozícii jednotlivé verzie základného, primárneho korpusu (prim0.1, prim0.2, prim1, prim-2.0), ako aj ďalšie jeho súčasti (napr. podkorpus vyvážený z hľadiska štýlovej distribúcie). Každý text v korpuse je podložený súhlasom autora alebo majiteľa autorských či distribučných práv na jeho spracovanie a zaradenie do celku korpusu podľa licenčnej zmluvy a má podrobnú bibliografickú a štýlovo-žánrovú anotáciu. Celý korpus je od verzie prim1 automaticky lematizovaný (každý slovný tvar má pri sebe informáciu o základnom tvare – leme) a automaticky morfológicky označovaný najskôr pomocou českého softvéru a pomocou českých značiek, postupne aj po natrénovaní značkovacieho softvéru na ručne morfológicky anotovaných textoch na báze vlastného tagsetu. Vybrané texty sa ručne anotujú aj syntakticky. Postupy pri získavaní textov, ako aj princípy ich spracovania od technického čistenia a konvertovania do jednotného formátu cez segmentáciu až po jednotlivé úrovne anotácie sú podrobne opísané na stránke Slovenského národného korpusu i v čiastkových štúdiách.

Slovenský národný korpus využíva každoročne vyše 200 registrovaných používateľov (s vlastným prístupom a možnosťou vyhľadávania pomocou korpusového manažéra Manatee s klientom Bonito), neregistrovaní používatelia v súčasnosti navštevujú stránku priemerne vyše 10-tisíckrát denne.

Slovenský národný korpus by sa mal ďalej rozrastať kvantitatívne i dopĺňať v kvalitatívnych kritériách. Jeho materiál sa bude využívať predovšetkým pri tvorbe výkladového 8-zväzkového Slovníka súčasného slovenského jazyka, ale plánuje sa na ňom aj príprava ďalších slovníkov, ktoré prispejú k exaktnějšímu poznaniu jazykového systému slovenčiny a typologickým i iným výskumom. Pripravuje sa spracovanie a distribúcia vybraných textov aj na CD/DVD nosičoch. Rozširovanie paralelných korpusov sa zameria najmä na česko-slovenský a slovensko-český paralelný korpus, ktorý môže poslúžiť ako materiálová báza na tvorbu prekladového slovníka. Osobitnou súčasťou bude tvorba Slovenskej terminologickej databázy (<https://data.juls.savba.sk/std/>), ktorá by sa mala zamerať najmä na terminológiu z oblasti práva a ekonómie.



# Processing XML Text with Python and ElementTree – a Practical Experience

**RADOVAN GARABÍK**

## 1 INTRODUCTION

XML format, despite its shortcomings, is attracting more and more attention as a format for text representation in corpus linguistics. XML is intended as a free extensible mark-up language for the description of richly structured textual information. The exact method of data description is unspecified and is usually designed according to specific requirements.

The Text Encoding Initiative (TEI) project[1] tries to establish a common XML schema for the general-purpose encoding of textual data. Following the relative success of SGML-based CES (Corpus Encoding Standard), an XML version of it was proposed[2] as a standard to store corpus compatible data.

XML as such gained quite a lot of popularity among different corpora (and corpus linguists); some of them use different XML schemas[3], but many of them use the XCES format.[4]

## 2 INFORMATION HIERARCHY IN TEXT DOCUMENTS

Logically, we can design a rather complicated hierarchy for a document, consisting of sections, each with its heading, each section consisting of subsections (each of those eventually with a heading of its own), then divided into paragraphs. Other types of texts (such as poems) can have different, often more complicated structure. We are talking now only about the structure of information flow in a document, not about other linguistic information (like sentence boundaries). When considering the features (styles) of common word processing and desktop publishing systems, one would expect that this kind of structure is present and in common use.

However, looking at actual texts that come into corpora, we find this kind of structure only very rarely. The overwhelming majority of word processing DTP software users do not use the facility offered by the software to create (or use those already existing) logical styles to format the document, but apply physical text attributes to the document parts instead – so, for example, the headers are distinguished from the rest of the text only by changes in font size or font weight. This makes it almost impossible to use universal tools to extract logical structure from the documents. Often, only very basic structure can be identified and kept in the corpus.

## 3 VARIOUS LEVELS OF TEXT REPRESENTATION

There are actually two different ways of putting texts into the XCES format. One way is to use XML tags to mark up the hierarchical structure of text flow and typographical information. The other way is to use XML to organise basic structural elements of the texts (usually words) together with additional linguistic information into a rigid structure for further processing – in this way, we are using XML format as a (rather inefficient) way of emulating a tabular format.

#### 4 WHY PYTHON

Our programming language of choice is Python[5], a high level object oriented programming language with a very clean syntax. Typically, using Python for software development leads to very short deployment times when compared with others, better promoted languages. The clarity of the syntax also contributes to very few language-oriented bugs in the software, leaving more time for debugging and optimisation of the algorithms used. Python also has an excellent standard library, covering most of the routine programming tasks connected with interfacing various levels of the operating system, user interaction and robust data manipulation. There are also many other external libraries (modules) covering more specialised tasks, and connecting to existing libraries in other programming languages (most notably C and C++) is easy, insofar as programming in C or C++ is easy.

The disadvantages of using Python stem mostly from the fact that it is an interpreted language, with the consequent negative effects relating to speed of execution. While several Python compilers, optimisers and JIT-compilers have been designed, at least theoretically, only Psyco[6] seems mature enough for production use, and its performance gain is not very impressive – thanks to Python's dynamic nature.

#### 5 STRUCTURE OF DATA IN THE SLOVAK NATIONAL CORPUS

Texts coming into the corpus are put into a hierarchical structure, each level corresponding to a different stage of text conversion and processing. Initially, texts are stored in the *Archive* in their original format. The texts are then converted into common text format, keeping some typographic information present in the original sources. We call this level of text processing the *Bank*. The data are then cleaned up and additional linguistic information is added to them, and the files are placed in the next level called the *Corpusoid*. The final step in data processing is a level called simply the *Data*, where the data are converted into binary format for the corpus manager.

File format in the Bank is in fact a simple subset of XCES-conforming XML. The files from the *Archive* are converted into this common Bank-format and these files are then converted on their way to the *Corpusoid*. In the *Corpusoid*, texts are already tokenised, tokens are grouped into sentences, and each token contains additional information about lemma and morphosyntactic categories. Therefore, XML is used here to implement this tabular-like structure.

#### 6 USING ELEMENTTREE

ElementTree[7], by Fredrik Lundh, is a Python implementation of an XML structure representation, in DOM-like style. The whole tree structure is represented by an ElementTree object, which can be created from scratch or read from an existing XML file. Parsing an XML file can be done in one line of code:

```
tree = ElementTree.parse('filename.xml')
```

Similarly, writing in-memory representation of an XML structure to a file can be done in this way:

```
tree.write(file('output_filename.xml', 'w'), encoding='utf-8')
```

Each XML node is represented by a dictionary-like object of an *Element* class. It is possible to loop through children of the node, to find a given subnode, to query attributes of the node

or to modify any of these in place. In order to start working with nodes, we have to create a reference to a top-level node in our XML structure:

```
root = tree.getroot()
```

root is now an *Element* object. Let's take as an example the following piece of an XML file:

```
<p style="plain">Paragraph with a <hi>highlighted</hi> word.</p>
```

This will be represented in Elementtree as an Element class with the following attributes (some are omitted for brevity):

```
element.name == 'p'  
element.text == 'Paragraph with a '  
element.attribs = {'style': 'plain'}  
element.tail = None  
element.children = [hi_element]
```

where *hi\_element* is another Element class:

```
element.name == 'hi'  
element.text == 'highlighted'  
element.attribs = {}  
element.tail = 'word.'  
element.children = []
```

The problem with this approach is obvious: while the text after the highlighted part in our example is logically and structurally on the same level as the rest of the text, in Elementtree XML representation it has been put into the *<hi>* element as a tail attribute, creating a lot of problems when trying to program a way of iterating through the text, because suddenly one has to be aware that parts of the text can be hidden in subordinate elements – and we have to go into arbitrary depths.

In fact, as our experience in parsing the bank format shows, this problem is really intimidating. We had to use complicated solutions, often including careful recursion into subnodes, and we learned that it is almost impossible to modify the document structure in place, because one has to be careful about putting the tail elements into the correct places when eliminating, adding or otherwise modifying the children nodes.

Fortunately, we need not to deal with the texts on this level, the only thing we have to do with texts in the Bank is to tokenise them and transform them into the XCES Corpusoid files.

Looking on the bright side, ElementTree turned out to be a very useful representation of XCES files in the corpusoid. Each token is represented by a *<tok>* node, containing several subnodes describing the token. At the first stage, just after converting the text from bank into XCES format, there is just an *<orth>* subnode with original wordform as a text attribute:

```
<tok>  
<orth>meč</orth>  
</tok>
```

The text is then lemmatised and morphologically annotated. We are using the software described in [8, 9]. The system consists of an external executable program, expecting data in its own SGML encoded format, transforming it and writing the output into an SGML output file. In order to utilise the tagger in our system, we convert our XCES file into input format, run the tagger, then iterate through tokens in the output SGML file and fill in lemmas and morphosyntactic tags into XML elements.

After the analyser run, XML in the Corpusoid looks like this (indentation has been added for clarity):

```
<tok>
  <orth>meč</orth>
  <disamb>
    <base>meč</base>
    <ctag>SSis1</ctag>
  </disamb>
  <lex>
    <base>meč</base>
    <ctag>SSis1</ctag>
  </lex>
  <lex>
    <base>mečať</base>
    <ctag>VMesb+</ctag>
  </lex>
</tok>
```

The results of the analyser run are stored in a sequence of `<lex>` elements. Each `<lex>` element describes one possible combination of a lemma (`base` node) and morphosyntactic tag (`ctag` node), corresponding to a given wordform. Out of these `<lex>` elements, one is chosen by a disambiguating module of the analyser as the right one for the given word, using statistical principles (see [8]), and is put into a `<disamb>` node.

Commented pseudocode (a valid python code) adding a `<ex>` element into the Elementtree corpusoid representation can look like this:

```
# tok variable refers to an element corresponding to
# a <tok> entry in XML file
#
# first, create a subnode of a <tok> node, with XML tag 'lex'
lex = SubElement(parent=tok, tag='lex')
# add newlines to make the XML look more pretty
lex.text = lex.tail = '\n'
# create a subnode of a <lex> node, with XML tag base
base = SubElement(parent=lex, tag='base')
# put the actual content into the <base> XML 'node'
base.text = lemma_from_analyser
# create a subnode of a <lex> node, with XML tag 'ctag'
ctag = SubElement(lex, 'ctag')
# put the actual content into the <ctag> XML node
```



```
ctag.text = tag_from_analyser
# that's all
```

It is possible to run other different analysers (e.g. semantic tagger) at this point; adding additional XML tags (i.e. subelements) into the <tok> node is really easy. Only if we need to modify the superior XML structure, we have to refrain from modifying the document in place because of the difficulties involved, and we should better create a new elementtree structure and create elements and subelements of it as needed.

## 7 COMPATIBILITY AND PERFORMANCE

ElementTree, having been written in pure Python, runs wherever Python can run, without any problems whatsoever. This includes almost all modern Unix operating systems together with Linux and MacOSX, and the Microsoft family of operating systems. Since XML has been designed from the beginning as a common format for textual data cross platform interchange, there are no problems at all in using documents transferred to/from other platforms. To avoid eventual problems with character encoding, we universally use UTF-8 encoding in NFKC canonical normalisation (as is the de-facto norm in the Unix world). The other, perfectly acceptable way would be to use just ASCII encoding, and have non-ASCII characters represented as XML entities. Being written in Python, one could expect ElementTree not to perform sufficiently well. However, in addition to the pure Python version, there is an alternative cElementTree module written in C, with ElementTree-compatible API, much better performance and lower memory requirements. As our experience shows, the speed of parsing is sufficient even for pure Python version on a modest 1200 MHz Pentium III CPU, an average speed of parsing a completely annotated XML file is about 1200 tokens per second. The morphological tagger on the above configuration is able to analyse 250 tokens per second, so the total overhead of using the ElementTree Python-based solution is not bad at all. ElementTree, being DOM-like, not SAX-like, requires the whole parsed document to be present in computer memory; therefore the memory requirements are going to be important. For example, representation of fully annotated document of about 200~000 tokens (one of the biggest continuous texts present in the Slovak National Corpus), being 16~MB of size, takes 410 MB of memory. The C version gives much better results parsing speed is about 80000 tokens per second, and the above mentioned document takes 62 MB of memory, which is perfectly adequate for modern computer systems.

There is also another implementation of the Python XML parsing library with API almost identical to ElementTree, called lxml[10], based on very fast libxml2 parsing library[11]. In addition to ElementTree capabilities, it exposes libxml2 and libxslt specific functionality, providing a way of handling XPath, Relax NG, XML Schema, XSLT and c14n. However, we did not evaluate this software.

## 8 CONCLUSION

Using Python has no doubt great advantages when used in general programming, especially considering its clean syntax, readability and extensive standard library and rich language features, all contributing to very rapid programming. Out of the different XML parser libraries existing for Python, ElementTree stands out because of its pure pythonic approach to the internal XML representation. Using ElementTree is not so straightforward during the first stages of text processing, with complex XML structures usually used to represent typographic information, but it really shines when processing and modifying already tokenised text, with

linear sequence of tokens (or other text units represented as data described by XML tags). The approach described is successfully used in the Slovak National Corpus, where Python is the programming language of choice, used at almost all levels of text processing and conversions.

## REFERENCES

- [1] <http://www.tei-c.org/>
- [2] IDE, N., BONHOME, P., ROMARY, L., XCES: *An XML-based Encoding Standard for Linguistic Corpora*. In: Proceedings of the Second International Language Resources and Evaluation conference. Paris, European Language Resources Association (2000).
- [3] ZAKHAROV, V., VOLKOV, V.: *Morphological Tagging of Russian Texts of the XIX<sup>th</sup> Century*. In: Text, Speech and Dialogue. Proceedings of the 7<sup>th</sup> International Conference TSD 2004. Brno, Czech Republic: (2004) 235–242.
- [4] PRZEPIÓRKOWSKI, A.: *The IPI PAN Corpus preliminary version*. Warszawa, Instytut Podstaw Informatyki PAN.
- [5] <http://www.python.org/>
- [6] RIGO, A.: *Representation-based Just-in-time Specialization and the Psycho prototype for Python*. In: Proceedings of the 2004 ACM SIGPLAN symposium on Partial evaluation and semantics-based program manipulation. Verona, Italy: (2004) 15–16.
- [7] <http://effbot.org/>
- [8] HAJIČ, J., HLADKÁ, B.: Czech Language Processing – POS Tagging. In: *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain: (1998) 931–936.
- [9] HAJIČ, J., HRIC, J., KUBOŇ, V.: Machine Translation of Very Close Languages. In: *Proceedings of the ANLP 2000*. Seattle, U.S.A. (2000) 7–12.
- [10] <http://codespeak.net/lxml/>
- [11] <http://xmlsoft.org/>

## ABSTRACT

In this paper, we evaluate the use of XML format as an internal format for storing texts in linguistic corpora, and describe our experience in using the ElementTree Python XML parser in the Slovak National Corpus.



# Morphological Analysis of the Slovak National Corpus

LUCIA GIANITSOVÁ

## 1 BASIS OF A MORPHOLOGICAL ANALYSIS OF THE SLOVAK NATIONAL CORPUS

The question of a morphological (or morphosyntactic) analysis has been a key problem for natural language processing (NLP) for several years. Automatic morphological annotation is a useful tool, especially with regard to corpus data-processing. In this respect, morphological annotation has also been considered in the course of the development of the Slovak National Corpus (SNC). The theoretical aspects of morphological analysis and its application in corpus-tagging associated with the morphological tag-set preparation for the manual tagging of SNC were outlined by M. Forróová and A. Horák (2004). Annotation, generally understood as the process of adding some information to texts, is – despite differing views – undoubtedly a convenient tool not only in the verification of linguistic theories, but also in carrying out various lexicographical projects.

Morphological analysis is generally understood as the assignment of a base-form (lemmatization), the classification of words into grammatical and semantic classes and the assignment of grammatical categories to words in texts (in the form of tags). In general, for a language-competent person, this kind of analysis is not difficult; by contrast, for computer processing it is a hard nut to crack (Forróová – Horák, 2004). We take into account at the same time the problem of formal homonymy resulting from automatic morphological annotation; a problem which requires a subsequent disambiguation. From the very beginning we were aware of the fact that a set of morphological tags should represent the language properties of a text; in other words, it should somehow interpret a text. We should decide whether the proposed tag-set would result in a new formal description of language, or whether we would reflect the current existing linguistic descriptions, and try to formalise them. Forróová and Horák (2004) point to seven maxims, as proposed by G. Leech, providing regularity of annotation and guaranteeing that annotation does not result in misinterpretations of corpus data. We would like to emphasize the maxim of accessibility and the maxim of a consensus of academic theories: theoretical “neutrality” which was accorded over-riding consideration when preparing a tag-set.

The central issue can be formulated as follows: to what extent can we refer to the traditional grammatical descriptions of Slovak morphology when preparing a lemmatization and a tag-set? We considered it relevant to take into account a systemic description made by academics: *Morfológia slovenského jazyka* (lit. Morphology of the Slovak language, MSJ, 1966), and ultimately some other works dealing with morphology (Oravec – Bajžíková – Furdík, 1984; Dvonč, 1984).

The conflict between the exponents of traditional grammatical categories and the possibility of automatic language processing is also reflected in the approach to a morphological tag-set for SNC. M. Forróová and A. Horák (2004) have already pointed to the vagueness of the criteria for morphological classification. They directed attention to the complex of

morphological, syntactical, lexical and semantic properties of words which serves as a criterion for their classification into parts of speech (MSJ, 1966; Oravec – Bajžíková – Furdík, 1984), and that, within the framework of morphology, linguists traditionally point to the lexical-grammatical categories such as intention, aspect or grade.

At the automatic processing level, this approach encounters numerous difficulties. Obviously, an alternative approach could be applied. In the field of corpus linguistics, it is possible to observe and analyse various approaches to morphological annotation operating in the text annotations of national corpora. M. Forróová and A. Horák (2004) were influenced by considering the advantages and disadvantages of the morphological tag-sets of the Czech National Corpus (CNC), Multext-East Project, and a corpus built in the Institute of Computer Science of the Polish Academy of Sciences (IPI PAN). They conclude that the SNC annotation should not be based solely on a formal approach. This approach led the authors of the CNC tag-set to specify in all 74 possible SubPOS values, including 19 pronoun values (Hajič, 2000; Hana – Hanová, 2002; Forróová – Horák, 2004). Similarly, in the IPI PAN tag-set, there are 29 grammatical classes (for details cf. Forróová – Horák, 2004).

When comparing the above-mentioned approaches with the tag-set types (linguistically optimistic tag-set types and linguistically pessimistic tag-set types)<sup>1</sup> specified in Forróová – Horák (2004) for the purpose of tagging in SNC, a formal-grammatical principle was designated. However, this principle is characterized by some specific features with regard to the domestic Slovak linguistic tradition.

A complete morphological tag-set proposal was introduced on October 24 2003 in Bratislava<sup>2</sup>. Later on it was contested on November 10 2003.<sup>3</sup> In addition, since then, the concept of morphological annotation has been subject to several changes attesting to its validity. After a manual annotation of the first text samples, the need to re-value some parts of the tag-set arose. We took into account the tagging of real texts in the SNC database and the requirements and demands of actual corpus-users (including possible users). The current tag-set version can be found on the SNC website.<sup>4</sup> In the following parts, we will identify only some general features of the morphological annotation of SNC and explain some particular problems and their possible solutions, influenced also by approaches to text tokenization. That is why some brief attention should first be paid to the problem of tokenization.

## 2 TOKENIZATION AS A BASIS FOR THE MORPHOLOGICAL ANNOTATION OF SNC

The approaches to morphological annotation as well as to a tag-set proposal are derived especially from the approaches to tokenization, e.g. the identification of the smallest text units (tokens), which equate neither to words nor to grammatical forms. Tokens are usually defined as chains of characters between two spaces. This concept includes words, numeral

<sup>1</sup> According to Forróová and Horák (2004), linguistically optimistic tag-set types include the implementation of a maximum number of grammatical categories, disambiguation based on syntactic rules; this approach is represented by e.g. V. Petkevič and K. Oliva; on the other hand, linguistically pessimistic tag-set types represent a compromise between a linguistic- and an “engineering”-based approach; eventually, the accommodation of tag-set contents to the mathematical model of a tagger: representatives are Hajič’s tag-set and the tag-set of Multext-East Project.

<sup>2</sup> Contribution to the international conference *Slovko – Slavic languages and their computer processing*; see Forróová – Garabík – Gianitsová – Horák – Šimková.

<sup>3</sup> Internal seminar meetings in SNC, see <http://korpus.juls.savba.sk/activities>.

<sup>4</sup> <http://korpus.juls.savba.sk/publications>

\* Editor’s note: The tag-set was slightly modified on 2005 and is also on the SNC website.

characters, punctuation characters and their combinations. All the reflections concerning a tag-set proposal, the meaning of individual marks and lemmatization, were predetermined by the means of tokenization chosen. The result is that chains of alphanumeric characters (such as letters or numeral characters) between the two spaces are merged into one unit (token). Punctuation (colons, dots, question marks, exclamation marks at the end of sentences, quotation marks, asterisks, mathematical symbols and others) are considered to be individual tokens in spite of the fact that they are not separated from a preceding or following token by a space, e.g. in a sentence „Win98 mi nefunguje!!!“ (lit. „Win98 does not work!!!“) there are 8 tokens (quotation marks „, Win98, mi, nefunguje, three exclamation marks !!!, quotation marks “).

Tokenization is an important phase in automatic text processing because morphological analysis and disambiguation are dependent on it. The proposed principles of tokenization can raise questions concerning analytical forms such as *v rámci, na bielo, a teda*, word-forms with hyphens or dashes (often used incorrectly) such as *kde-kto, čím – tým, 8 – krát, Košice-Bratislava*, analytical forms *menej lukratívny*, collocations *Spišská Nová Ves*, numeral characters such as *1 984*; on the other hand, there are agglutinated forms such as *oňho, akoby* and others. These words are divided into several tokens (despite the fact that they function as one language unit)<sup>5</sup> or integrated into one token (despite the fact that they function as two language units).

This kind of proposal of tokenization leads to an interpretation of words and grammatical forms which does not always accord with our linguistic tradition. However, the actual proposal does not exclude possibilities of the implementation of a logical module into text processing which would be employed afterwards as a more appropriate basis for lemmatization and tagging.

### 3 LEMMATIZATION OF SNC

A lemma (l) is often defined as a “dictionary” form of a token. The set of language features which a lemma should include can be described as follows: the so-called basic values of morphological categories and a distinction between upper and lower case characters. But the concept of a lemma is not only applied absolutely in its given semantic extension (which is analogous to the concept of token and tokenization): A lemma is always indicated by a lower case initial letter, e.g. *Alexander* = *alexander*. It is possible to argue that cancelling the distinction between upper and lower case characters may cause a loss of some semantic features of words; on the other hand, the range of results on the basis of searching for small-lettered lemmas is considerably larger. Moreover, it can be assumed that occurrences of words primarily written in capital letters can be easily found by means of accurately assigned queries in a corpus manager or by means of a negative filter. Information on

<sup>5</sup> For instance, the first parts of composite adjectives usually written with a hyphen (or, incorrectly, with a dash) are individual tokens. They are lemmatized on the basis of the word-form from the text: *česko – slovenský*, lemma (l) = *česko*; *bielo – červený*, l = *bielo*.

Composite pronoun forms such as *ten istý, tá istá, to isté, taký istý, tak isto, kolko ráz, -kolko ráz, kolký raz, -kolký raz, tolko ráz, toľký raz, tamto ten, tamto tá, tamto to*, and word-forms which can be found in the corpus with a hyphen or with (incorrect) dash, such as *čo-to, kolký-toľký, kolko-toľko, aký-taký, ako-tak, kde-tu, kade-tade, kedy-tedy, kdesi-čosi, čosi-kdesi, čosi-kamsi, ten-ktorý, tá-ktorá, to-ktoré*, numerals such as *3-krát* etc., are indicated as two (or three) tokens and each token is lemmatized, even though we respect the fact that it is one lexeme.

proper names is also indicated on the tag level (see Garabík – Gianitsová – Horák – Šimková, 2004, chapter 3.1).

Lemmas assigned to words belonging to inflected parts of speech can be of these values:

Substantives Pronouns <sup>6</sup> Numerals	Adjectives words with adjectival forms	Verbs <sup>7</sup>
particular gender	masculine	infinitive
singular (where it exists) <sup>8</sup>	singular	
nominative (where it exists)	nominative	
	base form	

Specific issues of the lemmatization of some grammatical forms have been solved within the SNC tag-set frame by detailed description (Garabík – Gianitsová – Horák – Šimková, 2004). Here we mention only some cases worthy of remark showing that homonymy (ultimately homography) of some tokens is already handled by manual annotation on the lemma level:

<i>od vedúcej jedálne</i>	l = vedúca	<i>vedúcej pretekárke</i>	l = vedúci
<i>otcovi priatelja</i>	l = otcov	<i>nepovedz otcovi</i>	l = otec
<i>jeho nedobehneš</i>	l = on	<i>jeho priatelja</i>	l = jeho
<i>nechali ho samého</i>	l = sám	<i>šaty zo samého zlata</i>	l = samý
<i>koľkí žiaci prišli</i>	l = koľko	<i>koľkí v poradí boli?</i>	l = koľký
<i>dávať si darčeky</i>	l = si	<i>kto si ty?</i>	l = byť
<i>tuším (asi)</i>	l = tuším	<i>niečo tuším</i>	l = tušiť
<i>začiatkom júna</i>	l = začiatkom	<i>so začiatkom zimy</i>	l = začiatok

#### 4 MORPHOLOGICAL TAGGING OF SNC

##### 4.1 MEANS OF ASSIGNMENT OF ATTRIBUTES AND THEIR VALUES

Various approaches to morphological annotation (CNC, Multext-East Project, IPI PAN, etc.) represent several methods of notation (Forróová – Horák, 2004):

1. **Position** (Hajič): Every position is assigned one character, encoding one grammatical category. Values of irrelevant categories are indicated by dashes, e.g. *politikou*, t = NNFS7---- -A---- (Noun, Noun-common, Feminine, Singular, 7th case, Affirmative).

<sup>6</sup> Blended forms (*oňho*, *preňho*, *naňho*, *oň*, *preň*, *zaň*) represent a special case; these forms are considered to be forms of the pronouns *on*, *ono*. Hence the lemma is composed of an independent preposition and a personal pronoun in the nominative singular and respective gender: *oňho* = *o\_on*, *preňho* = *pre\_on*, *naňho* = *na\_on*. These forms are considered to be agglutinated and this information is indicated on the tag level (see Garabík – Gianitsová – Horák – Šimková, 2004, chapter 3.3.13).

<sup>7</sup> Negated forms are lemmatized by a negative infinitive form, e.g. there are lemmas *vidieť*, *mať*, *chcieť* as well as lemmas *nevidieť*, *nemať*, *nechcieť*. Special attention is required for the lemmatization of the negated verb form *byť*. Negative past and future tense forms are expressed by synthetic means (*nebol*, *nebude*), negation in the present tense is expressed by means of a particle *nie* (*nie je*). The presence of this particle often influences the lemmatization of the verb (*nie je*; l = *nie*, *nebyť*) and at the same time indicates negation (see also Garabík – Gianitsová – Horák – Šimková, 2004, chapter 3.3.11).

<sup>8</sup> A singular lemma can be found even in some pluralia tantum words; otherwise they are lemmatized as nominative plurals: e.g. *nohavice* (lit. N. pl. trousers), l = *nohavica* (lit. N. sg. trouser), but *Alpy*, l = *alpy* (N. pl., not N. sg. *alpa*).



**2 Abbreviated/attributive** (Multext-East): Only the relevant categories for the given word-form are assigned, e.g. *budeme*, t = Vcif1pan (Verb, copula, indicative, future, 1st person, plural, active voice, non-negative).

The advantage of the position notation is that it is more appropriate for computer processing; the abbreviated notation is preferred because of a better understanding by users. Taking all the aspects into account, we have decided to make the best of both concepts.

The values of particular categories in SNC are encoded by one character taken from alphanumeric characters. A string of characters constitutes one tag assigned to one token and lemma. A tag is then a set of characters that encodes the values of formal categories regarded as relevant at the given word-form. The number of characters varies but their order is obligatory.

Every tag is composed of two parts. The first defines the morphological and grammatical properties of a token. It always begins with a character encoding part-of-speech, followed by characters for other categories, e.g. *Lingvista anotoval texty z korpusu*. (lit. A linguist annotated corpus texts.). There are 6 tokens (*lingvista*, *anotoval*, *texty*, *z*, *korpusu*, *.* (dot)); every token is assigned a lemma (l) and a tag (t):

<i>Lingvista</i>	l = <i>lingvista</i>	t = SSms1
<i>anotoval</i>	l = <i>anotoval'</i>	t = VLescm+
<i>texty</i>	l = <i>text</i>	t = SSip4
<i>z</i>	l = <i>z</i>	t = Eu2
<i>korpusu</i>	l = <i>korpus</i>	t = SSis2
<i>.</i>	l = <i>.</i>	t = Z

The second (facultative) part specifies the token as a part of specific word classes (proper names, defective forms). In most cases, the token does not belong to any of these specific classes, as the second part of the tag is missing at that time. In the cases of proper names, after the first part we assign: (U+003A COLON) and a special character **r**. In the cases of defective or wrong forms, a colon is followed by **q**:

<i>od Minárika</i>	l = <i>Minárik</i>	t = SSms2:r
<i>Goldsteinovú tvár</i>	l = <i>Goldsteinov</i>	t = AFfs4x:q

The detection of defective forms is also instructed by the frequency of occurrence in the corpus, ultimately by a type of an „error“. If a word-form is not standard but adequate in the given cases (*neni*, *do Košicoh*, *za prvé*, *postavím sa do rady*, *prádlo*), it is not regarded as wrong. Typos and obvious spelling mistakes are viewed as defective forms.

## 4.2 GENERAL TAGGING PRINCIPLES

A category is indicated by a character if it is relevant for the given form. E.g. for the pronoun *ako* the categories of gender, number and person are not relevant; therefore we indicate only POS and a paradigm: l = *ako*, t = PD. Verbs in a base (infinitive) form cannot be assigned a category of number, person and gender congruency, therefore we indicate only POS (verb), verb form (infinitive), verb mode (completive), affirmation (affirmative): l = *vníknúť*, t = VId+.

Characters are assigned to the values of (morphological) categories relevant for the given word-form even in those cases where the categories are not “visible” from the word-form, hence they are not formally transparent. In some cases these categories can be contextually determinable; since the context is unlimited for the purpose of manual tagging. First of all we take into account congruency within syntagmas or valency relations. Specifically we also indicate the category of person in -l-participle forms on the basis of the presence or absence of the grammatical morphemes *som*, *si*, *sme*, *ste*.

For example:

<i>Nechcem cestovať v tom <b>kupé</b>.</i> (neuter, sg., L)	l = <i>kupé</i>	t = SUn6
<i>Dozvedel sa to od <b>päť</b> chlapov.</i> (masc. anim., pl., G)	l = <i>päť</i>	t = NUmp2
<i>Pozdravil sa <b>jeho</b> sestru.</i> (fem., sg., D)	l = <i>jeho</i>	t = PUfs3
<i><b>Nenašiel</b> som ani kúsok.</i> (compl., sg., 1. pers., masc., neg.)	l = <i>nenašiel</i>	t = VLdsam-

In cases such as *kupé* (coupe) and *jeho* (his), essentially there is an absolute morphological homonymy, because these words have only one form through which they enter syntactic relations and thereby they are clearly defined by the context only.

Some forms usually referred to as inflexible, tend to be declined. Declined and non-declined forms of one lemma can occur in the same context and the user can find all the possibilities and discover the development of inflection. Frequency analysis can show their occurrences ratio. For example, the substantive *Philips* (l = *philips*) was observed in SNC in these forms of the genitive singular:

*Cieľom transakcie je transformácia **Philips**...*

*...hovori O. Š. z **Philips** Slovakia...*

*...s kapitálovou pomocou **Philipsu**...*

*...veľkých spotrebičov od **Philipsa**...*

Context is taken into account even when treating the homonymy (ultimately, homography) of some word-forms in a paradigm of one lexeme:

e. g. a form <i>pekné</i> (beautiful) – possibilities:	N, A pl. masc. inanim.	t = AAmp1x	t = AAmp4x
	N, A pl. fem.	t = AAfp1x	t = AAfp4x
	N, A sg. neuter	t = AAns1x	t = AAns4x
	N, A pl. neuter	t = AAnp1x	t = AAnp4x

Context: *Dievčatá sú **pekné**.* (lit. Girls are beautiful.) The only possibility is: t = AAnp1x.

Parts-of-speech homonymy is solved with the aid of codification books and dictionaries (e.g. MSJ, 1966; KSSJ, 2003); semantics is also taken into account:

<i><b>Tuším</b> vo vzduchu búрку.</i> (I feel)	l = <i>tušiť</i>	t = VKesa+
<i><b>Tuším</b> budú problémy.</i> (maybe)	l = <i>tuším</i>	t = T
<i><b>Prosím</b> si voľu.</i> (I beg for)	l = <i>prosiť</i>	t = VKesa+
<i>Podľa, <b>prosím</b>.</i> (please)	l = <i>prosím</i>	t = T
<i><b>Lepšie</b> to nebude.</i> (better, adj.)	l = <i>dobrý</i>	t = AAns1y
<i>Vieš to aj <b>lepšie</b>.</i> (better, adv.)	l = <i>dobré</i>	t = Dy
<i>Bolo <b>zima</b>.</i> (cold – adv.)	l = <i>zima</i>	t = Dx
<i>Prišla/Bola <b>zima</b>.</i> (winter – subst.)	l = <i>zima</i>	t = Ssfs1

Particular grammatical categories can even be assigned to those abbreviations, acronyms and units of measure coined from flexible parts of speech (*nám.*, *ul.*, *č.*, *l*, *cm*), or to forms functioning as declinable parts of speech (*do SND*, *v SR*).

#### 4.3 PARTS OF SPEECH

In SNC, a set of word-forms is divided into 19 classes, ten of them reflecting traditional word classes (parts of speech) – nouns (S), adjectives (A), pronouns (P), numerals (N), verbs (V), adverbs (D), prepositions (E), conjunctions (O), particles (T) and interjections (J) – nine of them representing various and specific language elements – formal participles (G), reflexive morphemes *sa/si* (R), the conditional morpheme *by* (Y), numbers (0), abbreviations and symbols (W), unclassifiable parts of speech (Q), citation forms (%), punctuation (Z) and non-word elements (#). The traditional parts of speech basically reflect the part-of-speech

classification in Slovak codification books (KSSJ, 2003; PSP, 2000). Disputable issues concerning part-of-speech classification required some compromise solutions:

1. **Verbal nouns** (*písanie* – writing, *hovorenie* – speaking, *rešpektovanie* – respect, etc.) are treated as nouns.

2. **Agglutinated forms** (*oňho, preňho, naňho, oň, preň, zaň*) are labelled as pronouns; the prepositional part is reflected as part of a lemma (see Garabík – Gianitsová – Horák – Šimková, 2004, chapter 2.2.3.3), their state of agglutination is conveyed by formal category and its value within a tag. E.g. *Starám sa oňho* (lit. I take care of him.) l = o\_on t = PPms4g

3. **Secondary prepositions** such as *s ohľadom na* (regarding), *v závislosti od* (in dependence on), *na rozdiel od* (unlike), *v prípade* (in case) are tagged as junctions of a preposition (or, prepositions) and a substantive.

4. **Active and passive participles** and adjectives converted from verbs (*písaný, otvorený, obutý, píšuci, hrajúci, stojaci*) are regarded as transitional groups; therefore, we decided to specify these adjectival forms as a group of formal participles. They are differentiated from adjectives on the grounds of their form and origin. (Formal) passive participles are considered to be adjectival forms coined from the infinitive stems of verbs by adding participle morphemes *-n+ý, -t+ý* (*sklad-a:t → sklad-a:n-ý*). (Formal) active participles are considered to be adjectival forms coined from the present tense stems (usually from 3rd person plural forms) of verbs by adding participle morphemes *-úc+i/-uc+i, -iac+i/-ac+i* (*sklad-a:j-ú → sklad-a:j-úc-i*), e.g. *píšuci, písaný, žnúci, žatý, bijúci, bitý, spiaci, šijúci, šitý, sejúci, siaty*. Deverbative adjectives created as a result of the word-formative process of derivation are not considered to be participles, e.g. *písací, skladací, žací, bicí, spací, šijací, sejací*.

5. At this level of annotation, we do not differentiate between **reflexive pronouns** *sa, si* and *sa, si* as verbal components. Their distinction needs to be the subject of individual papers written on the basis of a corpus database; this issue would also eventually be treated on the level of syntactic annotation. In this case, the possibility of specifying the morphological categories of *sa, si* as reflexive pronouns is excluded. On the other hand, it is possible to disambiguate reflexive verb components (*pospať si, zaspievať si, zaspieval si si*) and a 2nd person, present tense, indicative form of the verb *byť* – *to be* (*ty si klačal, zaspieval si si*).

6. **Morpheme by**, a part of the conditional verb-form, is tagged as an independent word class. Other forms with the morpheme *by* (e.g. conjunctions *keby, aby, žeby, akoby, staby*, particles *aby, keby*) are tagged as conjunctions or particles (according to the function they have) but we take into account the agglutination of the morpheme *by*. This fact is reflected by adding the (Y) character – conditionality.

7. **Numeral characters** (Roman as well as Arabic), eventually numeric symbols and combinations of numeral characters are assigned to an independent class: “numbers” (0).

8. **Symbols** such as *l, km, H<sub>2</sub>O, X569847* and **abbreviations** such as *atď., tzv., t. j., pod., kt., i., XML, SND* fall into the class of “abbreviations and symbols” (W). On the other hand, abbreviated words such as *Satur, Slovnaft, Rempo* fall into the class of substantives on the basis of their function and meaning.

9. **Multi-word lexemes** in SNC are composed of several tokens. The first and non-individual parts (words) of a larger lexical unit are often impossible to define. That is why they fall into the class of “unclassifiable part of speech”: *po slovensky, fast food, Los Angeles*. Forms such as *slovensko-český, 2-krát* are indicated as three tokens: *slovensko* (unclassifiable part of speech), – (punctuation mark), *český* (adjective); 2 (number), – (punctuation mark), *krát* (unclassifiable part of speech). A similar approach is applied when dealing with “juxtapositions” *až60* (unclassifiable part of speech).

10. **Citation forms** (%) include foreign multi-word phrases and sentences not adapted into a second language but functioning as parole units taken from a source language: *Take it easy!*; *Šaj pes dovakeras*; „správne vychladená dvanásťka“; *Ta naše povaha česká; náměstí*. These tokens do not need to be indicated by quotation marks. Individual words of foreign origin such as *kuskus, ska, sitar, djembe, česnečka, květák* do not fall within this class because in Slovak sentences they function in accordance with Slovak grammatical rules.

#### 4.4 CATEGORIES AND THEIR VALUES

In the matter of morphological annotation, our starting-point was the theory of grammatical categories introduced by MSJ (1966), and ultimately by other works dealing with morphology. In a tag-set there are indeed some categories with their values not explained and mentioned by traditional morphology (paradigm, verb form, agglutination, conditionality). There is a formal-morphological characteristic important in the process of token disambiguation.

In nominal parts of speech, the second position is occupied by a character indicating the type of **paradigm** with values: substantival, adjectival, pronominal, numeral, combined, uncompleted and adverbial.

The formal attribute “paradigm” is understood as the specification of a form of a particular word within a word class (e.g. *taký* (lit. such) is pronoun, but it has the form of an adjective, its tag is PA). Characters standing in for the substantival, adjectival, pronominal, numeral and adverbial paradigms are identical with the part-of-speech indicators (S, A, P, N, D).

**Combined paradigm** (F) is valid for words having a partial congruent paradigm. The development of their declension has undergone complicated processes and they are not unambiguously assignable to clearly-defined declension types. This class contains words such as *kuli, gazdiná* and nouns declined in the same way, *otcov, matkin* (all individual possessive adjectives), *on, ona, ono, kto, čo, nikto, nič...*, *môj (tvoj, náš, váš), ten (tá, to), sám, onen, žiaden, všetok, jeden*.

**Uncompleted paradigm** (U) is assigned to those substantives, adjectives, pronouns and numerals traditionally considered to be inflexible (*kupé, super, jeho*), or with a tendency to be declined (*kanoe* – only G pl. *od kanoí*), or usually inflexible (*pani, kolko, tolko, viacero, päť, sto, tisíc*). In these cases, the distribution of declined and inflexible forms depends on various circumstances. The form of such a substantive of foreign origin which is essentially declinable can be regarded as a noun with uncompleted paradigm (*Phillips, Tesco*) but the author of a text also prefers an inflexible form in a given case (genitive singular *od Phillips* – SU), even though the declined form is prevalent (genitive singular *od Phillipsu* – SS).

**Adverbial paradigm** (D) is assigned to inflexible pronouns and numerals, in KSSJ indicated by grammatical labels *neskl.* (inflexible) or *príslov.* (adverbial), eventually they function as adverbials (*kolkonásobne, kolkorako, tam, tu, vtedy, vždy, viacnásobne, dvojako*). In these cases, the tag contains only this category of information. Other categories such as gender, number or case are not indicated.

The first two characters of a tag present the following combinations: **SS** – substantive with substantival paradigm (*mama*), **SA** – substantive with adjectival paradigm (*vedúci*), **SF** – substantive with combined paradigm (*gazdiná*), **SU** – substantive with uncompleted (“invisible”) paradigm (*pani, kanoé*), **AA** – adjective with adjectival paradigm (*pekny*), **AF** – adjective with combined paradigm (*otcova*), **AU** – adjective with uncompleted (“invisible”) paradigm (*super*), **PS** – pronoun with substantival paradigm (*kolkrátka*), **PA** – pronoun with adjectival paradigm (*taký*), **PP** – pronoun with pronominal paradigm (*ja*), **PF** – pronoun with combined paradigm (*on, sám, žiaden*), **PU** – pronoun with uncompleted (“invisible”) paradigm (*jeho, jej, ich, kolko, tolko*), **PD** – pronoun with adverbial paradigm (*tam, niekedy*),

NS – numeral with substantival paradigm (*milión, raz*), NA – numeral with adjectival paradigm (*štvrtý*), NN – numeral with numeral paradigm (*tri*), NF – numeral with combined paradigm (*jeden*), NU – numeral with uncompleted (“invisible”) paradigm (*päť, sto, tisíc, päťoro, veľa*), ND – numeral with adverbial paradigm (*dvakrát, mnohonásobne*).

As far as nominal parts of speech are concerned, the indication of category is usually followed by the elementary morphological characteristics:

- **Gender:** masculine animate (**m**); masculine inanimate (**i**); feminine (**f**); neuter (**n**); unspecified (**o**); general (**h**), (this last holds true for pronouns and verbs);
- **Number:** singular (**s**), plural (**p**) and unspecified (**o**);
- **Case:** nominative (**1**), genitive (**2**), dative (**3**), accusative (**4**), vocative (addressing) (**5**), locative (**6**), instrumental (**7**), unspecified (**o**);
- **Grade:** base form (or irrelevant grade) (**x**), comparative (**y**), superlative (**z**); this holds true for adjectives, adverbs and formal participles.

Value “unspecified” (**o** character) in the position of gender, number or case is relevant for some morphologically non-transparent or homonymous forms if the context indicates several conflicting values for one category (*Mužov, žien a detí je päť. Kúpili kanoe. Mesto!*).

The second verb position “**verb form**” can be assigned to the following values: infinitive (**I**), formal present (indicative) (**K**), imperative (**M**), transgressive (**H**), *-l*-participle (**L**), future form (most commonly this is a form of the verb *byť*, also the synthetic future tense of uncompleted verbs – *poletím, ponesiem* etc.) (**B**). The establishment of the category of “verb form” resulted from our attempt at a description of analytic verb forms. Even though we do not regard this solution as ideal, for the time being it represents a systematic approach to this complicated issue. The categories of tense and modus are not indicated individually because they are included in particular definite verb forms being indicated as follows:

#### Indicative

	Verb form	Example	Lemma	Tag
present tense	formal present	<i>píšem</i>	<i>písať</i>	VK...
past tense	<i>-l</i> -participle + formal present	<i>písal som</i>	<i>písať</i> formal <i>byť</i> <sup>*</sup>	VL... VK...
future tense (uncompleted verbs)	future + infinitive	<i>budem písať</i>	formal <i>byť</i> <i>písať</i>	VB... VI...
	future	<i>ponesiem</i>	<i>niešť</i>	VB...
future tense (completed verbs)	formal present	<i>napíšem</i>	<i>písať</i>	VK...

#### Imperative

present tense	imperative	<i>píš!</i>	<i>písať</i>	VM...
---------------	------------	-------------	--------------	-------

#### Conditional

present tense	<i>-l</i> -participle + conditional morpheme + formal present	<i>písal by som</i>	<i>písať</i> <i>by</i> formal <i>byť</i> <sup>*</sup>	VL... Y VK...
past tense	<i>-l</i> -participle + conditional morpheme + formal present + <i>-l</i> -participle	<i>bol by som písal</i>	formal <i>byť</i> <i>by</i> formal <i>byť</i> <sup>*</sup> <i>písať</i>	VL... Y VK... VL...

\* if present



Other verb values (where relevant) are as follows:

- **Aspect:** perfective (**d**), imperfective (**e**), with both aspects (**j**); its indication is made on the basis of dictionary qualifiers (in KSSJ, 2003);
- **Number:** singular (**s**), plural (**p**);
- **Person:** first (**a**), second (**b**), third (**c**); indicated also in -l-participle forms and grammatical morphemes (forms of the formal verb *byť*);
- **Gender congruency:** congruency of masculine animate (**m**), inanimate (**i**); feminine (**f**); neuter (**n**); undefined (**o**), general gender (**h**) – relevant only for -l-participle verb forms;
- **Negation:** affirmative (+) and negative (-); relevant only for verbs.

When referring to pronouns, numerals, formal participles, prepositions, conjunctions, particles and the morpheme *by*, some other attributes are indicated. These are usually formal features we regarded as important when making a formal description of forms:

- **Agglutination:** a character for this attribute (**g**) is assigned to pronouns at the end of a tag, where relevant (forms such as *oňho, preňho, naňho, oň, preň, zaň* etc.).
- **Independent use of numerals** takes the second position in a tag (X), if it is an independent expression of quantity, e.g. mathematical operations ( $2 - 2 = 0$ ), specific nominative (*Dráma 2000*) etc.
- **Type of participle:** active (**k**) and passive (**t**) takes the second position so far as formal participles (**G**) are concerned (*píšuci, písaný*).
- **Form of preposition:** vocalized (**v**) or non-vocalized (**u**) takes the second position in a tag (*vo, v; ku, k; so, s*).
- **Conditionality:** holds true for conjunctions or particles *keby, aby, žeby, akoby, sľaby* (OY, TY).

Examples illustrating the previous categories from the SNC tag-set can also be found in the current version on the website.<sup>9</sup> There is a more detailed description of the system of categories, their values and concrete solutions. On the basis of this tag-set, the manual tagging of the Slovak National Corpus is carried out.

## 5 MANUAL TAGGING OF SNC – INITIAL RESULTS AND POSSIBLE PERSPECTIVES

Since the morphological annotation of the corpus requires, in addition to tag-set and computational tools (tagger), morphological dictionary and text data (namely, training corpus and testing data), we consider manual tagging to be an important step towards obtaining the material. In the first half of 2004, the manual tagging of the novel 1984 by G. Orwell and texts from the internet magazine InZine was started. The annotation is being carried out by three students from the Faculty of Philosophy of Comenius University in Bratislava and a working group of eleven students (Faculty of Arts, University of Prešov; the cooperation with Prešov is performed on the basis of the grant project *Morfosyntaktická analýza SNK* (VEGA 1/3149/04; *Morphosyntactic analysis of the SNC*).

Even though manual annotation is time-consuming, by the end of May 2004 we managed to obtain a set of texts including about 19,000 tokens from Orwell's novel 1984 and about 25,500 tokens from the internet magazine InZine. It should be emphasised that these are manual annotations that have not yet been checked and unified in accordance with current annotational principles. We are going to deal with this problem in the future. The average accuracy ratio of manual annotation is about 91.5 % and average speed is about 80 tokens

<sup>9</sup><http://korpus.juls.savba.sk/publications>



a minute. On the other hand, we regard manual annotation as important. The material so acquired would be useful for further tagger training. However, we consider it necessary to speed up the process of manual morphological annotation and render it more effective by means of automatically preprocessed annotated texts. The human annotator is subsequently given an automatically annotated text and he/she should decide whether or not the given tag is correctly assigned (in the latter case, a correction is required).

The next period of time should be devoted to manual annotation along with the testing of appropriate tools and their applications for the tagging of Slovak texts. On the basis of our co-operation with the Institute of Formal and Applied Linguistics (ÚFAL) at the Faculty of Mathematics and Physics, Charles University, Prague,<sup>10</sup> we have at our disposal a morphological analyzer and disambiguator developed by J. Hajič; we are also going to use a Slovak version of the morphological analyzer proposed by R. Sedláček and M. Grác (Masaryk University, Brno). We anticipate that this co-operation will be successful.

## REFERENCES

- DVONČ, L. (1984): *Dynamika slovenskej morfológie*. Veda, Bratislava.
- EAGLES (1996). Recommendations for the morphosyntactic annotation of corpora. EAG-CSG/IR-TR.1. ILC-CNR, Pisa. <<http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html/>>
- ERJAVEC, T. (2001): Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984. In: 6th Natural Language Processing Pacific Rim Symposium, NLP'01, Tokyo, pages 487-492. <<http://nl.ijs.si/et/Bib/NLP'01/mte-nlprs01.pdf>>
- FORRÓOVÁ, M., HORÁK, A. (2004): Morfológická anotácia korpusu. In: Proceedings of International Conference Slovenčina na začiatku 21. storočia. Prešov, 174 – 183.
- FORRÓOVÁ, M., GARABÍK, R., GIANITSOVÁ, L., HORÁK, A., ŠIMKOVÁ, M. (2003, to be published): Návrh morfológického tagsetu SNK. <<http://korpus.juls.savba.sk/publications>>
- GARABÍK, R., GIANITSOVÁ, L., HORÁK, A., ŠIMKOVÁ, M. (2004): Tokenizácia, lematizácia a morfológická anotácia Slovenského národného korpusu. (Current version of May 4, 2004). <<http://korpus.juls.savba.sk/publications>>
- HAIJČ, J. (2000): Popis morfológických značiek – poziční systém. ÚČNK – ÚFAL MFF UK, Praha. <<http://ucnk.ff.cuni.cz/manual/znacky.html>>
- HANA, J., HANOVÁ, H. (2002): Manual for morphological annotation. ÚFAL MFF UK, Praha.
- Krátky slovník slovenského jazyka (2003). 4<sup>th</sup> edition. eds. J. Kačala, M. Pisárčiková, M. Považaj. Veda, Bratislava. (KSSJ)
- LEECH, G. (2000): Anotáční systémy pro značkování korpusu. In: Studie z korpusové lingvistiky. Acta Universitatis Carolinae – Philologica 3 – 4. Karolinum, Praha, p. 185 – 197.
- Morfológia slovenského jazyka (1966). Ed. J. Ružička. Veda, Bratislava. (MSJ)
- ORAVEC, J. – BAJZÍKOVÁ, E. – FURDÍK, J. (1984): *Súčasný slovenský spisovný jazyk*. Morfológia. SPN, Bratislava.
- Pravidlá slovenského pravopisu (2000). Ed. M. Považaj. Veda, Bratislava. (PSP)
- PRZEPIÓRKOWSKI, A. – WOLIŃSKI, M. (2003): A Flexemic Tagset For Polish. In: Proceedings of Morphological Processing of Slavic Languages, EACL 2003, Budapest, pages 33 – 40. <<http://dach.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws12/ws12.pdf>>

<sup>10</sup> Grant project *Využitie spoločných vlastností češtiny a slovenčiny na budovanie anotovaných národných jazykových korpusov*; lit. *The Application of common properties of Czech and Slovak languages for the purpose of building annotated national language corpora*.

## RESUME

V príspevku sa zaoberáme problematikou morfolologickej analýzy Slovenského národného korpusu. Automatická morfológická anotácia textov predstavuje aplikáciu, ktorá sa teší veľkej pozornosti najmä v súvislosti so spracovaním korpusových dát a vyžaduje si prípravu morfológického tagsetu pre ručné značkovanie SNK. V úvodnej časti predstavujeme úvahy a východiská, ktoré viedli k voľbe spôsobu spracovania. Otázkam morfolologickej analýzy (vymedzeniu základných pojmov) a jej aplikácii pri značkovaní korpusu sa venujeme v teoretickej rovine. Zdôrazňujeme rozdiel medzi analýzou, ktorú vykonáva človek, ktorý je jazykovo kompetentný, a tou, ktorú vykonáva počítač. Stáli sme pred rozhodnutím, či navrhnutý súbor značiek bude výsledkom nového formálneho opisu jazyka alebo budeme vychádzať z lingvistických opisov, ktoré existujú a len sa ich pokúsime formalizovať. Za dôležitú považujeme najmä zásadu prístupnosti (koncovému používateľovi, ktorým býva aj nelingvista) a zásadu konsenzu vedeckých teórií, teoretickej „neutrality“. Preto sme sa rozhodli prihliadať najmä na systematický opis podaný v akademickej Morfológii slovenského jazyka (ďalej MSJ, 1966), prípadne na ďalšie morfológické práce (Oravec – Bajžíková – Furdík, 1984; Dvonč, 1984). Konflikt medzi reprezentáciou zaužívaných gramatických kategórií (často s nejasnými kritériami morfolologickej klasifikácie) a možnosťou automatického spracovania jazyka sa odráža aj v koncepcii morfológického tagsetu pre SNK, ktorý vychádza aj zo skúseností zahraničných tvorcov tagsetov pre morfológickú analýzu. Pre potreby značkovania bol preto zvolený formálno-gramatický princíp, ktorý však s ohľadom na domácu lingvistickú, resp. gramatickú tradíciu má isté špecifiká.


V nasledujúcich častiach upozorňujeme na všeobecné črty a niektoré konkrétne riešenia problémov morfolologickej anotácie textov SNK. Tie boli ovplyvnené i prístupom k segmentácii textu na tokeny, preto sa krátko najprv zmiňujeme o tokenizácii textov SNK a jej zásadách. Tokeny sa (s ohľadom na počítačový prístup) nemôžu kryť a ani sa nekryjú s pojmom slovo, či dokonca gramatický tvar. Tokenizácia je dôležitou etapou v automatickom spracovaní textu, pretože od jej výsledkov je priamo závislá morfológická analýza a dezambiguácia. Navrhnuté zásady tokenizácie môžu vyvolávať otázky pri zložených tvaroch, pri zápisoch so spojovníkom či pomlčkou (často aj chybné použitými), pri analytickom stupňovaní, združených pomenovaniach, zložených číslovkách a na druhej strane pri aglutinovaných podobách, keďže tieto lexikálne jednotky sú rozdelené na viac tokenov (napriek tomu, že tvoria jednu jazykovú jednotku) alebo zlúčené do jedného tokenu (napriek tomu, že ide o dve pôvodné jednotky). Tento návrh tokenizácie vedie k takej interpretácii slov a gramatických tvarov, ktorá nie vždy súhlasí s tradičnou lingvistickou tradíciou. Súčasné riešenie však nevylučuje možnosť zapojenia logického modulu do spracovania textu, ktorý by sa neskôr uplatnil ako vhodnejší základ lematizácie a morfológického značkovania. V časti o lematizácii textov SNK zdôrazňujeme, že aj pojem lemy neaplikujeme absolútne v jeho významovom rozsahu. Charakterizujeme niekoľko základných zásad a uvádzame niektoré zvláštne prípady.

Predstavenie zásad morfológického značkovania a formy zápisu značiek tvorí strednú časť príspevku. Pri morfolologickej notácii volíme kombinovaný pozično-atribútový spôsob. Zaujímavosťou je rozdelenie tagu na dve časti. Druhá (nepovinná) časť zaraďuje token do určitých špeciálnych skupín (ako sú vlastné mená alebo defektné zápisy). Ako všeobecnú zásadu sme prijali uvádzanie znakov pre jednotlivé atribúty aj vtedy, keď síce je hodnota pre daný tvar relevantná, ale nie je z formy slova „viditeľná“, teda nie je dostatočne formovo transparentná. V niektorých prípadoch ich môžeme určiť z kontextu, ktorý je pre potreby ručného značkovania neobmedzený. Prihliadame najmä na kongruenciu v rámci syntagmy alebo na valenčnú väzbu. Podobne sa rieši aj „tzv.“ nesklonnosť substantív, adjektív a i. Ich tvary totiž pova-

žujeme za absolútne morfológické homonymá, keďže tieto slová majú len jednu formu, ktorou sa zapájajú do syntaktických vzťahov a väzieb. Tým sú však jasne definovateľné z kontextu. Navyše, niektoré tvary, uvedené v odbornej literatúre ako nesklonné, sa v bežnej praxi začínajú skloňovať. Vysklňované a nevysklňované tvary jednej lexémy sa môžu v istom páde vyskytovať popri sebe a používateľ má možnosť nájsť všetky prípady a zistiť pokročilosť flektivizácie. Pri slovnodruhovej homonymii sa v zásade riadime kodifikačnými príručkami a sémantikou.

Nasleduje stručný opis súboru morfológických značiek. Množina slovných foriem používaných v slovenčine sa v morfológickom tagsete SNK rozdeľuje do 19 tried, z ktorých 10 v zásade zodpovedá tradične vydeľovaným slovným druhom a 9 obsahuje rôzne špecifické jazykové prvky. Pri klasických slovných druhoch sa v podstate rešpektuje slovnodruhové zaradenie podľa súčasných slovenských kodifikačných príručiek. V sporných otázkach určenia slovného druhu sme prijali niektoré kompromisné riešenia, ktoré v príspevku predstavujeme. Pri morfológickom značkovaní sme vychádzali z teórie gramatických kategórií, ako ich podáva akademická Morfológia slovenského jazyka (1966), príp. iné morfológické práce. V tagsete SNK sa však stretneme aj s kategóriami a ich hodnotami, o ktorých sa v tradičnej morfológii ako o kategóriách neuvažovalo (paradigma, slovesná forma, aglutinovanosť, kondicionálnosť). Ide o formálno-morfológické charakteristiky, ktoré sú dôležité na zjednotenie tokenu. Z nich najviac miesta dostáva najmä objasnenie kategórie paradigmy, ktorú chápeme ako vymedzenie špecifickej formy konkrétneho člena slovného druhu. Kategória slovesnej formy je zas výsledkom pokusu o uchopenie analytických tvarov slovies. I keď si uvedomujeme, že toto riešenie nie je ideálne, na súčasnej úrovni znamená systematické uchopenie tejto náročnej problematiky.

Na základe tohto tagsetu v súčasnosti prebieha ručné značkovanie textov, ktoré obsahuje Slovenský národný korpus. Ide o román Georgea Orwella 1984 a texty z internetového časopisu InZine. Prvé výsledky a okolnosti ručného značkovania SNK a možné perspektívy ďalšieho rozvoja a zefektívnenia práce predstavuje záverečná kapitola príspevku.



# Options for the Generation of a Corpus-Based Slovak Morphology (as Part of Corpus Morphosyntax)

MILOSLAVA SOKOLOVÁ

## 1 INTRODUCTION

The project I direct has been running since January 2004. It is a grant-funded project of the Ministry of Education of the Slovak Republic VEGA 1/3149/04 *Morphosyntax research within the Slovak National Corpus*. 15 solvers are involved, from University centres as well as from the Slovak Academy of Sciences (see project website). Thanks to Ms. Šimková (who engaged me for the corpus research) and to her team from the SNC Department and lecturers from the Department of Computation Linguistics in Prague, a series of lectures and briefings was held at Prešov University concerning work with the corpus. The audience comprised the project solvers and ca 20 students, whom I had engaged in the project, in the corpus linguistics seminars. The following workshops of the solver team took place up to June 2004:

Morphological corpus annotation;

Use of the corpus in the case of *Dictionary of root morphemes of the Slovak language*;

Research into the frequency of Slovak grammar forms and their valences;

Website generation.

One of the opinions on the tagset *Tokening, lemmatisation, and morphological annotation of SNC* was developed by me (Garabík – Gianitsová – Horák – Šimková, 2003) within the framework of the project Morphosyntactic research within the Slovak National Corpus, as well as the present draft of the Slovak Corpus Morphosyntax Concept, as a task for which I am in part responsible. Our work is based on English corpus grammars, in particular the *Longman Grammar of Spoken and Written English* (1999), but, unlike the above grammar, the Slovak Corpus Morphosyntax Concept will be based only on written texts from the following areas: journalistic, professional, and artistic texts, conversation (colloquial) texts being excluded.

## 2 BRIEF ANALYSIS OF THE EXISTING SITUATION IN SLOVAKISTICS

It is the case that no unified concept of a new grammar has been reached following several years of discussions on Slovak grammar (in particular in the period from 1993, when J. Dolník referred to natural morphology at the slavistics congress in Bratislava, to conferences in recent years – see the collection *Tradícia a perspektívy gramatického výskumu na Slovensku* (Tradition and Perspectives of Grammar Research in Slovakia), 2003, and the collection from the conference in 2002, which is under preparation), which would meet with general acceptance. There are further reasons for this situation:

A) On the one hand, there are only a limited number of slovakistics linguists. Slovakistics, being the linguistics of a small culture, cannot objectively produce grammars of so many

types and orientations as those of a large culture. Therefore, selection and generalisation are inevitable.

Specific position of linguistics of small cultures compared with large ones, with linguistics in Anglophone and/or Germanophone countries. The subjects of research in institutions with large teams, not just in the USA, Great Britain and Germany, but also in Russia or Poland, are covered by single persons in slovakistics. This situation has its pros and cons. A wider focus of interest is inevitably required from slovakistics researchers, who must be more universal. Their knowledge is more shallow (they focus either on the material or only on theory), they use greater selection, they lag behind European trends, etc.

B) The limited number of linguists is combined with a lack of readiness in the current generation to co-operate on a united grammar. They expend their energy in unconstructive disputes (e.g., including those between analogists and anomalists).

There lacks an Isačenko-like linguist. My opinion is that the high quality of the Slovak Language Morphology was influenced also by the fact that it was created on the back of Isachenko's Grammar Machine.

A solution to this situation is seen in the work of linguists in more areas, on more pillars, relying on the synergic effect of such research.

A concentration of organisation of research in JÚEŠ SAV would be ideal, with the participation of all University centres. The estimated time-horizon is ca 10 – 15 years.

The involvement of all morphologists and syntactists can be efficiently implemented through a website (see Štícha) and Internet workshops with a wide scope of morphosyntactists, (which would take place at least monthly). Involvement of foreign linguists (the following linguists can be considered at present: Lüdtke-Nolte, Späth; Giger, Musilová; Polish Slovakists Mieczkowska; Orwińska-Ruziczka, Szymczak).

What bases are currently available in slovakistics? Several areas will be listed with (in my view), their pros and cons:

## 2.1 system and linguistic pillar:

The good old Slovak Language Morphology (1966; Pauliny, 1981; Dvonč, 1984; Oravec et al., 1984; Najnowsze dzieje języków słowiańskich. Slovenský jazyk, 1998) as the base and new West-Slavic (Czech, Mluvnice češtiny, 1986 – 1987; Grepl et al., 1986; Příruční mluvnice češtiny, 1996; Čechová, M. et al., 2000; 2003, Štícha, 2003; Polish, Grzegorzczkova et. al., 1984; Upper-Lusatian, Faßke, 1981); Russian, Russkaja gramatika, 1980; German grammars (Grundzüge einer deutschen Grammatik. (Main Features of German Grammar.), 1981; Wurzel, 1984), and English corpus grammars (Coll. Longman Grammar of Spoken and Written English, 1999).

1. BĚLIČOVÁ, H.: Nástin porovnávací morfologie spisovných jazyků. (Outline of comparative morphology of standard languages.) Prague, Karolinum, Nakladatelství Univerzity Karlovy 1998. p. 217. (71 – 77);

2. Encyklopédia jazykovedy. (Encyclopaedia of Linguistics.) Editor J. Mistrík, 1<sup>st</sup> edition. Bratislava, Obzor 1993 (morphological headings).

3. *Teoretické základy synchronní mluvnice spisovné češtiny. (Theoretical Principles of Synchronous Czech Grammar.) Slovo a slovesnost, 1975, p. 18 – 46.*

**Cons:** The language material contained in the Slovak Language Morphology is 50 years old. It is based on artistic texts which tend towards folklorisation. Unlike J. Dolník, I do not feel that, in this context, new material is not necessary for explanation, while even Dvonč's material (1984) is not satisfactory, as theoretical and evaluation aspects are often missing. We

need a grammar which would utilize the technology of the 21st century to make up for the deficit in Slovakistic grammar researches (new material dating from 1955 onwards as a marked shift, quoting M. Dudok (2003), from texts belonging to the artistic domain to journalistic and scientific texts).

**2.2 Slovakistic works after 1966, which are unified by structuralism and are mainly system- and cognitive-linguistic-oriented:** Horák, 1993; Horecký et al., 1989; Kačala, 1989, 1998; Kořenský, 1984, 1998; Mieczkowska, 1994; Nábělková, 1993; Nemcová, 1990; Nižníková et al., 1998; Ondrejovič, 1989; Ondrus, 1978; Oravec, 1967; Orwińska-Ruziczka, 1992; Polański et al., 1984 – 1992; Ružicková, 1982; Sekaninová, 1980; Sokolová, J., 2004; Sokolová, M., 1993, 1995, 1999; Svozilová et al., 1997; Šikra, 1991.

**Cons:** The monographs do not cover all areas. There is a lack of monographs on particles, conjunctions, aspect, gradation (if foreign Slovakists are not taken into account). Interjections are not treated, as well as numerals and pronomines. Moreover, there is a problem with the different concepts and varying level of the monographs.

**2.3 The third pillar consists in principles of explication and the theory of natural grammar (Mayerthaler, 1981; 1998; Wurzel, 1984).**

There are in existence the results of more than ten years of research around J. Dolník and P. Žigo (Dolník et al., 2001; Dolník, 2000; the second collection, Dolník et al., 2003 denotes a marked qualitative jump), which can be used as an “explanation supervisor” when creating the new grammar (when interpreting corpus findings).

The theory of principles is very attractive to linguists. However, it only makes sense to have an integrated and hierarchicised system, not just fragmented research (2001). Therefore, the following actions are needed: unification of terminology; differentiation of the principle and the “principle”; selecting and hierarchicising the principles (universal principles and those with limited scope within the framework of languages tiers); definition of boundaries between description and explanation (a good description always contains elements of explanation, and is better than a bad explanation or an explanation without good language material, which is used inadequately, or only for illustrative purposes). The approach on the basis of principles increases the level of abstraction and universality of scope. Nevertheless, only a few linguists are able to perform such research on an equal level of quality, while form is abandoned here.

**Cons:**

- ⇒ Processing of principles on different levels and many unsorted principles not covered by any hierarchy\*;
- ⇒ Lack of unification of terminology; free substitution of the following terms: principle; canon; rule; law; regularity;
- ⇒ Different views on natural grammar (Mayerthaler – perception; Wurzel – systemisation; Werner – frequency);
- ⇒ Dolník's words in his introduction (2003) engender pessimism, when saying, that it is “still” a tentative foray – after ten years of research.

\*D. Slančová (1996) provides the most coherent hierarchy of rules, principles, and canons: Interaction rules: the principle of co-operative (the canons of quantity, quality, relation, and mode); of politeness (the canons of tact, generosity, satisfaction, sympathy, and contact), of irony; contextual rules: the principle of sequence, clarity, economy, expressiveness, ...), while other linguists work with those terms more or less loosely.



J. Dolník (1999) uses the term “principle” in 25 contexts as follows: the principle of analysis, analogy, anomaly, arbitrariness of language feature, distinctiveness, length of word, **dominance**, equality (when classifying phenomena into classes!), modifications of equivalence: imaginary, fictive), fictive classification, implication, cooperation, contrast, conversation implicatures, quality, quantity, mode, morphosemantic transparency, noesis, symptomatology, cognition of lexicon, cognition of semantic construction, relevance, rhythm, snores, language construction, – sonority.

By way of illustration, J. Dolník presents such problems as – **synergy of principles**: the principle of analogy in word formation, of sonority, of contrast on phonology, of conversation implications, of quality, quantity, relevance, and modus – **conflict of principles**: the principle of morphosemantic transparency, of word length, of anomaly, of shape analogy, rhythmic principle, principle of quantity and quality, of arbitrariness of language sign.

P. Žigo (1999) 11x: the principle of language development – the principle of analogy as a dominant development model, the principle of economy, of naturalness, development as removal of symptomatology, the principle of language functioning – cognitive, pragmatic, and parole principles, socio-linguistic, and stylistic principles.

E. Bajžíková (1999) 21x: **the principle of connectivity**: the principles of co-ordination, subordination, co-reference, connection, semantic modification, grammar modification, naturalness, simplicity – the principles of decomposability, dimerousness, connection, dominance, hierarchy, linearity, equality, equivalence, substitution, connection, gradual development, compatibility.

J. Kačála (1998) 23x: **the principle of construction**: the principle of update, grammar principles, the principle of activity, connectivity, word order, lexical-semantic connectivity, activity, word order, information principle, universal principle of core and periphery, hierarchic principle, **connectivity (defined as a principle on page 22, and as a rule on page 80)**, **complex principles**: principle of update, grammar principles, information principle, **partial principles**: **principle of connectivity**, order, dependence, syntagmatic principle, lexical principle, intention principle.

**2.4 The pillar which can be used as a base, as in the following communication-pragmatic researches:** monographs (Slančová, 1998, Kesselová, 2001, 2002, Kupcová, 2004, Kášová, 2003); Collections SOCIOLINGUISTICA SLOVACA 1 – 5.

The contribution of communicative researches is their attractive focus on practice, (the research into children’s language should result in practical works oriented towards the removal of speech defects, cf. Slančová, 1999, talkshow research as a manual of efficient talkshow, etc.). The research is based on communicative functions. There are complex relations between communicative functions and forms: some harmonious, but mainly discordant ones, cf. Dittmann J.: *Konstitutionsprobleme und Prinzipien einer kommunikativen Grammatik*. (Constitution Problems and Principles of Communicative Grammar.), 1994.

#### Cons:

- \* the existing researches are only partial and fragmented
- \* there are too few relevant researches
- \* complex relevant (organized) research cannot realistically be performed soon.

Communicative research introduces many questions for the creation of a Slovak communicative grammar.

For instance, I was very disappointed after reading the habilitation paper of J. Kesselová with the promising title *Morphology of Communication* (despite its undoubtedly high level), since it focused on a partial problem of morphology, i.e., the frequency of word-classes in children’s communication, in the limited scope of age.

#### 2.5 New computer and corpus researches create a pillar

The following sources bring new stimuli to linguistics: the inclusion of technology in linguistic researches from E. Páleš (1994), which is still inspirational; such papers as Benko – Hašanová – Kostolanský, 1998; essays and collection *Slovenčina a čeština v počítačovom spracovaní*, ed. A. Jarošová (2001); essays of E. Kostolanský, 2003; current results of the Slovak National Corpus Department (essays, tagset). The results of the following teams from Czech workplaces

can serve as a basis: F. Štícha et al. (2003: corpus era); Hajič, 1998; two valence corpus dictionaries of J. Panevová's team. Moreover, Lopatková et al., 2002; English corpus grammars (Col. *Longman Grammar of Spoken and Written English*, 1999); English – American, Polish and German corpus researches are available.

1. Col.: Longman Grammar of Spoken and Written English. Pearson Education Limited 1999.
2. LOPATKOVÁ, M. – ŽABOKRTSKÝ, Z. a kol.: Tektogramaticky anotovaný valenční slovník českých sloves. Praha, Universitas Carolina Pragensis 2002, 99 p.
3. PÁLEŠ, E.: Sapfo. Parafrázovač slovenčiny. (Sapfo: Paraphraser of Slovak Language.) Bratislava, Veda 1994. 305 p.
4. HAJIČ J. (1998): Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Hajičová, E. (ed.): Issues of Valency and Meaning, Praha: Karolinum, pp. 106 – 132.
5. ...

### Cons:

- ⇒ the absence of a corpus which would be balanced from the point of view of styles and genre;
- ⇒ limiting factors of the search- program BONITO; time demanding (experience from 10 weeks' intensive work with students);
- ⇒ real results require the performance of complex operations (Šimandl, Hošnová, 2004)

Problems following on from the difference between intuitive human deduction and consequent computer deduction (interference of language and of people – non-consequence and game as a normal situation; the computer is ineffective since it cannot recognise interference, and is limited only to logic and causality).

### Illustration:

Researches with FF PU students performed to date:

Except for GT of the future tense (*budem mať, budem sa mať, mať budem mať sa budem*) and of the past perfect tense (*bol som mal, by som robil, by som sa robil, by som si robil, je robený...*), real numbers can be obtained only on the basis of additional operations:

number from corpus	subtract frequency	subtract frequency	subtract frequency	real number
mám	–(mám sa	+ sa mám	+ sa.* (.* mám) =	real number

Using the method of trial and error, I arrived at the criteria for the selection of an ideal representative of conjugation classes:

- \* high frequency, but small number of lexions (single lexion in ideal case)
- \* regular paradigm (-ám, not -iam)
- \* primary in the aspect pair imperfective – perfective, or imperfectivum tantum
- \* verb without reflexive motivates (with DM, PLM sa and si.)

### ⇒ confirmation of hypotheses

from research into the paradigm of the verb *mať/robiť*;

negative forms are secondary;

reduced frequency of 2nd person forms ;

undocumented forms of past perfect tense and present perfect conditional

### ⇒ surprises

\* active participle of present perfect tense as a fictive grammar form (3778 tokens from 178 000 ones, reduced by filtration to 274 cases, which are repeated for ca 60 verbs!!!

CONCLUSION: this is a lexicalised grammar form, which does not belong to morphosyntax)

\* unexpectedly high frequency of directive adverbial *na Slovensko* 6 000 / and static adverbial *na Slovensku* 66 000

\* high disproportion between the “basic” forms from grammars and inversion forms from corpus:

mali sme 4661

sme mali 11 862

mal by som 366

by som mal 1 359

#### left collocations of *mať* (frequency more than 50)

Subject	Object	Adverbials
my, ja, vy, oni, on; kto, čo, ktoré, aké, ten, každý	to, aké, ktoré, ho,	tu, tak, teraz, kde, sa
ľudia, ženy, deti, strany, krajiny, človek, štát, Slovensko, zápas	taký pocit	conjunction keby, keď

#### right collocations *mať* (frequency more than 50)

*mať* categorical (compare KSSJ – formally *mať*) as an operator

Subject	Object	Adverbials
	pocit, dojem radosť, chuť, záujem pravdu právo, dôvod, čas, problémy, skúsenosti možnosť, šancu, k dispozícii význam, zmysel v pláne, na mysli strach, smolu (šťastie)	rád (rada, radi) voľno

Despite the above, corpus research has an extraordinary potential. As the characteristic tool of the current era of globalisation, computers can help to mitigate the handicap experienced in small cultures mentioned in the introduction. Moreover, computers help to free us from the „special case suggestion“ (Páleš, 1994) and they also mean „correction of theory“ (Štícha, 2003), as well as a great inspiration.

### 3 CORPUS MORPHOSYNTAX CONCEPT (KMS)

Any explanation of the Slovak corpus morphosyntax concept must be based on a correct understanding of the terms *corpus research* and *morphosyntax*.

In the case of corpus empirical research, we proceed from facts to inference. However, the starting point for morphosyntax includes the semantic aspect.

Structure of the paper:

- Semantic starting point;
- Total frequency in the corpus; frequency according to styles; frequency of collocations according to types of documents (frequency graphs);
- descriptive-explanation and communicative-pragmatic interpretation of corpus findings.

#### 3.1 Corpus Documents Base

Unlike *The Longman Grammar of Spoken and Written English* (1999), we have no opportunity for research into conversation style today.

As the starting point for morphosyntactic corpus research, the corpus was limited to written texts from the following three style areas: journalism (ca 5 + 5 mill., subjective-objective, or central and regional); science (ca 5 + 5 mill. popular-scientific); art (ca 7 mill. + TV and radio scripts) – 30 mill. representative electronic corpus (with style and genre external annotations), as a selection of the current ca 215 mill. in the Slovak National Corpus.

The following principle holds true: The greater the care paid to the selection of the initial texts, the more objective the results to be obtained even from less numerous corpuses.

### 3.1.1 Frequency of morphosyntax means

The strength of the Slovak Corpus Morphosyntax Concept (compared, for instance, with explanatory research) is data on frequency, frequent non-symptomatic phenomena being at the centre, in accordance with natural grammar theory.

Frequency dimension in the Slovak Corpus Morphosyntax Concept;

- The most frequent being as recorded over the whole corpus ;
- Frequency according to texts (UT, PT, OT) (morphopragmatic dimension)
- Frequency of collocations (generally probably only the first from the left and the right, with others in special cases).

### 3.1.2 Interpretation of corpus findings and their descriptive-explanations, communication-pragmatic and comparative dimensions

Descriptive-explanation interpretation will be at the centre of the Slovak Corpus Morphosyntax Concept, under the supervision of the principles, i.e., in close co-operation with J. Dolník and P. Žigo. In accordance with the naturalness theory (Werner, 1989), non-symptomatic and frequent means should be in focus, as well as symptomatic non-frequent ones, or absent means, such as, for instance, active present perfect participle should be located on the periphery (as a footnote, as a lexicalised one). When interpreting insulated phenomena, their removal from the natural model must be taken into account in any case.

#### 3.1.2.1 Descriptive and explanatory interpretation

Using the results of J. Dolník's team as an explanation framework when interpreting phenomena, even in the case of new declination and conjugation paradigms.

#### 3.1.2.2 Communicative and pragmatic dimension of interpretation

I recommend limiting the communicative and pragmatic dimension to such data on KF (compare Štícha, 2003) and data on the types of documents (pragmatic morphology and pragmatic syntax) which can be actually detected. The communicative and pragmatic character will be more markedly manifested in the case of such morphosyntactic phenomena, where the pragmatic dimension is constitutive, e.g., *modus*.

Communication-pragmatic starting point (compare List of References)

1. Conditions of communication situation
2. Acts of notice and communicative functions
3. Evaluation of validity of notice
4. Current classification

I agree with J. Kořenský and F. Mika that denotation is manifested at sentence level (in the case of co-operation of S and V), the naming of the denotate being, of course, noun (as the minimal denotation).

#### 3.1.2.2 Comparative aspect

The comparative aspect in relation with other languages will apply to those languages, most frequently taught i.e., English and German, as well as in comparison with English corpus grammar and Štícha's grammar (2003).

### Paradoxes of system and parole

For instance, the language system offers 30% concrete and 70% abstract items (Puškášová). Nonetheless, we assume frequency confirmation in the corpus of real use of concrete countable nouns – the natural aspect.

The language system offers, for instance, fewer qualitative evaluating quality adjectives and more relation ones. Nonetheless, we assume frequency confirmation in the corpus of real use of quality adjectives – the natural aspect (English corpus grammar, p. 511).

The language system offers, for instance, fewer circumstance adverbs and more property adverbs from adjectives. Nonetheless, we assume frequency confirmation in the corpus of real use of primary time and spatial circumstance adverbs – the natural aspect (English corpus grammar, p. 561).

The language system offers, for instance, fewer subject verbs of movement and more subject-object verbs. Nonetheless, we assume frequency confirmation in the corpus of real use of verbs of movement – the natural aspect (English corpus grammar, pp. 385, 388).

The language system offers, for instance, fewer paratactic and more hypotactic conjunctions. Nonetheless, we assume frequency confirmation of real use in the corpus of more concrete paratactic conjunctions (*a, či, aj, alebo* – or asyndetic conjunction) for highly abstract hypotactic subjunctions (*že, keď, lebo, aby*).

### 3.2 Morphosyntax character of the project

In accordance with W. Mayerthaler (1998), I understand morphosyntax as a discipline which is based on the interaction of morphology and syntax (in principle, in Morris' conception, as word syntax).

#### $M \leftrightarrow S$ (morphosyntax)

##### 3.2.1 Morphosyntax on semantic basis

I want to present morphosyntax on a semantic basis, within the framework of cognitive linguistics (in Kořenský's concept, see his argument, 1998: text  $\rightarrow$  morphosyntax  $\rightarrow$  formal (paradigmatic) morphology  $\rightarrow$  morphematics  $\rightarrow$  morphonology  $\rightarrow$  phonology).

As a morphologist, my starting point is morphology, but I am striving for its morphosyntactic classification and of exceeding language tiers by the mutual interconnection of syntactic and morphological structures, but mainly of the communicative functions.

Paradigmatics will be taken into account secondarily (as GT frequency and variantness)

- Semantic starting point;
- Frequency in the corpus (communicative – pragmatic aspect) and frequency of collocations according to types of documents (UT, PT, OT);
- Descriptive – explanation interpretation of corpus findings;
- Communicative – pragmatic functions of morphosyntactic means.

3.2.1.2 Semantic starting points (relation of semantic and morphosyntactic units), cf. Grundzüge, pp. 20 – 112; Kořenský, 1998, Karolák, 1984, Dolník, Žigo, 2003, etc.

Semantic structure:

- Reference (relation to denotate);
- Characterisation (different level of abstraction);
- Specific types: relation to classes (*Ortuť je jedovatá.*), to closed subjects (*ak..., tak...*); negation (*nič nepriniesol, nič sa nestalo*);
- **Level of lexions** (*jeť* transitive 47%, intransitive 36%).

Construction of semantic structures:

- Proposition as a reflection of reality, each semantic structure having at least one proposition;
- Proposition elements: semantic predicate (1 – n, question: *aký je?*) and semantic argument (1 – n; questions: *kto, čo, ktorý, kde, kedy?*), time and spatial frame;
- Single-argument semantic predicates – Functors (f)(x): *zelený, blond, kašľať, pes, spať, sedieť, kameň*; two-argument semantic predicates, functors (f)(x, y): *podobný, nosiť, pred, milovať, väčší, brat*; three-argument functors (f)(x, y, z): *predať, darovať, dar*; zero-argument functors: *pršať*;
- Predicates of predicates: Dom (argument) /je /veľmi/p2 /vysoký/p3./p1;
- Semantic arguments: (reference) objects, communication object (without denotation), closed (), classes (*strom, druh, cicavec*): arguments can be agens, paciens, experient, objective, recipient, source, target, etc.
- Logical predicates: identity (*je to Peter, kto to povedal*), membership of a class (*Bratislava je mesto.*), mutual connection (*sklenička ze skla...*), causality (*zabiť – spôsobiť, zapríčiniť, vyvolať, že je mŕtvy*).
- Operators as logical elements with scope (negation, performatives), empty operators (*a, alebo*), modal operators.

### 3.2.2 Theory of functional-semantic categories

The theory of functional-semantic categories (cf. List of References Bondarko, Chrakovskij), which are focused on grammar categories and based on them, but, on the other hand, when defining the means, the functional-semantic categories exceed the morphological framework in the hierarchy, as well as the framework of word classes. The theory of functional-semantic categories helps to overcome the insulation of grammar categories and of word classes, etc.

Functional-semantic category: gender (natural gender, animateness, animalies, feminatives, masculinatives, genre pronouns)

Functional-semantic category: quantitateness – quantification (number, numerals, repetition, collectiveness, propriality) quality as verity and materiality

Functional-semantic category: relation (case, prepositions, conjunctions, ...)

Functional-semantic category: causality

Functional-semantic category: aspectuality (secondary suffix imperfectives are the aspect in the focus; secondary prefix perfectives are the aspect in the centre; subsumption – redundancy, there is limit AV on the transition, actio verbi is hierarchised as well, temporalness, localising, modality as late as at the end, suffixing before prefixing)

Functional-semantic category: modus and modality

Functional-semantic category: temporalness (tense, aspect, actio verbi, ...)

Functional-semantic category: personalness (person, personal pronouns, performatives, ...)

Functional-semantic category: passivisation (passive perspective of sentence, non-action, processuality, ...)

Functional-semantic category: congruence (means for expression of syntactic relation of determination, hyponymy and hypernymy)

Functional-semantic category: intensity (gradation ... actio verbi, diminutives, augmentatives)

Functional-semantic category: space localising



**Functional-semantic categories** word classes + word categories + grammar categories with the same function and similar meaning

KM Lexical categories Semantic	DM Derivation	GM/MM Morphological categories	Member Syntactic categories
<i>natural sex</i>	<i>feminative masculinative</i>	<i>genus</i>	<i>congruence</i>
<i>numerals</i>	<i>collectives singulatives</i>	<i>numerus</i>	<i>congruence</i>
<i>prepositions conjunctions intensity</i>	<i>elation diminutives augmentatives actio verbi</i>	<i>case gradation</i>	<i>government / valence ako, než</i>
<i>action</i>	<i>process</i>	<i>intention</i>	<i>valence</i>
<i>resultativeness</i>	<i>actio verbi</i>	<i>aspect</i>	<i>phase, limit verbs</i>
<i>action AG process PT statics particles nech modal verbs modal adverbs</i>		<i>genus verbi AG – S active voice PT – S passive voice modus rob!</i>	<i>sentence segmentation sentence modality</i>
<i>time adverbs</i>	<i>odvčera</i>	<i>tempus</i>	
<i>participants of communication ja, ty on</i>		<i>person</i>	<i>double-clause single-clause sentences</i>

### 3.2.3 Basic principle of morphosyntax

The basic principle of morphosyntax is co-operation between theta-roles and case filter (Willi Mayerthaler, 1998), between verbal valence and substantive case.

Relevant principles according to (Mayerthaler):

Theta role principle: One theta role (functor) is assigned to any argument (semantic participant).

Case filter principle: any nominal phrase must contain at least one case.

More relevant morphosyntax principles cf. Dittmann, 1994:

The principle of relative arbitrariness of morphosyntactic means in their relation with communicative functions.

The principle of more-dimensional scope of language activity (not all pragmatic-communicative factors have their explicit indicators on the surface level; they are not in 1:1 ratio).

The morphosyntactic approach means focusing, within the framework of morphosyntax, on the phenomena which follow from the interaction between morphology and syntax:

A) Super-sentential syntax is deliberately excluded from syntax. It will be included in the morphosyntax secondarily, e.g., in the case of concurrence of memvers and VV, in the case of subjunctives, etc.

B) Morphosyntactic characteristics and word classes order

Morphosyntactic characteristics as used by J. Kořenský (1998)

Basic auto-paradigmatic word classes: verb; noun; adjective; adverb

<b>Vp</b>	Vs/o	Vatr	Vadv
(Sp)	<b>Ss/o</b>	Satr	Sadv
(ADJp)	((ADJs/o))	<b>ADJatr</b>	((ADJadv))
(ADVp)	((((ADVs/o)))	ADVatr	<b>ADVadv</b>

Superstructural non-auto-paradigmatic word classes: numerals; pronouns

synsemantic non-paradigmatic word classes: prepositions; conjunctions, particles

interjections

C) verb will be in the centre of Slovak morphosyntax;

D) valence is in the centre within the framework of verbal categories;

E) case is in the centre within the framework of noun categories;

F) further classification of word classes and morphosyntactic categories will use the results of the theory of functional-semantic categories, which transcends the boundaries of language tiers and of languages and connects means with different levels of abstraction.

Information on paradigmatics will be on the margin of any word class, in their connection with dynamic trends (variants).

### 3.3 International linguistic terminology (in accordance with the era of globalization)

In order to connect with contemporary developments and trends in the era of globalization (information control systems, corpuses, Wordnet, ...) I suggest the use of international linguistics terminology in grammars as well as in textbooks of Slovak. The teaching of foreign languages in schools will be simplified in this way, as well as achieving an approximation of slovakistic works to the intercultural (European) community. I feel that the Enlightenment domestic („understandable“) terminology may act as a disincentive in the era of globalization – compare the comic suggestion of using the Slovak equivalents of indications of cases (Očenáš, 2003).

3.3.1 Definitions of relevant terms of morphematic, morphological, and syntactic structures (cf. List of References)

Besides an index, a glossary of international terms and their Slovak equivalents is appended (A. HORÁK)

This requirement raises the following challenges which must be addressed:

it is uncommon;

international terms are not adopted (*modu / modusu, verbum, neuters: tempus – tempora (temporá), genus – genera (generá), numerus – numera (numerá), but korpusy (not korporá); plurále, numerále, pronomen/pronominá*)

#### 4 CORPUS-BASED STRUCTURE OF CORPUS MORPHOLOGY OF THE SLOVAK LANGUAGE

The team of solvers is open to other interested parties, who will be logged onto the project's website on the basis of their written application.

M → S (classic approach: form → contents → function)

The Czech approach was not quite respected in MČ (M II – S III), which was recommended in Theoretical Foundations from function through contents to form (function → contents → form). Nevertheless, the approach is relevant even today, i.e., 30 years later. The direction S → M is used by F. Štícha (2003), in German grammars, as well as in *Longman Grammar of Spoken and Written English* (1999). Its Bondarko modification: function/contents → form is used in the monograph. MSJ is consequent in the context with the starting point form – contents – function (as in Russian grammar), but it does not mean a deterioration of its quality. Therefore, I agree with J. Bosák (2003) who indicated the equivalence of these procedures. To put it simply, we should choose whichever will be the most efficient for KSM.

M → S (classical approach: form → contents → function) S → M (function → contents → form)

When comparing F and O, we will find out that F are fixed, but O are universal. Therefore, the procedure which starts with O is ideal for comparison of languages (Bondarko).

Morphosyntax will be used by me mainly when creating a morphosyntax of the Slovak language as a foreign language. Compare also the theory of functional-semantic categories and its pros.

##### 4.1 *Verbum / verbá (verb)*

(compare List of References)

Solvers: Ivanová (static verbs), Kášová (modus), Nižníková (valence), Sokolová (semantic classification), Giger ? (genus verbi), Szymczak ? (verb-nominal predicates) (open to other solvers, including diploma students)

**Word-class characteristics** (using the theory of functional-semantic categories, natural grammar, and UG principles)

##### **Syntactic functions of finite and infinite verbal forms**

(Vvp, Vv-n p, Vo/s, Vatrib, Vadv, Vkomplement)

- Semantic, morphosyntactic, and pragmatic aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect UT, PT, OT);
- descriptive – explanative interpretation of corpus findings

##### 4.1.1 Semantic classification of verbs in Slovak

Solvers: Ivanová, Kášová, Sokolová... (open to other solvers)

Auto-semantic, auto-syntactic, and auto-paradigmatic verbs (action, process, static), synsemantic and synsyntagmatic, auto-paradigmatic verbs (copula and categorical operators, phase, modal, and limit modifiers).

**a) auto-semantic (predicates):**

- **action** (dynamic action (*robiť*), communicative, mental, causal, ... )
- **process** (dynamic non-action, *chudnúť*)
- **static:**
  - relation (*patriť, mať*)
  - qualification (*vyzeráť*)

**b) desemantised: MODIFIERS**

- **phase** (*začať, neprestať, končiť*)
- **limit** (*ísť, mať*)
- **modal** (*môcť, musieť, smieť, chcieť*)

**OPERATORS**

- **copula** (*byť, mať, stať sa, cítiť sa*)
- **categorical** (*dať, sedieť*)
- **performance** (*myslím*) Wurzel

**ATTITUDE/CERTITUDE MODALITY**

verbs with contents VV (*povedal, že pride, myslím, že tomu rozumie; zdá sa, že tomu rozumie*)

Research into performatives which have no truth-value, but successfulness, their allocation not depending on the context (they occur only in the first person, present tense, formula: present tense (2<sup>nd</sup> person *aby* + sentence)) 1.

attitude of author of the statement leads to proposition – they describe situation (epistemic *vedieť, myslieť*, doxastic *veriť, domnievať sa, predpokladať*; normative *musieť*, motivation *želať, priať*, intentional *chcieť, zamýšľať*, preference *uprednostniť*, evaluation (*pokladať za...*), expectative *očakávať*, parative *môcť*...functors); statement (description, reactivity))

Research of performatives *a aby, keby, žeby... by* (*Tvrdí, vie, vidím, počujem....*) *Prší*.

Relativisation: *Myslím, domnievam sa, že prší*.

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- descriptive – explanatory interpretation of corpus findings

Semantic base, semantic relations and derivations of base structures (predication; actualisation; identification; phasing; quantification; intensification; modality – volunative; attitude; emotionality; negation; sentence perspective, cf. Kořenský, 1998, Štícha, 2003).

Functional-semantic categories: causality (adverbs; conjunctions; prepositions; causatives), modality, temporalness, localness.

**4.1.2 Syntactic and semantic valences – intention (theta role principle)**

Solvers: Ivanová, Kášová, Nižníková, Sokolová... (open to other solvers).

Syntactic and pragmatic dimensions (morphosyntactically constitutive category).

**I. VT subject – object**

I.ITa Ag– D– Pt *robiť*

(transitive (Sakuz) and I.ITb Proc– D– Pt *mrzieť* intransitive)

I.ITc Stat– D– Pt *patriť*

II. **subject** (object-less)

II.IT Ag=Pt– D *ísť*

III.IT Ag– DPt *smiať sa*

IV.IT Ag/Proc– D *baničiť*

V.IT Proc– D *starnúť*

VI.IT Proc/Stat – D *belieť sa*

III. VT **object** (subject-less)

VII.IT D– Proc/Stat *smädiť*

IV. **subject-object-less** VIII.IT D *blýskať sa*

• Semantic, syntactic, and morphological aspects;

• Frequency in the corpus (selection of 30 mill. tokens);

• Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

• Descriptive – explanative interpretation of corpus findings

Table 1 (Sokolová, 1995)

Number of actants	Valence structure	Examples	Modified valence	
			Double-member structures	Single-member structures
1	Sn ← VFpers Sn/g	riadiť/ovať kričať	<i>Kričať slová.</i>	<i>Ako sa ti riadiť/uje?</i> <i>Potom sa kričí.</i>
2 1+(1)	Sn ← VFpers (ADVloci) Sn/g (ADVdir)	sediť ísť, bežať	<i>Bežala sa stovka.</i>	<i>Stálo sa tam aj sedelo.</i> <i>Išlo sa do kina.</i>
2	Sn ← VFpers → Sa/g Sn/g Sd Sg Si Sp. Oinf	piť, jesť predísť dotýkať sa pohrdať čakať (na). ísť	<i>Pilo sa víno.</i>	<i>Pije sa tam.</i> <i>Predchádza sa chorobe.</i>  <i>Pohrda sa smrťou.</i> <i>Čaká sa na smrť.</i> <i>Ide sa nakupovať.</i>
3	Sn ← VFpers → Sa → Sa Sn/g Sa Sd Sa Sg  Sa Si Sa Sp Sd Sp Si Sp Sa VV/inf Sd VV/inf	naučiť poslať ušetriť  ponúkať prebrať pomáhať s zaplatiť prinútiť prikázať	<i>Bola naučená.</i> <i>Bol poslaný jemu.</i> <i>Bol ušetrený bolesti.</i> <i>Boli ponúkané.</i> <i>Preberie sa to.</i>  <i>Bola prinútená ísť.</i> <i>Prikázalo sa im ísť.</i>	<i>Pomáha sa mu s...</i> <i>Zaplatilo sa životom.</i>
3	S ← VFp → Sa → (ADVdir) Sn/g Sa/g	odviezť položiť	<i>Boli odvážaní.</i> <i>Bola položená nabok.</i>	
4	4 S ← VFp → Sa → Sd → Sp Sn/g Sa/VV/inf Sa/g(Sd)	hovoriť napísať čítať	<i>Reči sa hovoria.</i> <i>Je napísaný.</i> <i>Knihy sa čítajú.</i>	<i>Lahko sa mu hovorilo.</i> <i>Písalo sa mu dobre.</i> <i>Číta sa nám dobre.</i>
4	Sn/g ← VF → Sa → Sp → Sp	vymeniť s, za	<i>Súčiastky sa vymenia za iné.</i>	

#### 4.1.3 Genus verbi (morphosyntactically constitutive category)

Solvers: Sokolová, Giger... (open to other solvers)

Syntactic and pragmatic dimensions (hierarchisation of statement).

Relations between semantics of verb, its valence, and transformates are indicated in the following scheme:

Basic Structure:

Static verbs		Process verbs		Action verbs	
←V		←V	29,9%	←V	
←V→	89,5%	←V→	39,5%	←V→	90,9%
0		←V→→		←V→→	
0		0		(←V→→→)	
(V→→)		←V		0	
0		←V		0	
0		(V)		0	

Modified structure

Static verbs		Process verbs		Action verbs	
0		←V		←V	
0		←V→		←V	
0		←V→→ADV		←V→→	
0		V→		V→	
(V→→→)		V→→→		V→→→	
0		(V)		(V)	

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- Descriptive – explanative interpretation of corpus findings

Functional-semantic category passivisation (passive sentence perspective, non-action, processualness), hierarchicisation.

#### 4.1.4 Congruence (numerus, genus) (morphosyntactically constitutive category)

Solvers: Ivanová, Nižníková... (open to other solvers)

Syntactic and pragmatic dimensions.

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- Descriptive – explanative interpretation of corpus findings

#### 4.1.5 Actio verbi and verbal aspect (morphosyntactically non-constitutive lexical-grammar category and communication-pragmatic category)

Solvers: Ivanová, Sokolová... (open to other solvers)

Semantic and pragmatic dimensions.

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);



- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- Descriptive – explanative interpretation of corpus findings

Functional-semantic category aspectuality (the aspect secondary suffix imperfectives are in focus, secondary prefix perfectives are in the centre, subsumption – redundant, limit AV is in transition)

Actio verbi – the following are hierarchicised as well: temporalness, localness, modality, suffixing before prefixing, functional-semantic category of intensity (gradation, actio verbi, diminutives, augmentatives).

#### 4.1.6 Modus (morphosyntactically non-constitutive grammar category)

Solvers: Kášová, Nižníková... (open to other solvers)

Pragmatic dimension (update, attitude).

Empiricism (past (reality, unreality) / non-past (reality, unreality) / non empiricism (appeal / non-appeal).

- Semantic, syntactic, and morphological aspects;

- Frequency in the corpus (selection of 30 mill. tokens);

- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- Descriptive – explanative interpretation of corpus findings

Functional-semantic category modus / modality (modus modal verbs, modal predicatives, particles, interjections); morphological modus and syntactic modality.

Morphological modus	Syntactic modality
Indicative	Declarative sentences Interrogative sentences Optative sentences Imperative sentences
Imperative	Imperative sentences
Conditional mood	Optative sentences Declarative sentences Interrogative sentences

#### 4.1.7 Tempus (morphosyntactically non-constitutive grammar category)

Solvers: Ivanová... (open to other solvers)

Pragmatic dimension (update).

Time of speech (past / not past)

- Semantic, syntactic, and morphological aspects;

- Frequency in the corpus (selection of 30 mill. tokens);

- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- Descriptive – explanative interpretation of corpus findings

Functional-semantic category temporalness (tense, aspect, actio verbi, adverbs, ...).

#### 4.1.8 Persona (morphosyntactically non-constitutive grammar category)

Solvers: Papierz... (open to other solvers)

Pragmatic dimension (update, identification).

Functional-semantic category personalness (person, personal pronouns, performatives)

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);
- Descriptive – explanative interpretation of corpus findings

#### 4.1.9 Form verbal structure (conjugation, declination)

Solvers: Sokolová, students of FF PU 124523 (open to other solvers)

(morphosyntactically non-constitutive area)

Verbs with full and defect paradigms.

Infinitive verbal forms (morphological and syntactic use).

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);
- descriptive – explanative interpretation of corpus findings

#### 4.2 Noun (cf. List of References)

Solvers: Benko, Hašanová, Gianitsová, Ološtiak... (open to other solvers)

Word-class characteristics (using the theory of functional-semantic categories, natural grammar, and UG principles)

Ss, Sadj, Sadv, Sv

Syntactic functions

(Sp, Sv– n p, So/s, Satrib, Sadv, Skomplement)

##### 4.2.1 Semantic classification

Solvers: Ološtiak, Sokolová... (open to other solvers)

**Appelatives – propria (group antroponyms as improper propria); concretes (singulative – materialia, collectiva), abstracts (improper abstracts)**

##### 4.2.2 Case / cases (morphosyntactically constitutive category)

Solvers: Gianitsová, Ološtiak, Sokolová...

Morphosyntactic characteristics

S	Ss	Intention cases S, O	Ps		Ns
Adj	Sadj	adnominal G	Padj		Nadj
Adv	Sadv	Circumstantive L, pP	Padv	Pnum	Nadv
V .....	Sv	Contentual N, I	to je on		I+I = 2

Syntactic dimension of substantive GT (Sokolová, 1995):

Case	Characteristics	Valence position	Syntactic functions
N	Does not indicate participation in action Active participant	<b>left-valence</b> (right-valence, non-valence, adnominal)	<b>Basic form: Subject</b> Verbal-nominal predicate, nominal phrase, adverbials, (non)congruent attribute ?? complement
A	Unlimited effect Non-active participant	<b>right-valence</b> non-valence, adnominal	<b>Basic form: Direct object</b> (Adverbials, non-congruent attribute)

D	Marginal participation in action Non-active target participant	<b>right-valence</b> non-valence, adnominal	<b>Basic form: Indirect object</b> (Adverbials, non-congruent attribute)
G.	Limits the scope of participation (non)active participant	<b>adnominal</b> right-valence non-valence	<b>non-congruent attribute</b> (object, subject, adverbials)
I	Temporary participation in action (non)active participant	<b>non-valence</b> (right-valence, adnominal)	<b>Adverbials</b> verbal-nominal predicate, object, non-congruent attribute)
L	Absence in the intention structure Non-active participant	<b>non-valence</b> (right-valence, adnominal)	<b>Adverbials</b> (object, non-congruent attribute)

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (*na Slovensko* 6 000 – *na Slovensku* 66 000);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- descriptive – explanative interpretation of corpus findings

Case filter principle: any nominal phrase must contain at least one case. Case hierarchy N (independent), others being dependent (accusative – the most), others in limited scope, dative – marginally, others not peripherally, genitive (limits the scope), instrumental (), local, and other preposition cases. Function-semantic category relation (case, prepositions, subjunctions)

#### 4.2.3 Numeral(s) (morphosyntactically non-constitutive grammar category)

Solvers: Benko, Hašanová, Gianitsová, Ološtiak... (open to other solvers)

Singular – plural, singulare tantum, plurale tantum

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);
- descriptive – explanative interpretation of corpus findings

Functional-semantic category quantitateness / quantification (quantification operators of proposition semantics: Number, numerals, repetitiveness, collectiveness, propriateness)

#### 4.2.4 Genus / Genera (morphosyntactically non-constitutive grammar category)

Natural grammar gender. Gender and feminatives/masculinatives/animalies. Gender studies (gender linguistics).

Solvers: Benko, Hašanová, Gianitsová, Ološtiak, Sokolová... (open to other solvers)

Syntactic, semantic, and pragmatic dimensions.

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);
- Descriptive – explanative interpretation of corpus findings

Functional-semantic category genre (natural genre, animateness, animalies, feminatives, masculinatives, genre pronouns)

#### 4.2.5 Form structure (declination)

Solvers: Benko, Hašanová, Gianitsová, Ološtiak, Sokolová... (open to other solvers)

Auto-paradigmatic word class. New noun declination system.

#### 4.3 Adjective (cf. List of References)

Solvers: Benko, Hašanová...

Word-class characteristics (using the theory of functional-semantic categories, of natural grammar, and UG principles)

ADJadj, (ADJs??), ADJadv, ADJv (ADJatrib, ADJkomplement, ADJadv, ADJv–nom. p)

##### 4.3.1 Classification

Solvers: Benko, Hašanová... (open to other solvers)

Semantic features: qualitateness, relation, appreciativity, intensity.

##### 4.3.2 Congruence of adjectives with superior noun in categories genre, number, and case (morphosyntactically constitutive category)

Solvers: Benko, Hašanová... (open to other solvers)

Syntactic dimension (attributiveness)

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- Descriptive – explanative interpretation of corpus findings

Functional-semantic category congruence (a tool of expression of syntactic relation of determination, subordinality, superordinality, case, number, genre)

##### 4.3.3 Gradation (morphosyntactically non-constitutive lexical-grammar category)

Solvers: Benko, Hašanová... (open to other solvers)

Semantic feature of intensity (increasing and decreasing).

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);

- Descriptive – explanative interpretation of corpus findings

Functional-semantic category intensity (gradation, actio verbi, diminutives, augmentatives, comparison)

##### 4.3.4 Valence (morphosyntactically constitutive category)

Solvers: Benko, Hašanová... (open to other solvers)

##### 4.3.5 Form structure (declination, gradation)

Solvers: Benko, Hašanová... (open to other solvers)

Auto-paradigmatic adjectives. Non-auto-paradigmatic adjectivals.

#### 4.4 Adverb (cf. List of References)

Solvers: Šimková... (open to other solvers)

Word-class characteristics (using the theory of functional-semantic categories, natural grammar, and UG principles)

ADVadv, ADJv– nom. p, ADJatrib

#### 4.4.1 Classification

Solvers: Šimková, Horák... (open to other solvers)

Functional-semantic category temporalness (temporal adverbs, tempus, verbal aspect, actio verbi)

Functional-semantic category space localising (space adverbs, actio verbi, determining pronouns)

Functional-semantic category modus/modality (modus, modal verbs, predicatives, modal particles, and interjections)

Functional-semantic category causality (adverbs, conjunctions, prepositions, causatives)

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);
- descriptive – explanative interpretation of corpus findings

#### Form structure

Non-paradigmatic inflective word class.

#### 4.5 Numerales / numerals (cf. List of References)

Solvers: Garabík, Gianitsová, Horák... (open to other solvers)

Word-class characteristics (using of theory of functional-semantic categories, natural grammar, and principles)

Quantifiers (numerals, numerus)

NUMs/o, NUMatrib, Num adv

#### 4.5.1 Classification

Solvers: Gianitsová, Sokolová... (open to other solvers)

Functional-semantic category quantitativeness – quantification (number, numerals, repetitiveness, collectiveness, propriality)

Congruence Nadj.

- Semantic, syntactic, and morphological aspects;
- Frequency in the corpus (selection of 30 mill. tokens);
- Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);
- Descriptive – explanative interpretation of corpus findings

#### 4.5.2 Form structure (declination, trend to inflectivity)

Solvers: Garabík, Sokolová, students 124523 (open to other solvers)

Non-auto-paradigmatic Ns, Nadj, Nadv

#### 4.6 Pronoun(s) (cf. List of References)

Solvers: Papierz, Šimková, Hašanová... (open to other solvers)

Word-class characteristics (superstructure word class)

PRONs, PRONadj, PRONadv, PRONnum (PRONs/o, PRONatrib, PRONadv, PRONv-n p)

#### Classification

Solvers: Papierz, Šimková... (open to other solvers)

Syntactic, semantic, and pragmatic dimensions.

- Semantic, syntactic, and morphological aspects;
  - Frequency in the corpus (selection of 30 mill. tokens);
  - Frequency of collocations according to types of documents (communicative – pragmatic aspect: UT, PT, OT);
  - Descriptive – explanative interpretation of corpus findings
- Functional-semantic category deixis (gender, number, case, congruence)

**Form structure** (according to the parallel word class the pronouns refer to)

Solvers: Šimková, Hašanová, students... (open to other solvers)

Non-auto-paradigmatic PRONs, PRONadj, PRONadv

#### 4.7 Prepositions (cf. List of References)

Solvers: Horák, Šimková... (open to other solvers)

Word-class characteristics

Classification

Primary and secondary prepositions.

Local and temporal relations.

Prepositions and prefixes.

Functional-semantic category relation (case, preposition operators, conjunctions)

Aparadigmatic.

#### 4.8 Conjunction (cf. List of References)

Solvers: Šimková... (open to other solvers)

Word-class characteristics

##### Classification

Conjunctions and subjunctions as operators. Sentential and super-sentential contexts.

Function-semantic category relation (case, prepositions, conjunctions)

Aparadigmatic

#### 4.9 Particle (cf. List of References)

Solvers: Šimková... (open to other solvers)

Word-class characteristics

##### Classification

Particles with communicative function and syntactic function.

Particles transcending boundaries of members – *nie, áno, možno* – sentences and particles which modify members.

+/- sentential equivalent (operators, commentators)

Functional-semantic category modus / modality (modus modal verbs, predicatives, particles, interjections)

Functional-semantic category quality as truth and materiality

Functional-semantic category evaluation (in relation to members – *to je pekný kabát*, in relation to addressee – *dobré, že*)

Aparadigmatic



#### **4.10 Interjection** (cf. List of References)

Solvers: Orwińska-Ruziczka, Šimková... (open to other solvers)

Word-class characteristics (reactors, sentential equivalents)

Amorphous features which express relation to author, addressee, and to the object of communication.

Syntactic dimension.

#### **Classification**

Onomatopoeia, volitive and emotional interjections.

Functional-semantic category emotionality (diminutives, vulgarisms, imperatives, particles, interjections)

Aparadigmatic

### **5 CONCLUSIONS**

#### **Corpus-based Slovak morphology is a part of corpus morphosyntax**

Corpus document base

Morphosyntactic character

International linguistic terminology

Corpus grammars are based on existing theories of grammar and real non-fabricated texts. While exact frequency expression of use of morphosyntactic phenomena can be used also in prescription (at least as an orientational indicator), this will not be the primary task of Slovak corpus morphosyntax. The morphosyntactic approach to corpus researches follows from the situation in slovakistics, too. It is also convenient due to the fact that the existing works from research in morphology are rather uncritically accepted in slovakistics, but discussion is continuing on the character of the Slovak syntax. The corpus morphosyntax does not come with a new theory; rather, it is oriented towards the précising and tuning of the existing grammar research works in the Slovak language and their verification on the basis of the corpus materials. Nevertheless, it will use the results of theoretical research works of Slovakists, with whom our authorial team has to co-operate closely (and, indeed, so wishes). Such co-operation, based as it is, on the mutual exchange of information and on constructive professional discussions, should be useful for all the participating parties. Moreover, the Slovak corpus morphosyntax can have an application in the context of drafting a new university grammar (which is really essential) and when creating a grammar of the Slovak language as a foreign language, which is in particular demand abroad. Corpus research has great potential. Nevertheless, its final effect depends on extraneous factors, such as the quality of computer equipment in the participating workplaces, and the preparedness and ability of Slovakists to use that potential.

## BIBLIOGRAPHY

1. BĚLIČOVÁ, H.: Nástin porovnávací morfologie spisovných jazyků. (Outline of Comparative Morphology of Standard Languages.) Praha, Karolinum, Nakladatelství Univerzity Karlovy 1998. 217 p.
2. BENKO, V. – HAŠANOVÁ, J. – KOSTOLANSKÝ, E.: Počítačové spracovanie jazyka. Časť: Morfológia podstatných mien. (Computer Language Processing. Part: Morphology of Nouns.) Bratislava, Pedagogická fakulta Univerzity Komenského 1998. 79 s.
3. BONDARKO, A. V.: Grammaticeskije kategorii i kontekst. (Grammar Categories and Context.) Moskva, 1971. 116 p.
4. BONDARKO, A. V.: Klassifikacija morfologičeskich kategorij. (Classification of Morphological Categories.) In: Tipologija grammaticeskich kategorij. (Typology of Morphological Categories.) Moskva, 1975, p. 56 – 76.
5. BONDARKO, A. V.: Teorija morfologičeskich kategorij. (Theory of Morphological Categories.) Leningrad, Nauka 1976, pp. 14 – 25, 41 – 129, 223 – 245.
6. BOSÁK, J. – BUZÁSSYOVÁ, K.: Východiská morfémovej analýzy. (Bases of Morpheme Analysis.) (Morfematika. Slovtvorba) In: Jazykovedné štúdie. 19. Bratislava, Veda 1985. 131 p.
7. BRINKMANN, H.: Die Wortarten im Deutschen. (Word classes in German.) In: Das Ringen um eine neue deutsche Grammatik. (The Struggle for the New German Grammar.) Darmstadt 1962, p. 118 – 122.
8. BUZÁSSYOVÁ, K.: Kategória určenosti a zhoda spony vo vetách s menným prísudkom. (Category of Definiteness and Agreement of copula in Sentences with Nominal Predicate.) Jazykovedné štúdie, 13, Ružičkov zborník. Bratislava, Veda 1977, p. 61 – 72.
9. Col.: Longman Grammar of Spoken and Written English. Pearson Education Limited 1999.
10. ČECHOVÁ, M. et al.: Čeština – řeč a jazyk. (Czech – Speech and Language.) 2. Re-edition Praha, ISV nakladatelství 2000. 407 p.
11. DANEŠ, F.: Pokus o strukturní analýzu slovesných významů. (Attempt at Structural Analysis of Verbal Meanings.) In: Slovo a slovesnost, 32, 1971, pp. 193 – 207.
12. DANEŠ, F. – HLAVSA, Z.: Větné vzorce v češtině. (Sentential Patterns in Czech.) Praha, Academia 1981.
13. DANEŠ, F.: Věta a text. (Sentence and Text.) Praha, Academia 1985. 236 p.
14. DITTMANN, J.: Konstitutionsprobleme und Prinzipien einer kommunikativen Grammatik. (Constitution Problems and Principles of Communicative Grammar.) Berlín, 1994. 90 p.
15. DOLNÍK, J.: Vývin morfológie súčasnej spisovnej slovenčiny. (Development of Morphology of Contemporary Standard Slovak.) In: SAS, Bratislava, Stimul 2000. pp. 277 – 287.
16. DOLNÍK, J. et al.: Princípy stavby, vývinu a fungovania slovenčiny. (The Principles of Construction, Development, and Function of Slovak.) Bratislava, Stimul 1999. pp. 7 – 21
17. DOLNÍK, J. et al.: Princípy jazyka. (The Principles of Language.) Bratislava, Stimul 2003. 137 p.
18. DVONČ, L.: Dynamika slovenskej morfológie. (Dynamics of Slovak Morphology.) Bratislava, VEDA 1984, 124 p.
19. Encyklopédia jazykovedy. (Encyclopaedia of Linguistics.) Editor J. Mistrík, 1<sup>st</sup> Edition. Bratislava, Obzor 1993 (headwords from morphology). 513 p.
20. Encyklopedický slovník češtiny. (Encyclopaedic Dictionary of Czech.) Praha, Nakladatelství Lidové noviny 2002, p. 377.
21. FASSKE, H.: Grammatik der obersorbischen Schriftsprache der Gegenwart. (Grammar of Contemporary Upper Serbian Standard Language.) Morphologie. Bautzen, Domowina – Verlag 1981. 881 p.
22. GREPL, M. – KARLÍK, P.: Skladba spisovné češtiny. (Syntax of Standard Czech.) Praha, Státní pedagogické nakladatelství 1986. 474 p.
23. GRZEGORCZYKOWA, R. – LASKOWSKI, R. – WRÓBEL, H.: Gramatyka współczesnego języka polskiego. Morfologia. Warszawa, Państwowe wydawnictwo naukowe 1984, 556 p.
24. Grundzüge einer deutschen Grammatik. (Main Features of German Grammar.) Editor K. E. Heidolph et al. Berlin, Akademie – Verlag 1981. 1028 p.

25. HAJIČ J. (1998): Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Hajičová, E. (ed.): *Issues of Valency and Meaning*, Praha: Karolinum, pp. 106 – 132.
26. HELBIG, G. – BUSCHA, J.: *Deutsche Grammatik. (German Grammar.)* Leipzig 1979. 629 p.
27. HORÁK, G.: *Slovesné kategórie osoby, času, spôsobu a ich využitie. (Verbal Categories of Person, Tense, and Mode, and their Use.)* Bratislava, Veda 1993. 174 p.
28. HORECKÝ, J. – BUZÁSSYOVÁ, K. – BOŠÁK, J. et al.: *Dynamika slovnej zásoby súčasnej slovenčiny. (Dynamics of Vocabulary of Contemporary Slovak.)* Bratislava, Veda 1989. 430 p.
29. <http://korpus.juls.savba.sk>
30. ISAČENKO, A. V.: *Grammaticeskij stroj russkogo jazyka v sopostavlenii s slovackim II. (Grammar Structure of Russian in Comparison with Slovak II.)* Bratislava 1960, pp. 345 – 406, 540 – 570.
31. KAČALA, J.: *Sloveso a sémantická štruktúra vety. (Verb and Semantic Structure of Sentence.)* Bratislava, Veda 1989. 250 p.
32. KAČALA, J.: *Syntakticko-sémantický výklad viet typu je hmla. (Syntactic and Semantic Interpretation of Sentences of the type "je hmla".)* In: *Jazykovedný časopis*, 41, 1990, pp. 3 – 14.
33. KAČALA, J.: *Kategoriálne slová v slovných spojeniach. Príspevok k teórii jazykového významu. (Categorical Words in Idioms: A Contribution to Theory of Linguistic Meaning.)* In: *Jazykovedný časopis*, 44, 1993, pp. 14 – 24.
34. KAČALA, J.: *Syntaktický systém jazyka. (Syntactic System of Language.)* Pezinok, Formát 1998. 144 p.
35. KAROLAK, et. al.: *Gramatyka współczesnego języka polskiego. Składnia. Warszawa, Państwowe wydawnictwo naukowe* 1984. 397 p.
36. KÁŠOVÁ, M.: *Komunikačné funkcie nemeckého konjunktívu a ich vyjadrenie v slovenčine. (Communication Functions of German Conjunctive and their Expression in Slovak.)* (Theses.) Prešovská univerzita v Prešove FHPV 2003. 163 p.
37. KESSELOVÁ, J.: *Morfológia v komunikácii. (Morphology in Communication.)* Doctoral thesis, Prešov, Prešovská univerzita v Prešove, FHPV 2002, p. 181.
38. KESSELOVÁ, J.: *Lingvistické štúdie o komunikácii detí. (Linguistic Studies in Children's Communication.)* Prešov. Náuka 2001. 104 p.
39. KOŘENSKÝ, J.: *Konstrukce gramatiky ze sémantické báze. (Constructing Grammar from a Semantic Basis.)* Praha, Academia 1984. 164 p.
40. KOŘENSKÝ, J.: *Proměny myšlení o řeči. (Metamorphoses of Thinking on Language.)* Praha, FF UK 1998. 310 p.
41. KOSTOLANSKÝ, E.: *Formálny opis syntaxe slovenského jazyka. In: Tradícia a perspektívy gramatického výskumu na Slovensku. Edit. M. Šimková. Bratislava, Veda 2003, s. 162 – 173.*
42. LOPATKOVÁ, M. – ŽABOKRTSKÝ, Z. a kol.: *Tektogramaticky anotovaný valenční slovník českých sloves. Praha, Universitas Carolina Pragensis* 2002, 99 p.
43. MAYERHALER, W.: *Natürliche Morphologie. (Natural Morphology.)* Frankfurt. Athenaion 1981. 129 p.
44. MAYERHALER, W. – FLIEDL, G. – WINKLER, CH.: *Lexikon der Natürlichkeits- theoretischen Syntax und Morphosyntax. (Lexicon of Natural-Theoretic Syntax and Morphosyntax.)* Tübingen, Stauffenburg Verlag 1998. 408 p.
45. MIECZKOWSKA H.: *Kategoria gramatyczna liczebników w ujęciu konfrontatywnym polsko-słowackim. Rozprawy habilitacyjne* 267. Kraków. Uniwersytet Jagielloński 1994. 146 s.
46. *Mluvnice češtiny I. (Czech Grammar I.)* Editor J. Petr. Praha, Academia 1986. 539 p.
47. *Mluvnice češtiny II. Tvarosloví. (Czech Grammar II. Morphology.)* Prague, Academia 1986. 536 p.
48. *Mluvnice češtiny III. Skladba. (Czech Grammar III. Syntax.)*, Prague, Academia 1987. 738 p.
49. *Morfológia slovenského jazyka. (Slovak Language Morphology.)* Editor J. Ružička. Bratislava, Vydavateľstvo SAV 1966. 896 p.
50. NÁBĚLKOVÁ, M.: *Vztahové adjektiva v slovenčine. (Relational Adjectives in Slovak.)* Bratislava, Veda 1993. 208 p.

51. Najnowsze dzieje języków słowiańskich. Editor. J. Bosák. Opole 1998.
52. NEMCOVÁ, E.: Sémantická analýza verb dicendi. (Semantic analysis of verba dicendi.) Bratislava, Veda 1990. 133 p.
53. NIŽNÍKOVÁ, J. – SOKOLOVÁ, M.: Valenčný slovník slovenských slovies. (Valence Dictionary of Slovak Verbs.) Prešov, Slovacontact 1998. 290 p.
54. OČENÁŠ, I.: Fonetika so základmi fonológie a morfológie slovenského jazyka. (Phonetics with Bases of Phonology and Morphology of Slovak Language.) Banská Bystrica, UMB 2003. 124 p.
55. ONDREJOVIČ, S.: Medzi slovesom a vetou. Problémy slovesnej konverzie. (Between Verb and Sentence: Problems of Verb Conversion.) Bratislava, Veda 1989. 123 p.
56. ONDRUS, P.: Kapitoly zo slovenskej morfológie. (Chapters from Slovak Morphology.) 1. 1<sup>st</sup> Edition Bratislava, SPN 1978, 219 p.
57. ORAVEC, J. – BAJZÍKOVÁ, E. – FURDÍK, J.: Súčasný slovenský spisovný jazyk. Morfológia. (Contemporary Slovak Standard Language. Morphology.) Bratislava Slovenské pedagogické nakladateľstvo 1984. 227 p.
58. ORAVEC, J. – BAJZÍKOVÁ, E.: Súčasný slovenský jazyk. Syntax. (Contemporary Slovak Language: Syntax.) Bratislava, Slovenské pedagogické nakladateľstvo 1982. 261 p.
59. ORAVEC, J.: Väzba slovies v slovenčine. (Relation of Verbs in Slovak.) Bratislava, Vydavateľstvo SAV 1967. 392 p.
60. ORWIŃSKI-RUZICZKA, E.: Funkcje językowe interjekcji w świetle materiału słowackiego i polskiego. Kraków 1992. 145 p.
61. PÁLEŠ, E.: Sapfo. Parafrázovač slovenčiny. (Sapfo: Paraphraser of Slovak Language.) Bratislava, Veda 1994. 305 p.
62. PANEVOVÁ, J.: Formy a funkce ve stavbě české věty. (Forms and Functions in the Construction of Czech Sentences.) Praha 1980. 222 p.
63. PAULINY, E.: Slovenská gramatika. (Slovak Grammar). Bratislava, Slovenské pedagogické nakladateľstvo 1981. 323 p.
64. POLAŃSKI, R. et al.: Słownik syntaktyczno – generatywny czasowników polskich. I. – V. Polska Akademia Nauk. Instytut Języka Polskiego. Wrocław, Kraków 1984 – 1992.
65. Příruční mluvnice češtiny. (Manual of Czech Grammar.) Praha, Nakladatelství Lidové noviny 1996. p.
66. Russkaja gramatika I. (Russian Grammar I.) Moskva, Akademia nauk SSSR 1980. 783 p.
67. RUŽIČKOVÁ, E.: Slovesá pohybu v slovenčine a angličtine. (Movement Verbs in Slovak and English.) Bratislava, Veda 1982. p. 244.
68. Retrospektívne a perspektívne pohľady na jazykovú komunikáciu. (Retrospective and Perspective Views on Language Communication.) Banská Bystrica, Univerzita M. Bela 1999, p. 190 – 198.
69. SABOL, J.: Syntetická fonologická teória. (Synthetic Phonological Theory). Bratislava, Jazykovedný ústav Ľudovíta Štúra SAV 1989. 253 p.
70. SEKANINOVÁ, E.: Sémantická analýza predponového slovesa v ruštine a slovenčine. (Semantic Analysis of Prefix Verb in Russian and Slovak.) Bratislava, Veda 1980. 199 p.
71. Semantika i sintaksis konstrukcij s predikatnymi aktantami. Materialy vsesojuznoj konferencii Tipologičeskije metody v sintaksise raznosistemnyh jazykov (14. – 16. 4. 1981). Leningrad, Akademija Nauk SSSR 1981. 108 p.
72. SGALL, P. – HAJIČOVÁ, E. – PANEVOVÁ, J. (2000): Manuál pro tektogramatické značkování, Tech. Report 7, UFAL/MFF.
73. SCHUMACHER, H.: Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben. (Verbs in Roles: Valence Dictionary to Syntax and Semantics of German Verbs.) Berlin. New York, Walter de Gruyter 1986. 830 p.
74. SLANČOVÁ, D.: Reč authority a lásky. Reč učitelky materskej školy orientovaná na dieťa – opis registra. (The Language of Authority and Love. Child-Oriented Language of Teachers in Kindergartens – Description of Register.) Prešov. Filozofická fakulta Prešovskej univerzity 1998. 224 p.

75. Slovenčina a čeština v počítačovom spracovaní. (Slovak and Czech Languages in Computer Processing.) Collection of reports from the workshop (Bratislava 26 – 27 October 2001). Editor: A. Jarošová. Bratislava Veda 2001.
76. Sociolingvistické aspekty výskumu súčasnej slovenčiny. Sociolinguistika Slovaca 1. Zost. S. Ondrejovič – M. Šimková. Bratislava, Veda 1995, 240 s.
77. Sociolinguistika a areálová lingvistika. Sociolinguistika Slovaca 2. Zost. S. Ondrejovič. Bratislava, Veda 1996, 170 s.
78. Slovenčina na konci 20. storočia, jej normy a perspektívy. Sociolinguistika Slovaca 3. Zost. S. Ondrejovič. Bratislava, Veda 1997, 352 s.
79. Slovenčina v kontaktoch a konfliktoch s inými jazykmi. Sociolinguistika Slovaca 4. Zost. S. Ondrejovič. Bratislava, Veda 1999, 200 s.
80. Mesto a jeho jazyk. Sociolinguistika Slovaca 5. Zost. S. Ondrejovič. Bratislava, Veda 2000, 310 s.
81. SOKOLOVÁ, J.: Sémantika kvalifikačných adjektív. (Semantics of Qualification Adjectives.) Nitra, FF UKF 2004. 100 p.
82. SOKOLOVÁ, M.: Sémantika slovesa a slovesný rod. (Semantics of Verbs and Verb Voice.) Bratislava, Veda 1993. 110 p.
83. SOKOLOVÁ, M.: Kapitoly zo slovenskej morfológie. (Short Chapters from Slovak Morphology.) Prešov, Š. Franko, Slovacontact 1995. 180 p.
84. SOKOLOVÁ, M. – MOŠKO, G. – ŠIMON, F. – BENKO, V.: Morfematický slovník slovenčiny. (Morphothematic Dictionary of Slovak.) Prešov, Náuka 1999.
85. SVOZILOVÁ, N. – PROUZOVÁ, H. – JIRSOVÁ, A.: Slovesa pro praxi. (Verbs for Practice.) Valenční slovník nejčastějších českých sloves. (Valence Dictionary of the Most Frequent Czech Verbs.) Praha, Akademie věd České republiky 1997. 360 p.
86. ŠIKRA, J.: Sémantika slovenských prísloviak. (Semantics of Slovak Adverbs.) Bratislava, Veda 1991.
87. ŠIMKOVÁ, M.: Možnosti využitia SNK na štúdium slovenského jazyka. In: Studia Academica Slovaca. 33. Red. J. Mlacek, Bratislava, Stimul – Centrum informatiky a vzdelávania FF UK 2004. s. 204 – 217.
88. ŠTÍCHA, F.: Utvárení a hierarchizace struktury větného znaku. (Creating and Hierarchicising the Structure of Sentential Feature.) Praha, Univerzita Karlova 1984. 132 s.
89. ŠTÍCHA, F.: Současný český jazyk. Význam a konkurence v syntaxi. (Contemporary Czech Language: Meaning and Competition in Syntax.) Praha, Univerzita Karlova 1989. 106 s.
90. ŠTÍCHA, F.: Česko-německá srovnávací gramatika. (Czech – German Comparative Grammar.) Praha, Argo 2003. 842 p.
91. Teoretické základy synchronní mluvnice spisovné češtiny. (Theoretical Principles of Synchronous Czech Grammar.) In: Slovo a slovesnost, 1975, pp. 18 – 46.
92. Tradícia a perspektívy gramatického výskumu na Slovensku. (Tradition and Perspectives of Grammar Research in Slovakia.) Bratislava, Veda 2003. 243 p.
93. WAGNER, C.: Pragmatik der deutschen Sprache. Frankfurt a. M. 2001. 495 s.
94. WERNER, O.: Sprachökonomie und Natürlichkeit. Im Bereich der Morphologie. (Speech Economy and Naturalness: In the area of Morphology.) In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 42, 1989, 1, pp. 34 – 47.
95. WÜRZEL W. U.: Flexionmorphologie und Natürlichkeit. (Flexion-Morphology and Naturalness.) Berlin, Akademie Verlag 1984. 324 p.
96. WOLIŃSKI, M. – PRZEPIÓRKOWSKI, A.: Projekt anotacji morfosyntaktycznej korpusu języka polskiego. Prace IPI PAN 938. Instytut Podstaw Informatyki PAN 2001. <<http://dach.ipipan.waw.pl/~adamp/Papers/2001 - tagset/ipi938.pdf>>

## ABSTRAKT

Korpusové gramatiky vychádzajú z existujúcich gramatických teórií a z reálnych nevykonštruovaných textov. Hoci presné frekvenčné vyjadrenie používania morfosyntaktických javov sa môže využívať aj v preskripcii (príjajmenšom ako orientačný ukazovateľ), nebude to primárna úloha slovenskej korpusovej morfosyntaxe. Morfosyntaktický prístup ku korpusovým výskumom vyplýva tiež zo situácie v slovakistike; vyhovuje aj preto, že v slovakistike sa pomerne jednoznačne akceptujú doterajšie výskumy z morfológie, ale diskusie sa vedú o charaktere slovenskej syntaxe. Korpusová morfosyntax, ktorej cieľom je spresnenie a doladenie doterajších gramatických výskumov v slovenčine a ich verifikácia korpusovým materiálom, neprichádza s novou teóriou, preto bude využívať výsledky teoretických výskumov slovakistov, s ktorými autorský kolektív musí (a chce) úzko spolupracovať. Nazdávam sa, že takáto spolupráca založená na vzájomnej výmene informácií aj na konštruktívnej odbornej diskusii bude na úžitok všetkých zúčastnených strán. Slovenská korpusová morfosyntax sa bude dať využiť aj pri koncipovaní veľmi potrebnej novej vysokoškolskej gramatiky a pri tvorbe gramatiky slovenčiny ako cudzieho jazyka požadovanej predovšetkým v zahraničí. Korpusový výskum ponúka veľké možnosti, jeho konečný efekt však závisí od kvality úrovne počítačového vybavenia pracovnísk, ale aj od pripravenosti a schopnosti slovakistov tento potenciál využiť.



**ZOZNAM PREDNÁŠOK A WORKSHOPOV V ODDELENÍ SLOVENSKEHO NÁRODNÉHO  
KORPUSU JÚTŠ SAV BRATISLAVA  
LIST OF LECTURES AND WORKSHOPS HELD AT THE DEPARTMENT OF THE SLOVAK  
NATIONAL CORPUS JÚTŠ SAV BRATISLAVA**

4. 11. 2002

BENKO, VLADIMÍR: Vyhľadávacie a konkordančné programy (1) [Retrieval and Concordance Programs (1)]

18. 11. 2002

BENKO, VLADIMÍR: Vyhľadávacie a konkordančné programy (2) [Retrieval and Concordance Programs (2)]

25. 11. 2002

HORÁK, ALEXANDER: Prehľad korpusov slovanských jazykov [An Outline of Corpora of Slavonic Languages]

9. – 10. 12. 2002

HAJIČ, JAN – UREŠOVÁ, ZDENKA: Tri úrovne anotácie v Pražskom závislostnom korpuse a využitie anotovaného korpusu pri tvorbe inteligentných lingvistických nástrojov. Workshop [Three Levels of Annotation within the Prague Dependency Treebank and the Application of Annotated Corpus in Generation of Intelligent Linguistic Tools. Workshop]

16. 12. 2002

GARABÍK, RADOVAN: Jemný úvod do programovacieho jazyka python [Smooth Introduction into Python Programming Language]

13. – 14. 1. 2003

ČERMÁK, FRANTIŠEK: Všeobecne o korpusoch [On Corpora in General]

SCHMIEDTOVÁ, VIERA: Korpus a slovníky [Corpora and Dictionaries]

KŘEN, MICHAL: Příprava frekvenčního slovníka [Preparation of the Frequency Dictionary]

ČERMÁK, FRANTIŠEK: Hovorené korpusy [Spoken Corpora]

27. 1. 2003

BELICA, CYRIL: Korpusové technológie v Ústave nemeckého jazyka [Corpus Technologies at the Institute of German Language]

30. 1. 2003

GARABÍK, RADOVAN: Reprezentácia textových údajov v elektronickej forme a implikácie z toho vyplývajúce (1) [Representation of Textual Data in Electronical Form and its Implications (1)]

3. 2. 2003

GARABÍK, RADOVAN: Reprezentácia textových údajov v elektronickej forme a implikácie z toho vyplývajúce (2) [Representation of Textual Data in Electronical Form and its Implications (2)]

10. 2. 2003

PETKEVIČ, VLADIMÍR: Lingvisticky založené morfológické značkovanie jazykových korpusov [Linguistic-Based Morphological Annotation of Language Corpora]

24. 3. 2003

PALA, KAREL – SEDLÁČEK, RADEK: Morfológické značkovanie (českých) korpusových textov (nástroje Ajka a I\_Par) [Morphological Annotation of (Czech) Corpus Texts (Tools Ajka and I\_Par)]

31. 3. 2003

GARABÍK, RADOVAN: USENET News alebo menej známy spôsob ako prostredníctvom internetu zabíjať čas [USENET News or Less Known Way How to Kill Time by Means of Internet]

28. 4. 2003

OLIVA, KAREL: Teoretické základy morfolologickej dezambiguácie korpusu lingvistickými metódami [Theoretical Basis of Morphological Disambiguation of Corpus by Means of Linguistic methods]

15. 5. 2003

PALA, KAREL – MRÁKOVÁ, EVA: Môžu byť slovníkové definície konzistentné? [Can Dictionary Definitions Be Consistent?]

26. 5. 2003

PETKEVIČ, VLADIMÍR: Jazyky na explicitné značkovanie textov – SGML a XML [Languages for the Explicit Annotation of Texts – SGML and XML]

9. 6. 2003

SKOUMALOVÁ, HANA: Valenčný slovník češtiny [Valency Dictionary of the Czech Language]

16. 6. 2003

ROSEN, ALEXANDER: Nástroje pre paralelné korpusy [Tools for Parallel Corpora]

22. 9. 2003

HAJIČOVÁ, EVA: K otázkám hloubkové syntaktické anotace velkých textových korpusů [Towards the Underlying Structure Annotation of a Large Corpus of Texts]

6. 10. 2003

PANEVOVÁ, JARMILA – LOPATKOVÁ, MARKÉTA: Valence a Pražský závislostní korpus (PDT) [Valency and the Prague Dependency Treebank (PDT)]

27. 10. 2003

GIGER, MARKUS: Problém delimitácie analytických slovesných tvarov [On the Delimitation of Analytic Verbal Forms]

10. 11. 2003

Návrh morfológického tagsetu Slovenského národného korpusu – oponentské konanie [Proposal of Morphological Tag Set of the Slovak National Corpus – external examination process]

1. 12. 2003

TADIĆ, MARKO: Hrvatski narodni korpus i njegovo obilježavanje (Chorvátsky národný korpus a jeho značkovanie) [Croatian National Corpus and Its Annotation]

8. 12. 2003

GORJANC, VOJKO: Sledenje leksikalnim spremembam v referenčnem korpusu slovenskega jezika (Sledovanie lexikálnych zmien v referenčnom korpuse slovinského jazyka) [Monitoring of Lexical Changes within Reference Corpus of the Slovenian Language]

7. 6. 2004

RYCHLÝ, PAVEL: Budoucnost korpusového manažeru Manatee [Future of the Corpus Manager Manatee]

28. 6. 2004

SOKOLOVÁ, MILOSLAVA: Možnosti vytvorenia slovenskej morfológie na korpusovom základe [Options for the Generation of a Corpus-Based Slovak Morphology]

# INSIGHT INTO THE SLOVAK AND CZECH CORPUS LINGUISTICS

Editor  
MÁRIA ŠIMKOVÁ

Návrh obálky Eva Kovačevičová-Fudala

Zodpovedný redaktor Emil Borčín

Prvé vydanie. Vydala VEDA, vydavateľstvo Slovenskej akadémie vied, v Bratislave roku 2006 ako svoju 3565. publikáciu z tlačových podkladov Jazykovedného ústavu Ľudovíta Štúra. 208 strán.

ISBN 80-224-0880-8