NATURAL LANGUAGE PROCESSING CENTRE
FACULTY OF INFORMATICS
MASARYK UNIVERSITY
BRNO

After Half a Century of Slavonic Natural Language Processing

D. Hlaváčková, A. Horák, K. Osolsobě, P. Rychlý (Eds.)

Published by Tribun EU
Brno 2009

Proceedings Editors

Dana Hlaváčková, Aleš Horák, Pavel Rychlý
Faculty of Informatics, Masaryk University, Brno
Department of Information Technologies
Botanická 68a
CZ-602 00 Brno, Czech Republic
E-mail: {hlavack,hales,pary}@fi.muni.cz

Klára Osolsobě
Faculty of Arts, Masaryk University, Brno
Department of Czech Language
Arna Nováka 1
CZ-602 00 Brno, Czech Republic
E-mail: osolsobe@phil.muni.cz

ISBN 978-80-7399-815-8

Printed in Czech Republic

Preface

Professor Karel Pala – Septuagenarian

Karel Pala, now celebrating the jubilee of entering his seventies, was born in Zlín on July 15, in the ominous year of 1939, when the world stood on the edge of World War II, the consequences of which imprinted its form and pressure onto the post-war years of totality, in which he grew up, was educated, and lived through for more than half of his life.

After having finished secondary school in Hlučín he attended the College of Russian Language (Vysoká škola ruského jazyka) in Prague, specialized in Czech/Russian translation. Under the guidance of professor Petr Sgall, he completed his post-graduate training in Mathematic Linguistics, Logic, and Informational Science at Charles University in Prague (1962-64), obtaining the academic degree of *candidatus scientiarum* (CSc.) in Czech Language. At the same time, he concentrated on research in the field of the formal description of Czech syntax and functional sentence perspective. In 1964 Pala returned to his native Moravia, this time to Brno, where he worked in diverse functions at the Department of Czech Language at Jan Evangelista Purkyně University (previously and currently Masaryk University). Together with Jiří Levý (+1967), a pioneer of translation-theory, he formed the Group for Exact Methods and Interdisciplinary Studies.

For his political views he was prohibited from teaching regular lessons. Nevertheless, he tutored seminars in Czech Syntax and yielded optional lessons in Mathematic and Computational Linguistics, which became the field of his future research and publishing. He developed (with Pavel Materna and Aleš Svoboda) the so-called Three-Part Theory for the Description of Natural Language, and he critically examined the usage of formal grammar for the description of Czech language and its analyses, both syntactic and semantic. Pala also participated in the implementation of the experimental syntactic-semantic analyzator for Czech (implementation in the LISP programming language). Since 1988 he has been the leader of a work group developing instruments for the automatic morphological analysis of Czech language.

Though Pala made little effort to hide his antipathy to the political regime of the time, he succeeded in traveling abroad for the purpose of study and experience. His refusal, however, to collaborate with the Communist State Security (StB) as a secret agent at the School of Slavonic and East European Studies at London University, resulted in his dismissal from his lectorate at the same institution (1972-1973). He employed his London experience afterwards in Brno as a lector at the Summer School of Slavonic Studies (1978-1991), and even more considerably, in his teaching years during the political thaw of 'perestrojka'. In those years he managed only once to move beyond the Iron Curtain, for the Summer School in Computational Linguistics in Pisa, Italy. In 1988, at the close of the communist era, Pala went to the Hungarian Academy of Sciences in Budapest for a month sojourn. When in 1989 the Czech borders were re-opened after forty years, Pala received a three-month invitation to the Institut für Slavistik at the University of Regensburg, Germany.

The Velvet Revolution in the former Czechoslovakia brought a new beginning for those whose professional carriers has been suspended for decades by the communist

regime. Pala submitted and defended his habilitation theses on computational linguistics and automatic processing of natural language and has been awarded the degree of Associate Professor (*Docent*) in 1993.

The new era of political normalcy brought many new opportunities for academic research and pedagogy and Pala was still at his apogee. He then started teaching regular courses of linguistic methodology and computational linguistics, and with a circle of colleagues, joined through common interest in the field of automatic processing of the Czech language, became the founder of the Computational Fund of the Czech Language (*Počítačový fond češtiny* – PFČ), an association for the support and co-ordination of manifold efforts in natural language processing. The team headed or co-headed by Pala was successively awarded by several academic grants, and he bore responsibility for a number of the grant projects. The first of these was entitled *Corpus of Czech Written Texts*, the subsequent *Czech in the Age of Computers*, as well as many others. They were the fruit of a collaboration of experts from Charles University, Prague, Masaryk University, and the Institute for Czech Language under the Czech Academy of Sciences.

In 1993-1995, twenty years after his first sojourn, Karel Pala returned to the School of Slavonic and East European Studies in London. Having returned to Brno, he accepted Jiří Zlatuška's invitation to the newly founded Faculty of Informatics at which he has been working since 1995 at the Department of Informatic Technologies (which he himself has successfully headed since 1998). Pala and Jiří Zlatuška were acquaintances from the time of their earlier collaboration on the book *Logical Analysis of Natural Language*, published in 1990 with Pavel Materna). In 1997 Pala founded the Laboratory for Natural Language Processing (*Laboratoř pro zpracování přirozeného jazyka*). He continues to teach the following courses: Introduction to Academic Writing, Introduction to Computational Linguistics, Introduction to Corpus Linguistics and Lexicography, Introduction to Machine–Translation and its Theory, Semantics and Communication. He is also dedicated to teaching post-graduate students and supervises their doctoral projects with enormous success; twelve of his doctoral students have already obtained their Ph.D degrees. He continues to publish indefatigably in many respected journals, the proceedings of international conferences, tirelessly participates in *EuroWordNet-2, Balkanet*, and the Language Advisory Centre (*Jazyková poradna*), and is also in charge of the grant project *Semantic Analysis of the Natural Language – TIL*. He is a member of the Academic Senate at the Faculty of Informatics and of many professional boards and scientific councils. He was awarded with an honorary membership in the scientific boards of the *Text, Speech, Dialogue* and *Global Wordnet* conferences, having co-edited both of their proceedings (2004).

Karel Pala is currently the chair of the Centre for Natural Language Processing at the Faculty of Informatics at Masaryk University.

Our Festschrift is to pay honours to our beloved university teacher and to one of the leading experts in the field of natural language processing (NLP), the improvement of which Pala's professional life was generously dedicated to.

An usual part of such retrospective occasions as a jubilee are congratulations, grateful thanks, best wishes, and personal memories – so, what follows is one of my most endearing memories.

When I entered Karel Pala's office at the Faculty of Arts for the first time, my eyes met, and I took well note of, and not without surprise, a provocative inscription on his door: *Order is for idiots, an intelligent person will cope with confusion.*

I have ruminated its playfully arrogant message quite often. Nowadays, with a distance of twenty-five years, I have come to find it deeper than it seemed to be to me at that time. The Latin verb *inter-legere (intellegere)* is related to the substantive *intellectus*, and by virtue of its etymology is also connected with the Czech (as well as English) expression "intellectual" (person). The verb *intellegere*, however, originally meant "to choose between alternatives", "to discern", or "to see a distinction". On this occasion, I would like to express my sincere thanks to Karel Pala, as my teacher and supervisor, for the fact that he has never held computers as a means to *simplify* things, not to say *confound* our natural language, quite the contrary, he saw them as a means of discernment and distinction-making. It is a matter of course that any instrument, computers not excepting, fulfills its function only when it is used properly and with discretion, by the experienced hands of one who is fully conscious of his/her purpose, the purpose of a linguist in general, and of Karel Pala in particular, always having been and being – *to discern*.

Klára Osolsobě

Sedmdesátiny doc. PhDr. Karla Paly, CSc.

Docent PhDr. Karel Pala, CSc., se narodil ve Zlíně 15. června 1939, v roce kdy naše vlast i celý svět stál na prahu války, jejíž důsledky pak měly vtisknout ráz následujícím létům totality, v nichž jubilant rostl, studoval a prožil víc než polovinu svého života.

Po skončení gymnaziálních studií v Hlučíně nastoupil na Vysokou školu jazyka ruského a literatury v Praze se specializací překladatelskou (čeština a ruština). Pod vedením profesora Petra Sgalla absolvoval postgraduální kurs z matematické lingvistiky, logiky a informatiky na Karlově universitě (1962-64) a v roce 1973 získal titul CSc. v oboru český jazyk. Jeho školitelem na Karlově univerzitě byl profesor Sgall. Paralelně byl zaměstnán v Ústavu pro jazyk český ČS AV v Praze jako praktikant na studijním pobytu a věnoval se výzkumu v oblasti formálního popisu české syntaxe a funkční perspektivy větné. V roce 1964 se odebral zpět na rodnou Moravu, tentokrát do Brna, kde působil v letech 1964-1995 v různých funkcích na Katedře nyní Ústavu českého jazyka Filozofické fakulty tehdejší Univerzity Jana Evangelisty Purkyně (od roku 1990 Masarykovy).

Ačkoliv se pro politické názory nesměl věnovat pedagogické činnosti, vedl semináře z české skladby a posléze i výběrové přednášky a semináře z matematické a počítačové lingvistiky. K těmto disciplinám byla zaměřena jeho badatelská i publikační činnost. V rámci tzv. státních výzkumných úkolů se zabýval spolu s Pavlem Maternou a Alešem Svobodou vytvořením trojsložkové teorie pro popis přirozeného jazyka, dále využitím formálních gramatik pro popis češtiny a výzkumy v oblasti automatické syntaktické

a sémantické analýzy češtiny. Spolupracoval na implementaci experimentálního syntaktického a sémantického analyzátoru pro češtinu (implementováno v programovacím jazyce LISP). Od roku 1988 pak vedl tým pracující na vybudování nástrojů automatické morfologické analýzy češtiny.

Přestože příliš neskrýval své antipatie k stávajícímu režimu, podařilo se Karlu Palovi získat tolik potřebné odborné zkušenosti i v zahraničí. Odmítnutí spolupracovat se státní bezpečností vedlo v roce 1973 k předčasnému ukončení práce lektora českého jazyka na School of Slavonic and East European Studies na londýnské univerzitě (1972-1973). Zkušenosti z Londýna mohl využít až v době „politického oteplení – perestrojky" jako vyučující na Letní škole slovanských studií v Brně (1978-1991). Za železnou oponu se ovšem dostal již jen na měsíční kurs na Letní školu počítačové lingvistiky v italské Pise. Na sklonku komunismu odjel ještě v roce 1988 na měsíční stáž na Jazykovědný ústav Maďarské Akademie věd do Budapešti. Poté, co se roku 1989 otevřely hranice, nabídl mu v následujícím roce 1990 tříměsíční studijní pobyt Institut für Slavistik na Univerzitě v bavorském Řezně.

Se sametovou revolucí nastal obrat v životě mnoha jubilantových vrstevníků, pro které, tak jako pro něj, nonkonformní postoje vůči komunistickému režimu znamenaly nonkomfortní postavení v rámci toho, čemu se dnes říká kariérní růst. Docentskou habilitaci v oboru český jazyk (se zaměřením na komputační lingvistiku a počítačové zpracování přirozeného jazyka) tudíž Karel Pala podal a obhájil až na počátku 90. let na tehdy již opět Masarykově univerzitě v Brně v roce 1993.

Skutečnost, že se Karel Pala, bezprostředně po převratu – tehdy na vrcholu života – v „normálním" a nikoli „normalizovaném" světě občansky angažuje, jej ovšem nikterak neodvádí od vědecké práce. Po roce 1989 se otevírají další možnosti pro bádání i pedagogickou činnost. Karel Pala vede kursy zaměřené na lingvistickou metodologii a počítačovou lingvistiku. V roce 1992 zakládá spolu se skupinou kolegů badatelů, které spojuje zájem o rozvoj počítačového zpracování češtiny, zájmové sdružení *Počítačový fond češtiny* (PFČ), jehož cílem je koordinovat úsilí a zajišťovat komunikaci a spolupráci odborníků, kteří mají zájem o počítačové zpracování českého jazyka. Posléze tyto snahy nabyly institucionalizované podoby ve formě řady grantových projektů. (Vůbec první grantový projekt nesl název „Počítačový korpus českých psaných textů", další „Čeština ve věku počítačů". Spolupracovali na nich odborníci Univerzity Karlovy v Praze, Masarykovy univerzity v Brně a Ústavu pro jazyk český). Karel Pala stál na předním místě řešitelských týmů.

V letech 1993-1995 se Karel Pala po dvaceti letech vrátil jako lektor českého jazyka na School of Slavonic and East European Studies do Londýna. Po návratu z Anglie přijal návrh svého kolegy, Jiřího Zlatušky, (v roce 1990 spolu s P. Maternou a J. Zlatuškou vydal Karel Pala monografii „*Logická analýza přirozeného jazyka*") a přešel v září 1995 na nově založenou Fakultu informatiky Masarykovy univerzity. Působí zde na Katedře informačních technologií, kterou od roku 1998 úspěšně vedl. V roce 1997 založil Laboratoř zpracování přirozeného jazyka. Vyučuje kursy Základy odborného stylu, Úvod do počítačové lingvistiky, Úvod do korpusové lingvistiky a počítačové lexikografie, Úvod do strojového překladu a kurs Sémantika a komunikace a úspěšně se věnuje celé řadě doktorandů (12 obhájilo doktorskou práci, 8 pokračuje v doktorandském studiu). Publikuje nepřehlednou řadu studií ve sbornících z řady mezinárodních prestižních konferencí a podílí se (EuroWordNet-2, Balkanet, Velké

korpusy, Jazyková poradna) i vede (Sémantická analýza přirozeného jazyka, TIL) grantové projekty. Je členem Akademického senátu FI MU, členem oborové rady, vědecké rady a konkursní komise tamtéž. Mimo svoji mateřskou fakultu působí jako člen oborové rady MFF UK (počítačová lingvistika) a člen konkursní komise FF MU (bohemistika a počítačová lingvistika) a též člen vědecké rady ÚJČ AV ČR. O mezinárodním ocenění jeho odborných schopností a zkušeností svědčí členství v programovém výboru konference TSD (Text, Speech, Dialogue – redaktorství konferenčních sborníků) a Global Wordnet Conference (editor sborníku z roku 2004).

V současné době stojí v čele Centra zpracování přirozeného jazyka FI MU.

Předkládaným sborníkem bychom rádi poctili jednoho z předních českých vědců, badatele i učitele, který zasvětil svůj život tomu, čemu se dnešním jazykem vědy říká *Natural Language Processing* (NLP).

Nedílnou součástí bilancování při příležitostech životních jubileí bývají vzpomínky a přání; patří se však i poděkovat. Dovolte mi tedy závěrem malý osobní dodatek.

Když jsem poprvé vstoupila do kanceláře dr. Karla Paly na Filozofické fakultě tehdejší UJEP v Brně, padl mi do oka nápis na dveřích: „*Pořádek je pro blbce, inteligent zmatek zvládne.*“ Často jsem se k obsahu tohoto na první pohled furiantsky působícího sdělení vracívala. Dnes si s odstupem téměř čtvrt století myslím, že je hlubší, než zprvu vypadalo. Latinské sloveso *inter-legere* (*intellergere*), které souvisí se substantivem *intellectus* a je etymologicky spojeno s českým *inteligent*, znamená doslova *vybírat si mezi možnostmi*, tedy *roz-lišovat*. Na tomto místě bych ráda svému učiteli, Karlu Palovi, poděkovala, že (navzdory dennodenní jazykové realitě zjednodušující a matoucí pojmy až k samé ztrátě smyslu řečeného) nikdy neviděl v počítačích nástroj zjednodušení, či spíše *zmatení*, přirozeného jazyka, ale vždy nástroj *rozlišování*. To, že nástroj má správné uplatnění pouze v rukou toho, kdo si je vědom, že má pěstovat svou schopnost *rozlišovat*, bylo a je samozřejmé.

Dík, Karle, a vše nejlepší do dalších let!

Klára Osolsobě

August 2009 D. Hlaváčková, A. Horák, K. Osolsobě, P. Rychlý (Eds.)

Table of Contents

Presupposition vs. allegation

Marie Duží

VSB-Technical University Ostrava,
17. listopadu 15, 708 33 Ostrava
marie.duzi@vsb.cz

Abstract. In the paper we examine two kinds of entailment, viz. presupposition and allegation. We show that the topic of a sentence is associated with the existence of the object denoted by the topic, which is not only entailed but also presupposed by the sentence. On the other hand, allegation is often triggered by the focus of a sentence, and this is the case of mere entailment. Hence sentences differing only in their topic-focus articulation may have different truth-conditions and should thus have assigned different logical forms. To this end we apply procedural semantics of Transparent Intensional Logic (TIL), and furnish sentences with hyperpropositions that are precisely defined in terms of TIL *constructions*. These are *procedures* assigned to sentences as their context-invariant structured meanings. We analyse the phenomenon of *topic-focus* distinction, in particular the case of a presupposition connected with the topic and allegation triggered by the focus of a sentence, in such a way that relevant consequences can be formally derived.

1 Introduction

There has been much dispute among theoretical linguists and logicians on whether the problem of topic-focus articulation is the problem of semantics rather than pragmatics. In this paper we are going to demonstrate the *semantic* nature of the topic-focus difference by its *logical* analysis. To this end we apply procedural semantics of Transparent Intensional Logic (TIL) and assign (algorithmically structured) procedures to expressions as their meanings. As a result, we furnish sentences differing only in the topic-focus articulation with different structured meanings producing different PWS-propositions.

By analysing sample sentences we are going to show that while the clause standing in the topic often generates the case of a presupposition, a focus-clause usually entails rather than presupposes another proposition. The difference between the two cases is this. A sentence Q is entailed by a sentence P if whenever P is true, Q is true as well. If P is not true, we cannot deduce anything about the truth of Q. On the other hand, Q is not only entailed by also presupposed by P, if Q is entailed both by P and non-P. In other words, if a presupposition Q of a given proposition P is not true, then P as well as negated P have no truth-value. Since our logic is a hyper-intensional logic of *partial functions*, we analyse sentences with presuppositions in a natural way. We furnish them with hyper-propositions that produce PWS-propositions with truth-value gaps. Having a rigorous, fine-grained analysis at our disposal, we can then easily infer the relevant consequences.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 1–15, 2009.

The paper is organised as follows.[1] After briefly introducing TIL philosophy and its basic notions in Section 2, the main Section 3 introduces the method of analysing sentences with presuppositions induced by a sentence topic, and allegation triggered by a focus clause. Concluding Section 4 presents the direction of future research and a few notes on TIL implementation via the *TIL-Script* functional programming language.

2 TIL in brief

Transparent Intensional Logic (TIL) is a system with procedural semantics primarily designed for the logical analysis of natural language.[2] Traditional non-procedural theories of formal semantics are less or more powerful logical languages, from the extensional languages based on the first-order predicate logic paradigm, through some hybrid systems up to intensional (modal or epistemic) logics. Particular systems are well suited for analysing restricted sublanguages. Yet there are hard cases like attitudes, anaphora, or the topic-focus articulation that are stumbling blocks for all of them. This is due to the fact that any intensional logic without hyper-intensional features can individuate meanings only up to equivalence. Equivalent but non-synonymous expressions are indistinguishable.

On the other hand, TIL, due to its *procedural semantics* based on strong typing, operates smoothly with the three levels of granularity: the extensional level of truth-functional connectives, the intensional level of modalities and finally the hyper-intensional level of attitudes.[3] The sense of a sentence is an algorithmically structured *construction* of a proposition denoted by the sentence. The denoted proposition is a flat mapping with the domain of possible worlds. Our motive for working 'top-down' has to do with anti-contextualism: any given unambiguous term or expression (even one involving indexicals or anaphoric pronouns) expresses the same construction as its sense (meaning) in whatever sort of context the term or expression is embedded within. And the meaning of an expression determines (possibly dependently on the situation of utterance in case of a pragmatically incomplete meaning) the respective denoted entity (if any), but not vice versa. Thus we strictly distinguish between a procedure (construction) and its product (constructed function), and between a function and its value.

Intuitively, construction C is a *procedure* (a generalised algorithm). When assigning constructions to expressions as their meanings, we specify *procedural know-how*, which must not be confused with the respective *performatory know-how*.[4] Understanding a sentence S involves procedural know-how; one can spell out instructions for evaluating the truth-conditions of S in any state-of-affairs w at any time t. But, of course, one can know how to evaluate S without actually being able to do so, that is, without having the performatory skills that enable them to determine the truth-value of S in a particular possible world w and time t. Constructions are structured in the following way. Each construction C consists of sub-instructions (constituents), the execution of which is involved when executing C. It is an instruction on how to proceed in order to obtain the output entity given some input entities.

[1] The previous version of this paper was read by Marie Duží at CICLing 2008, *Computational Linguistics and Intelligent Text Processing* conference in Haifa, Israel, see [1].

[2] See, for instance, [2], [7], [8], [12] and [13].

[3] For TIL analysis of anaphoric references, see, e.g., [3], and for attitudes [4].

[4] See [9], pp.6–7.

There are two kinds of constructions, atomic and compound (molecular). Atomic constructions (*Variables* and *Trivializations*) do not contain any other constituent but themselves; they supply objects (of any type) on which compound constructions operate. *Variables x, y, p, q, ...,* construct objects dependently on a valuation; they *v*-construct. *Trivialisation* of an object X (of any type, even a construction), in symbols ${}^0\boldsymbol{X}$, constructs simply X without the mediation of any other construction. *Compound* constructions, which consist of other constituents, are *Composition* and *Closure*. *Composition* $[\boldsymbol{F}\, \boldsymbol{A}_1 \ldots \boldsymbol{A}_n]$ is the instruction to apply a function f (*v*-constructed by F) to a tuple argument A (*v*-constructed by $A_1 \ldots A_n$).[5] Thus it *v*-constructs the value of f at A, if the function f is defined at A, otherwise the Composition is *v-improper*, i.e., it *fails* to *v*-construct anything. *Closure* $[\lambda \boldsymbol{x}_1 \ldots \boldsymbol{x}_n\, \boldsymbol{X}]$ is the instruction to *v*-construct a function by abstracting over values of variables $x_1,\ldots,x_n$ in the ordinary manner of λ-calculi. Finally, higher-order constructions can be used twice over as constituents of composed constructions. This is achieved by a fifth construction called *Double Execution*, ${}^2\boldsymbol{X}$, that behaves as follows: If X *v*-constructs a construction X', and X' *v*-constructs an entity Y, then ${}^2X$ *v*-constructs Y; otherwise ${}^2X$ is *v-improper*, it *fails* to *v*-construct anything.

TIL constructions, as well as the entities they construct, all receive a type. The formal ontology of TIL is bi-dimensional; one dimension is made up of constructions, the other dimension encompasses non-constructions. On the ground level of the type-hierarchy, there are non-constructional entities unstructured from the algorithmic point of view belonging to a *type of order 1*. Given a so-called *epistemic (or 'objectual') base* of *atomic types* (*o*-truth values, ι-individuals, τ-time moments / real numbers, ω-possible worlds), the induction rule for forming functional types is applied: where α, $\beta_1,\ldots,\beta_n$ are types of order 1, the set of partial mappings from $\beta_1 \times \ldots \times \beta_n$ to α, denoted $(\alpha\beta_1 \ldots \beta_n)$, is a type of order 1 as well.[6] Constructions that construct entities of order 1 are *constructions of order 1*. They belong to a *type of order 2*, denoted by $*_1$. This type $*_1$ together with atomic types of order 1 serves as a base for the induction rule: any collection of partial mappings, type $(\alpha\beta_1 \ldots \beta_n)$, involving types of order 1 and $*_1$ in their domain or range is a *type of order 2*. Constructions belonging to a type $*_2$ that identify entities of order 1 or 2, and partial mappings involving such constructions, belong to a *type of order 3*. And so on *ad infinitum.*

The sense of an empirical expression is a *hyper-intension*, i.e., a construction that produces a possible-world intension defined as follows:

(α-)intensions are members of type $(\alpha\omega)$, i.e., functions from possible worlds to an arbitrary type α.

(α-)extensions are members of a type α, where α is not equal to $(\beta\omega)$ for any β, i.e., extensions are not functions with the domain of possible worlds.

Intensions are frequently functions of a type $((\alpha\tau)\omega)$, i.e., functions from possible worlds to *chronologies* of the type α (in symbols: $\alpha_{\tau\omega}$), where a chronology is a function of type $(\alpha\tau)$.

Some important kinds of intensions are:

[5] We treat functions as mappings, i.e., set-theoretical objects, unlike the *constructions* of functions.

[6] TIL is an open-ended system. The above epistemic base $\{o, \iota, \tau, \omega\}$ was chosen, because it is apt for natural-language analysis, but the choice of base depends on the area to be analysed.

Propositions, type $o_{\tau\omega}$. They are denoted by empirical sentences.
Properties of members of a type α, or simply α*-properties,* type $(o\alpha)_{\tau\omega}$.[7] General terms, some substantives, intransitive verbs ('student', 'walking') denote properties, mostly of individuals.
Relations-in-intension, type $(o\beta_1 \dots \beta_m)_{\tau\omega}$. For example transitive empirical verbs ('like', 'worship'), also attitudinal verbs denote these relations.
α*-roles, offices*, type $\alpha_{\tau\omega}$, where $\alpha \neq (o\beta)$. Frequently $\iota_{\tau\omega}$. Often denoted by concatenation of a superlative and a noun ('the highest mountain').

An object A of a type α is denoted A/α. That a construction $C/*_n$ *v*-constructs an object of type α is denoted $C \rightarrow_v \alpha$. We use variables $w, w_1, \dots$ as *v*-constructing elements of type ω (possible worlds), and $t, t_1, \dots$ as *v*-constructing elements of type τ (times). If $C \rightarrow_v \alpha_{\tau\omega}$ *v*-constructs an α-intension, the frequently used Composition of the form [[*Cw*]*t*], the intensional descent of an α-intension, is abbreviated as $\boldsymbol{C}_{wt}$.

We invariably furnish expressions with their procedural structural meanings, which are explicated as TIL constructions. The analysis of an expression thus consists in discovering the logical construction encoded by the expression. *TIL method of analysis* consists of three steps:[8]

1. *Type-theoretical analysis*, i.e., assigning types to the objects that receive mention in the analysed sentence.
2. *Synthesis*, i.e., combining the constructions of the objects *ad* (1) in order to construct the proposition of type $o_{\tau\omega}$ denoted by the whole sentence.
3. *Type-theoretical checking.*

As an example we are going to analyse the proverbial sentence "The King of France is bald". The sentence talks about the office of the King of France (topic) ascribing to the individual (if any) that occupies this office the property of being bald (focus). Thus there is a presupposition that the King of France exists, i.e., that the office is occupied. If not, then the proposition denoted by the sentence has no truth-value.[9] This fact has to be revealed by our analysis. Here is how.

Ad (1) King_of/$(\iota\iota)_{\tau\omega}$; *France*/$\iota$; *King_of_France*/$\iota_{\tau\omega}$; *Bald*/$(o\iota)_{\tau\omega}$.

Ad (2) Now we combine *constructions* of the objects *ad (1)* in order to construct the proposition of type $o_{\tau\omega}$ denoted by the whole sentence. The simplest constructions of the above objects are their Trivialisations: ^{0}*King_of*, ^{0}*France*, ^{0}*Bald.* The attribute *King_of* has to be extensionalised first *via* Composition ^{0}*King_of*$_{wt}$, and then applied to *France*; we get [^{0}*King_of*$_{wt}$ ^{0}*France*]. Finally by abstracting over values of w, t we obtain the office, $\lambda w \lambda t$ [^{0}*King_of*$_{wt}$ ^{0}*France*]. But the property of being bald cannot be ascribed to an individual office. Rather, it is ascribed to an individual occupying the office.

[7] We model α-sets and $(\alpha_1 \dots \alpha_n)$-relations by their characteristic functions of type $(o\alpha)$, $(o\alpha_1 \dots \alpha_n)$, respectively. Thus an α-property is an empirical function that dependently on states-of-affairs $(\tau\omega)$ picks-up a set of α-individuals, the population of the property.

[8] For details see, e.g., [8].

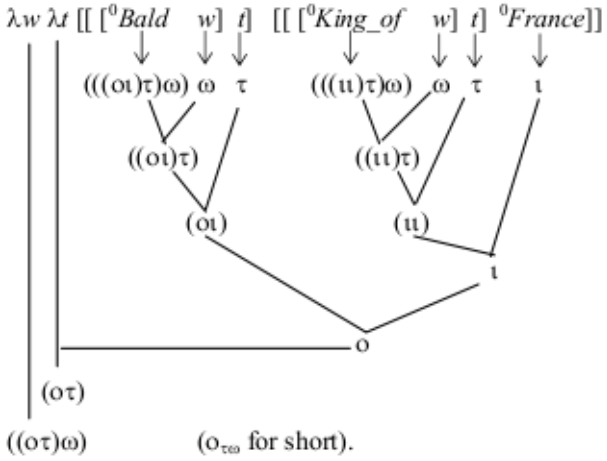
[9] On our approach this does not mean that the sentence is meaningless. The sentence has its sense, namely the instruction how to evaluate in any possible world w at any time t its truth-conditions. Just that if we evaluate this instruction in such a state-of-affairs where the King of France does not exist, the process of evaluation yields a truth-value gap.

Thus the office has to be subjected to the intensional descent to *v*-construct an individual (if any) occupying the office: $\lambda w \lambda t$ [^{0}*King_of*$_{wt}$ ^{0}*France*]$_{wt}$. The property itself has to be extensionalised as well, ^{0}*Bald*$_{wt}$. Composing these two constructions, we obtain a truth-value **T**, **F**, or nothing, according as the King of France is or is not bald, or does not exist.[10] Finally, abstracting over values of w, t, we construct the proposition:

$$\lambda w \lambda t \, [\, ^0Bald_{wt} \, \lambda w \lambda t \, [\, ^0King\_of_{wt} \; ^0France]_{wt}].$$

This Closure can be equivalently simplified into $\lambda w \lambda t$ [^{0}*Bald*$_{wt}$ [^{0}*King_of*$_{wt}$ ^{0}*France*]]. *Gloss*. In *any* world at *any* time ($\lambda w \lambda t$) do these: First, find out who is the King of France by applying the extensionalised attribute *King_of* to *France* ([^{0}*King_of*$_{wt}$ ^{0}*France*]). If there is none, then finish with the truth-value gap because the Composition [^{0}*King_of*$_{wt}$ ^{0}*France*] is *v-improper*. Otherwise, check whether the so-obtained individual has the property of being bald ([^{0}*Bald*$_{wt}$ [^{0}*King_of*$_{wt}$ ^{0}*France*]]). If so, then **T**, otherwise **F**.

Ad (3). Drawing a type-theoretical structural tree,[11] we check whether particular constituents of the above Closure are combined in a type-theoretically correct way.



So much for the semantic schema of TIL logic. Now we are going to apply this formal apparatus to analyse the phenomena of presupposition and allegation. We will focus in particular on the way how the topic-focus distinction determines which of the two phenomena the case is.

3 Topic-Focus articulation and presuppositions

In this section we propose the method of *logically* analysing sentences with presuppositions. The input for our analysis is the result of linguistic analysis, such that it reflects the topic-focus articulation. When used in a communicative act, the sentence communicates something (the focus F) about something (the topic T). Thus the schematic structure of a sentence is $F(T)$.The topic T of a sentence S is often associated with a presupposition

[10] For details on predication of properties see [6].

[11] See [2].

P of S such that P is entailed both by S and non-S. To start up, let us analyse some examples.[12]

(1) All *John's children* are asleep.

(1') All *John's children* are not asleep.

Strawson in [11, 173ff.] demonstrates that (1) as well as (1') entail that

(2) John has children.

In other words, (2) is a *presupposition* of (1), as well as of (1'). If each of John's children is asleep, then the sentence (1) is true and (1') false. If each of John's children is not asleep, then the sentence (1) is false and (1') is true. However, if John does not have any children, then (1) and (1') are neither true nor false. If (1) was false than (1') would be true, which means that "Some of John's children are still up", which entails that John has children contrary to the assumption. On the other hand, if (1) was false in the situation when John does not have any children, then (1') should be true, which again entails that John has children, contrary to the assumption.

However, applying a classical translation of (1) into the language of first-order predicate logic, we get

$$\forall x[JC(x) \supset S(x)].$$

But this formula is true under every interpretation assigning an empty set of individuals to the predicate *JC*. We need to apply a richer logical system in order to express the instruction how to evaluate the truth-conditions of (1) in the above described way.

Reformulating the specification of the truth-conditions of (1) in a rather technical jargon of English, we get

If John has children *then* true or false according as all John's children are asleep,
else fail (to produce a truth-value).

Since TIL meets the principle of Universal Transparency, i.e., TIL analysis is fully compositional, we first need to analyse particular constituents of this instruction, and then combine these constituents into the construction expressed by the sentence. As always, we start with assigning types to the objects that receive mention in the sentence:

John/ι, an individual;

Have/$(o\iota(o\iota)_{\tau\omega})_{\tau\omega}$, a relation-in-intension between an individual and a property of individuals, some instances of which the individual has;

(to be a) *Child*/$(o\iota)_{\tau\omega}$, a property of individuals;

All/$((o(o\iota))(o\iota))$, the restricted general quantifier that assigns to a given set the set of all its supersets;

[12] Some of the examples we are going to analyse were taken from [5], where Hajičová argues that topic-focus articulation is a matter of semantics rather than pragmatics. We agree, and in order to put her arguments on a still more solid ground, we explicitly demonstrate different logical constructions assigned to sentences differing only in topic/focus articulation. In what follows we mark the topic of a sentence in italics.

Children_of/$((o\iota)\iota)_{\tau\omega}$,a function-in-intension that dependently on states-of-affairs (possible world/ω and time/τ) assigns to an individual (of type ι) a set of individuals (of type $(o\iota)$), its children;

Sleep/$(o\iota)_{\tau\omega}$, a property of individuals.

The presupposition that John has children receives the analysis

$$\lambda w \lambda t[\ ^0\textit{Have}_{wt}\ \ ^0\textit{John}\quad ^0\textit{Child}].$$

The literal analysis of "All John's children are asleep" is best obtained by using the restricted quantifier *All*. Composing the quantifier with the set of John's children, [^{0}*All* [^{0}*Children_of*$_{wt}$ ^{0}*John*]], we obtain the set of all supersets of John's children population in a possible world w at time t. The sentence claims that the population of those who are asleep, ^{0}*Sleep*$_{wt}$, is one of such supersets:

$$\lambda w \lambda t[[\ ^0\textit{All}\ [\ ^0\textit{Children\_of}_{wt}\ ^0\textit{John}]]\ \ ^0\textit{Sleep}_{wt}].$$

So far so good; yet there is a problem how to analyse the connective '*if-then-else*'. There has been much dispute on the semantics of 'if-then-else' among computer scientists. We cannot simply apply material implication '$\supset$'. For instance, it might seem that the instruction "If 5=5 then output 1 else output the result of 1 divided by 0" receives the analysis $[[[^05 = {}^05] \supset [n = {}^01]] \wedge [\neg[^05 = {}^05] \supset [n = [\ ^0\textit{Div}\ {}^01\ {}^00]]]]$, where n is the outputted number.[13] But the output of the above instruction should be the number 1 because the 'else-clause' is never executed. However, due to strict compositionality, the above analysis fails to produce anything, the construction is improper. The reason is this. The Composition $[\ ^0\textit{Div}\ {}^01\ {}^00]$ does not produce anything, it is improper because the division function has no value at the argument $\langle 1, 0\rangle$. Thus the Composition $[n = [\ ^0\textit{Div}\ {}^01\ {}^00]]$ is *v-improper* for any valuation v, because the identity relation $=$ does not receive an argument, and so is any other Composition containing the improper Composition $[\ ^0\textit{Div}\ {}^01\ {}^00]$ as a constituent (partiality is strictly propagated up). This is the reason why the 'if-then-else' connective is often said to be a non-strict function.

However, there is no cogent reason to settle for non-strictness. We need to apply a mechanism known in computer science as *lazy evaluation.* The *procedural* semantics of TIL operates smoothly even in the level of constructions. Thus it enables us to specify a strict definition of 'if-then-else' that meets the compositionality constraint. The definition of "If P then C_1 else C_2" is a procedure that decomposes into two phases. First, on the basis of the condition P, select one of C_1, C_2 as the procedure to be executed. Second, execute the selected procedure.

First, the selection is realized by the Composition

$$[\ ^0\iota\lambda c\ [[P \supset [c = {}^0C]] \wedge [\neg P \supset [c = {}^0D]]]].$$

The Composition $[[P \supset [c = {}^0C]] \wedge [\neg P \supset [c = {}^0D]]]$ *v*-constructs either C or D. If P constructs **T** then the variable c receives as its value the *construction C*, and if P constructs **F** then the variable c receives the *construction D* as its value. In any case the

[13] For the sake of simplicity, we use infix notation without Trivialization when applying truth-value connectives and equality. Thus, for instance, instead of '$[\ ^0 \supset [\ ^0 = [\ ^05\ {}^05]][\ ^0 = [n\ {}^01]]]$' we write '$[[\ ^05 = {}^05] \supset [n = {}^01]]$'.

set constructed by λc [[$P \supset [c = {}^0C]] \wedge [\neg P \supset [c = {}^0D]]$] is a singleton. Applying singulariser ι on this set returns as its value the only member of the set, i.e., either the construction C or D. Thus we have [${}^0\iota\lambda c$ [[$P \supset [c = {}^0C]] \wedge [\neg P \supset [c = {}^0D]]$]].

Second, the chosen construction c is executed. As a result, the schematic analysis of "If P then C else D" is

(*) $$ {}^2[\,{}^0\iota\lambda c[[P \supset [c = {}^0C]] \wedge [\neg P \supset [c = {}^0D]]]]. $$

Types: $P \to o$, the condition of the choice between the execution of C or D, $C/*_n$, $D/*_n$; variable $c \to *_n$; $\iota/(*_n(o*_n))$, the singulariser function that associates a singleton set of constructions with the only construction that is an element of this singleton, otherwise (i.e., if the set is empty or many-valued) it is undefined.

Note that we do need a hyperintensional, procedural semantics here. We need variable c ranging over constructions. Moreover, evaluation of the first phase does not involve the execution of constructions C and D. These constructions are only arguments of other constructions.

Returning to the analysis of (1), in our case the condition P is that John has children, [${}^0Have_{wt}$ ${}^0John$ ${}^0Child$], the construction C that is to be executed if P yields **T** is [[${}^0All$ [${}^0Children\_of_{wt}$ ${}^0John$]] ${}^0Sleep_{wt}$], and if P yields **F** then *no* construction is to be chosen. Thus the analysis of the sentence (1) comes down to this Closure:

(1*) $$ \begin{array}{l} \lambda w\lambda t\, {}^2\big[\,{}^0\iota\lambda c[\,{}^0Have_{wt}\,{}^0John\,{}^0Child] \supset \\ \quad \big[c = {}^0[[\,{}^0All[\,{}^0Children\_of_{wt}\,{}^0John]]\,{}^0Sleep_{wt}] \\ \quad\quad \wedge\neg[\,{}^0Have_{wt}\,{}^0John\,{}^0Child] \supset {}^0\mathbf{F}\big]\big] \end{array} $$

The evaluation of (1*) in any world/time pair $\langle w, t\rangle$ depends on whether the presupposition condition [${}^0Have_{wt}$ ${}^0John$ ${}^0Child$] is true in $\langle w, t\rangle$.

a) [${}^0Have_{wt}$ ${}^0John$ ${}^0Child$] $\to_v$ **T**.
Then λc[${}^0\mathbf{T} \supset [c = {}^0$[[${}^0All$ [${}^0Children\_of_{wt}$ ${}^0John$]] ${}^0Sleep_{wt}$] $\wedge$ [${}^0\mathbf{F} \supset {}^0\mathbf{F}$]] =
{ ^{0}[[${}^0All$ [${}^0Children\_of_{wt}$ ${}^0John$]] ${}^0Sleep_{wt}$]}. Hence
${}^2\iota\lambda c$[${}^0\mathbf{T} \supset [c = {}^0$[[${}^0All$ [${}^0Children\_of_{wt}$ ${}^0John$]] ${}^0Sleep_{wt}$] $\wedge$ [${}^0\mathbf{F} \supset {}^0\mathbf{F}$]] =
${}^2\,{}^0$[[${}^0All$ [${}^0Children\_of_{wt}$ ${}^0John$]] ${}^0Sleep_{wt}$] =
[[${}^0All$ [${}^0Children\_of_{wt}$ ${}^0John$]] ${}^0Sleep_{wt}$].

b) [${}^0Have_{wt}$ ${}^0John$ ${}^0Child$] $\to_v$ **F**.
Then λc[${}^0\mathbf{F} \supset [c = {}^0$[[${}^0All$ [${}^0Child\_of_{wt}$ ${}^0John$]] ${}^0Sleep_{wt}$] $\wedge$ [${}^0\mathbf{T} \supset {}^0\mathbf{F}$]] = $\lambda c\,{}^0\mathbf{F}$.
The *v*-constructed set is *empty*. Hence, ${}^2[\iota\lambda c\,{}^0\mathbf{F}]$ is *v-improper*, *fails*.

To generalise, we now present a ***general analytic schema*** *of an (empirical) sentence S associated with a presupposition P.* The analysis is an instruction of the form

If P then S else *Fail*.

The corresponding schematic TIL construction is

(**) $$ \lambda w\lambda t\ {}^2[\iota\lambda c[[P_{wt} \supset [c = {}^0S_{wt}]] \wedge [\neg P_{wt} \supset {}^0\mathbf{F}]]]. $$

The evaluation in any $\langle w, t\rangle$-pair depends on the value *v*-constructed by P_{wt}.

(a) $P_{wt} \rightarrow_v \mathbf{T}$. Then
$^2[\iota\lambda c[[\,^0\mathbf{T} \supset [c = \ ^0S_{wt}]] \wedge [\,^0\mathbf{F} \supset \ ^0\mathbf{F}]]] = \ ^{2\,0}S_{wt} = S_{wt}$
(b) $P_{wt} \rightarrow_v \mathbf{F}$. Then
$^2[\iota\lambda c[[\,^0\mathbf{F} \supset [c = \ ^0S_{wt}]] \wedge [\,^0\mathbf{T} \supset \ ^0\mathbf{F}]]]$ is *v-improper*.
(c) P_{wt} is *v-improper*, then
$^2[\iota\lambda c[[P_{wt} \supset [c = \ ^0S_{wt}]] \wedge [\neg P_{wt} \supset \ ^0\mathbf{F}]]]$ is *v-improper*.

Another phenomenon we encounter when analysing sentences with topic-focus articulation is *allegation*.[14] Consider another group of sample sentences.

(3) *The King of France* visited London yesterday.

(3') *The King of France* did not visit London yesterday.

The sentences (3) and (3') talk about the (actual and current) King of France (the topic), ascribing to him the property of having (not having) visited London yesterday (the focus). Thus both the sentences have the presupposition that the King of France actually exists *now*. If it is not so, then none of the propositions expressed by (3) and (3') have any truth-value. The situation is different in case of sentences (4) and (4'):

(4) *London* was visited by the King of France yesterday.

(4') *London* was not visited by the King of France yesterday.

Now the property (in focus) of having been visited by the King of France yesterday is predicated of London (the topic). The existence of the King of France (now) is not presupposed by (4), and thus also not by (4'), of course. The sentences can be read as "Among the visitors of London was (was not) yesterday the King of France". The existence of the King of France *yesterday* is *implied*, but *not presupposed*, by (4).

To describe the difference between the cases such as (3) and (4), Hajičová in [5] characterizes *allegation* like this: while (i) *presupposition* is characterised as an assertion A entailed by an assertion carried by a sentence S, and also by the negation of S, (ii) an *allegation* is an assertion A entailed by an assertion carried by a sentence S, but the negative counterpart of S entails neither A nor its negation. Schematically,

(i) $(S \models A)$ and $(\neg S \models A)$ (A is a ***presupposition*** of S);
Corollary: If $\neg A$ then *neither* S *nor* $\neg S$ have any truth-value.
(ii) $(S \models A)$ and *neither* $(\neg S \models A)$ nor $(\neg S \models \neg A)$ (***allegation***).

Our analyses respect these conditions. Let *Yesterday*/$((o\tau)\tau)$ be the function that associates a given time t with a time-interval (that is yesterday with respect to t); *Visit*/$(o\iota\iota)_{\tau\omega}$; *King_of*/$(\iota\iota)_{\tau\omega}$; *France*/$\iota$.
Remark. Now we are going to make use of (unrestricted) quantifiers, existential, $\exists^\tau$ and general one, $\forall^\tau$. They are functions of type $(o(o\tau))$. The existential quantifier assigns to a given set of times the truth-value **T** if the set is non-empty, otherwise **F**. The general quantifier assigns to a given set of times the truth-value **T** if the set is the

[14] The term 'allegation' is due to B. Partee. See [5, 248-249].

whole type τ, otherwise **F**. In what follows we will use an abbreviated notation without Trivialisation; instead of '$[\,^0\forall^\tau \lambda x A]$', '$[\,^0\exists^\tau \lambda x A]$' we write '$\forall x A$', '$\exists x A$'. Thus the analyses of sentences (3), (3') come down to

(3*) $\lambda w\lambda t\,[\lambda x \exists t'[[[\,^0\mathit{Yesterday}\; t]t'] \wedge [\,^0\mathit{Visit}_{wt'}\; x\; {}^0\mathit{London}]][\,^0\mathit{King\_of}_{wt}\; {}^0\mathit{France}]]$

(3'*) $\lambda w\lambda t\,[\lambda x[\exists t'[[[\,^0\mathit{Yesterday}\; t]t'] \wedge \neg[\,^0\mathit{Visit}_{wt'}\; x\; {}^0\mathit{London}]][\,^0\mathit{King\_of}_{wt}\; {}^0\mathit{France}]]$

In such a $\langle w, t\rangle$-pair in which the King of France does not exist both the propositions constructed by (3*) and (3'*) have no truth-value, because the Composition $[\,^0\mathit{King\_of}_{wt}\; {}^0\mathit{France}]$ is *v-improper*. On the other hand, the sentences (4), (4') express

(4*) $\lambda w\lambda t\,\exists t'[[[\,^0\mathit{Yesterday}\; t]t'] \wedge [\,^0\mathit{Visit}_{wt'}[\,^0\mathit{King\_of}_{wt'}\; {}^0\mathit{France}]\; {}^0\mathit{London}]]$

(4'*) $\lambda w\lambda t\,\exists t'[[[\,^0\mathit{Yesterday}\; t]t'] \wedge \neg[\,^0\mathit{Visit}_{wt'}[\,^0\mathit{King\_of}_{wt'}\; {}^0\mathit{France}]\; {}^0\mathit{London}]]$

Now in such a $\langle w, t\rangle$-pair in which the proposition constructed by (4*) is true, the Composition $\exists t'[[[\,^0\mathit{Yesterday}\; t]t'] \wedge [\,^0\mathit{Visit}_{wt'}[\,^0\mathit{King\_of}_{wt'}\; {}^0\mathit{France}]\; {}^0\mathit{London}]]$ *v*-constructs the truth-value **T**. This means that the second conjunct *v*-constructs **T** as well. Hence $[\,^0\mathit{King\_of}_{wt'}\; {}^0\mathit{France}]$ is not *v*-improper, which means that the King of France *existed in some time* t' belonging to *yesterday*. On the other hand, if the King of France did not exist at any time of yesterday, then the Composition $[\,^0\mathit{King\_of}_{wt'}\; {}^0\mathit{France}]$ is *v-improper* for any t' belonging to yesterday. Thus the time interval *v*-constructed by $\lambda t'[[[\,^0\mathit{Yesterday}\; t]t'] \wedge [\,^0\mathit{Visit}_{wt'}[\,^0\mathit{King\_of}_{wt'}\; {}^0\mathit{France}]\; {}^0\mathit{London}]]$, as well as by $\lambda t'[[[\,^0\mathit{Yesterday}\; t]t'] \wedge \neg[\,^0\mathit{Visit}_{wt'}[\,^0\mathit{King\_of}_{wt'}\; {}^0\mathit{France}]\; {}^0\mathit{London}]]$, is empty, and the existential quantifier takes this interval to the truth-value **F**. This is as it should be, because (4*) *only implies yesterday's existence* of the King of France but *does not presuppose* it.[15]

Note that here we utilised the singularity of the office of King of France, i.e., of the function of type $\iota_{\tau\omega}$. If the King of France does not exist in some world W at time T, the office is not occupied and the function does not have any value in W at T.Thus we did not have to explicitly specify the presupposition of (3) that the King exists using the schema (**). As explained above, due to partiality of the office constructed by $\lambda w\lambda t$ $[\,^0\mathit{King\_of}_{wt}\; {}^0\mathit{France}]$ and compositionality, (3*) and (4*) behave as desired.

Consider now another pair of sentences differing only by topic-focus articulation.

(5) *Our defeat* was caused by John.

(6) *John* caused our defeat.

While (5) not only implies but also presupposes that we were defeated, the truth-conditions of (6) are different, as our analysis clarifies.

First, (5) as well as (5')

(5') *Our defeat* was not caused by John.

entail (7):

(7) We were defeated.

[15] Using medieval terminology, we also say that the concept of the King of France occurs with *de re* supposition in (3) and (3'), and with *de dicto* supposition in (the τ-intensional context of) constructions (4) and (4').

These two sentences are about our defeat. Thus (7) is a presupposition of (5) and (5'), and the schematic logical form of (5) is the instruction "If we were *Defeated* then it was *Caused by John*, else *Fail.*" Simplifying a bit by ignoring the indexical character of 'we', let the proposition that we were defeated be constructed by ${}^0\textit{Defeat} \rightarrow o_{\tau\omega}$.[16]

In order to make use of the general schema (**), we need to specify P and S. The presupposition P that we were defeated is constructed by ${}^0\textit{Defeat}$ and S is that it was caused by John, $\lambda w \lambda t[\,{}^0\textit{Cause}_{wt}\,{}^0\textit{John}\;{}^0\textit{Defeat}]$.

Types: *Defeat*/$o_{\tau\omega}$; *Cause*/$(o\iota o_{\tau\omega})_{\tau\omega}$; *John*/$\iota$.

As a result, (5) expresses

(5*)
$\lambda w \lambda t\;{}^2[\iota\lambda c[[\,{}^0\textit{Defeat}_{wt} \supset [c = {}^0[\,{}^0\textit{Cause}_{wt}\,{}^0\textit{John}\,{}^0\textit{Defeat}]]] \wedge [\neg\,{}^0\textit{Defeat}_{wt} \supset {}^0\mathbf{F}]]]$.

The evaluation of the truth-conditions in any w, at any t thus follows these cases:

a) ${}^0\textit{Defeat}_{wt} \rightarrow_v \mathbf{T}$.
Then ${}^2\,{}^0[\,{}^0\textit{Cause}_{wt}\,{}^0\textit{John}\;{}^0\textit{Defeat}] = [\,{}^0\textit{Cause}_{wt}\,{}^0\textit{John}\;{}^0\textit{Defeat}]$;

b) ${}^0\textit{Defeat}_{wt} \rightarrow_v \mathbf{F}$.
Then ${}^2\,{}^0[\iota\lambda c\;{}^0\mathbf{F}] \rightarrow$ *Fails.*

On the other hand, the truth-conditions of (6) and (6') are different.

(6) *John* caused our defeat.

(6') *John* did not cause our defeat.

Now the sentence (6) is about the topic John, ascribing to him the property that he caused our defeat (focus). Thus the scenario of truly asserting (6') can be, for instance, this. Though it is true that John has a reputation of a rather bad player, Paul was in a very good shape and we won. Or, the other scenario is thinkable. We were defeated not because of John but because the whole team performed badly.

Hence, that we were defeated is not presupposed by (6), and the analyses of (6) and (6') are:

(6*) $\lambda w \lambda t\,[\,{}^0\textit{Cause}_{wt}\,{}^0\textit{John}\,{}^0\textit{Defeat}]$

(6'*) $\lambda w \lambda t\,\neg[\,{}^0\textit{Cause}_{wt}\,{}^0\textit{John}\,{}^0\textit{Defeat}]$

Yet, if (6) is true, then (7) can be validly inferred. In other words, (7) is entailed by (6) but not by (6'). This indicates that (7) is an allegation associated with (6) rather than a presupposition. As Hajičová says, the '(be)cause-clause' in focus triggers an allegation. To capture such truth-conditions, we need to refine the analysis. A plausible explication of this phenomenon is this: x is a cause of a proposition p iff p is true and if so then x affected p to be true. Schematically,

$$\lambda w \lambda t[\,{}^0\textit{Cause}_{wt}\,x\,p] = \lambda w \lambda t[p_{wt} \wedge [p_{wt} \supset [\,{}^0\textit{Affect}_{wt}\,x\,p]]].$$

[16] If we want to take into account the indexical character of these sentences, we use free variable '*we*' and obtain an *open* construction that constructs a proposition only after a valuation of *we* is supplied by a context of utterance. Thus (6) expresses $\lambda w \lambda t\,[\,{}^0\textit{Defeated}_{wt}\textit{we}]$. However, this is irrelevant here, as well as the past tense used in the example.

Types: *Cause*, *Affect*/$(o\alpha o_{\tau\omega})_{\tau\omega}$; $x \rightarrow \alpha$, α – any type; $p \rightarrow o_{\tau\omega}$.

If x is not a cause of p, then either p is not true or p is true but x did not affect p so that to be true:

$$\lambda w\lambda t\, \neg[\, {}^0\mathit{Cause}_{wt}\, x\, p] = \lambda w\lambda t[\neg p_{wt} \vee [p_{wt} \wedge \neg[\, {}^0\mathit{Affect}_{wt}\, x\, p]]].$$

Applying such an explication to (6), we get

$$(6^{**}) \qquad \lambda w\lambda t\, [\, {}^0\mathit{Defeat}_{wt} \wedge [\, {}^0\mathit{Defeat}_{wt} \supset [\, {}^0\mathit{Affect}_{wt}\, {}^0\mathit{John}\, {}^0\mathit{Defeat}]]],$$

which entails that we were defeated, $\lambda w\lambda t\, [\, {}^0\mathit{True}_{wt}\, {}^0\mathit{Defeat}]$, as it should be.

Similar phenomenon also crops up in case of seeking and finding. Imagine that one is referring on the tragedy in Dallas, November 22, 1963, by "The police were seeking the murderer of JFK but never found him". The sentence is ambiguous due to different topic-focus articulation.

(8) The *police* were seeking the murderer of JFK but never found him.

(9) The police were seeking *the murderer of JFK* but never found him.

The existence of the murderer of JFK is not presupposed by (8) unlike (9). The sentence (8) can be true in such states-of-affairs when JFK was not murdered, unlike the sentence (9). The latter can be reformulated in a more unambiguous way as "*The murderer of JFK* was looked for by the police but never found". This sentence expresses the construction

$$(9^{*}) \qquad \begin{array}{c} \lambda w\lambda t[[\, {}^0\mathit{Look\_for}_{wt}\, {}^0\mathit{Police}[\lambda w\lambda t\, [\, {}^0\mathit{Murder\_of}_{wt}\, {}^0\mathit{JFK}]]_{wt}] \wedge \\ \neg[\, {}^0\mathit{Find}^{L}_{wt}\, {}^0\mathit{Police}[\lambda w\lambda t\, [\, {}^0\mathit{Murder\_of}_{wt}\, {}^0\mathit{JFK}]]_{wt}]]. \end{array}$$

Types: *Look_for*, *Find*^{L}/$(o\iota\iota)_{\tau\omega}$; *Police*/$\iota$; *Murder_of*/$(\iota\iota)_{\tau\omega}$; *JFK*/$\iota$.[17]

On the other hand, the analysis of (8) relates police to the *office* of the murderer rather than to its holder. The police aim at finding who the murderer is. Thus we have *Seek*, *Find*^{S}/$(o\iota\iota_{\tau\omega})_{\tau\omega}$; and (8) expresses:

$$(8^{*}) \qquad \begin{array}{c} \lambda w\lambda t\, [[\, {}^0\mathit{Seek}_{wt}\, {}^0\mathit{Police}[\lambda w\lambda t[\, {}^0\mathit{Murder\_of}_{wt}\, {}^0\mathit{JFK}]]] \wedge \\ \neg[\, {}^0\mathit{Find}^{S}_{wt}\, {}^0\mathit{Police}[\lambda w\lambda t\, [\, {}^0\mathit{Murder\_of}_{wt}\, {}^0\mathit{JFK}]]]]. \end{array}$$

If the police did not find the murderer then either the murderer did not exist or the search was not successful. However, if the foregoing search was successful, then it is true that police found the murderer

$$\lambda w\lambda t\, [\, {}^0\mathit{Find}^{S}_{wt}\, {}^0\mathit{Police}\, [\lambda w\lambda t[\, {}^0\mathit{Murder\_of}_{wt}\, {}^0\mathit{JFK}]]]$$

and the murderer exists. Hence a successful search, i.e. *finding* after a foregoing search, also *triggers an allegated existence*

$$\frac{\lambda w\lambda t\, [\, {}^0\mathit{Find}^{S}_{wt}\, {}^0\mathit{Police}[\lambda w\lambda t[\, {}^0\mathit{Murder\_of}_{wt}\, {}^0\mathit{JFK}]]]}{\lambda w\lambda t\, [\, {}^0\mathit{Exist}_{wt}[\lambda w\lambda t[\, {}^0\mathit{Murder\_of}_{wt}\, {}^0\mathit{JFK}]]]}$$

where *Exist*/$(o\iota_{\tau\omega})_{\tau\omega}$ is the property of an individual office of being occupied. In order to render allegation, we explicate finding after a foregoing search in a similar way as the above causing ($x \rightarrow \iota$; $c \rightarrow \iota_{\tau\omega}$; *Success_Search*/$(o\iota\iota_{\tau\omega})_{\tau\omega}$):

[17] For the sake of simplicity, past tense and anaphora reference are ignored. For a more detailed analysis of this kind of seeking and finding, see, for instance, [3].

$\lambda w\lambda t[\ ^0\textit{Find}^S_{wt}\ x\ c] = \lambda w\lambda t[[\ ^0\textit{Exist}_{wt}c]\wedge[[\ ^0\textit{Exist}_{wt}c]\supset[\ ^0\textit{Success\_Search}_{wt}\ x\ c]]]$
$\lambda w\lambda t\neg[\ ^0\textit{Find}^S_{wt}\ x\ c] = \lambda w\lambda t[\neg[\ ^0\textit{Exist}_{wt}c]\vee[[\ ^0\textit{Exist}_{wt}c]\wedge\neg[\ ^0\textit{Success\_Search}_{wt}\ x\ c]]]$

The last example we want to adduce is again the ambiguity of a topic-focus.

(10) John *only* introduced *Bill* to Sue.

(11) John *only* introduced Bill to *Sue.*

Leaving aside possible disambiguation "John introduced only *Bill* to Sue" vs. "John introduced Bill only to *Sue*", (10) can be truly affirmed only in a situation when John did not introduce other people to Sue except for Bill; (10) says that *only Bill* (topic) was introduced by John to Sue (focus). This is not the case of (11). This sentence can be true in a situation when John introduced other people to Sue, but the only person Bill was introduced to by John was Sue.

Recalling the general schema of analysis of sentences with presupposition

$$\lambda w\lambda t\ ^2[\iota\lambda c\ [[P_{wt}\ \supset[c=\ ^0S_{wt}]]\wedge[\neg P_{wt}\ \supset\ ^0\mathbf{F}]]],$$

we have:

ad (10). Presupposition $P = \lambda w\lambda t\ [\forall x\ [[\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ x\ ^0\textit{Sue}]\supset[x=\ ^0\textit{Bill}]]]$

ad (11). Presupposition $P = \lambda w\lambda t\ [\forall y\ [[\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ ^0\textit{Bill}\ y]\supset[y=\ ^0\textit{Sue}]]]$

The construction C that is to be executed in case the presupposition is true is here

$$\lambda w\lambda t\ [\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ \ ^0\textit{Bill}\ \ ^0\textit{Sue}].$$

Types: *Int_to*/$(o\iota\iota\iota)_{\tau\omega}$ - *who* introduced *who* to *whom*; *John*, *Sue*, *Bill*/ι; $\forall/(o(o\iota))$.

The resulting analyses are

$$\begin{aligned}(10^*)\qquad &\lambda w\lambda t\ ^2[\iota\lambda c[[\forall x[[\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ x\ ^0\textit{Sue}]\supset[x=\ ^0\textit{Bill}]]\supset\\ &\quad[c=\ ^0[\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ ^0\textit{Bill}\ ^0\textit{Sue}]]]\wedge\\ &\quad[\exists x[[\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ x\ ^0\textit{Sue}]\wedge\neg[x=\ ^0\textit{Bill}]]\supset\ ^0\mathbf{F}]]];\end{aligned}$$

$$\begin{aligned}(11^*)\qquad &\lambda w\lambda t\ ^2[\iota\lambda c[[\forall y[[\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ ^0\textit{Bill}\ y]\supset[y=\ ^0\textit{Sue}]]\supset\\ &\quad[c=\ ^0[\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ ^0\textit{Bill}\ ^0\textit{Sue}]]]]\wedge\\ &\quad[\exists y[[\ ^0\textit{Int\_to}_{wt}\ ^0\textit{John}\ ^0\textit{Bill}\ y]\wedge\neg[y=\ ^0\textit{Sue}]]\supset\ ^0\mathbf{F}]]].\end{aligned}$$

Using a technical jargon, the truth conditions expressed by the construction (10*) are "If the only person that was introduced by John to Sue is Bill, then it is true that John introduced *only Bill* to Sue, otherwise undefined". Similarly for (11*).

4 Concluding Remarks

We demonstrated the semantic character of topic-focus articulation. This problem is connected with the ambiguity of natural language sentences. Logical analysis cannot disambiguate any sentence, because it presupposes full linguistic competence. Thus the input for our method is the output of a linguistic annotation providing labels for the topic-focus articulation. Yet, our fine-grained method can contribute to a language

disambiguation by making these hidden features explicit and logically tractable. In case there are more non-equivalent senses of a sentence we furnish the sentence with more than one different TIL constructions. Having a formal fine-grained encoding of a sense, we can then automatically infer the relevant consequences. Thus in our opinion theoretical linguistics and logic must collaborate and work hand in hand.

Using the expressive logical system of TIL, we were able to provide rigorous analyses such that sentences differing only in the topic-focus articulation are assigned different constructions producing different propositions and implying different consequences. We analysed the phenomena of presupposition connected with a topic and allegation triggered by a focus so that relevant consequences can be formally derived. Thus, in principle, an inference machine can be built on the basis of TIL analysis such that it neither over-infers (by inferring something that does not follow from the assumptions) nor under-infers (by not being able to infer something that does follow). Currently we develop a computational variant of TIL, the *TIL-Script* functional programming language. TIL constructions are encoded by natural-language expressions in a near-isomorphic manner and for the needs of real-world human agents *TIL-Script* messages are presented in a standardised natural language. *Vice versa*, humans can formulate their requests, queries, etc., in the standardised natural language that is transformed into *TIL-Script* messages. Thus the provision of services to humans can be realised in a form close to human understanding. From the theoretical point of view, the inference machine for TIL has been specified. However, its full implementation is still work in progress.

The direction of further research is clear. We are going to develop the *TIL-Script* language in its full power, and examine other complex features of natural language. Yet the clarity of this direction does not imply its triviality. The complexity of the work going into building a procedural theory of language is almost certain to guarantee that complications we are currently unaware of will crop up. Yet we are convinced that if any logic can serve to solve such problems, then it must be a logic with hyperintensional (most probably procedural) semantics, such as TIL.

References

1. Duží, M.: Topic-Focus Articulation from the Semantic Point of View. In *CICLing 2009*. Ed. Gelbukh Alexander, Berlin Heidelberg: Springer-Verlag LNCS, vol. 5449 (2009), 220-232.
2. Duží, M. and P. Materna: 'Logical form', in: *Essays on the Foundations of Mathematics and Logic*, vol. 1, G. Sica (ed.), Monza: Polimetrica International Scientific Publisher (2005), pp. 115-53.
3. Duží, M.: 'TIL as the Logic of Communication in a Multi-Agent System'. In *Research in Computing Science*, vol. 33 (2008), pp. 27-40.
4. Duží, M., B. Jespersen, and J. Müller: 'Epistemic closure and inferable knowledge', in: *The Logica Yearbook 2004*, L. Běhounek and M. Bílková (eds.), Czech Academy of Science, Prague: Filosofia (2005), pp. 125-140.
5. Hajičová, E.: What we are talking about and what we are saying about it. In *Computational Linguistics and Intelligent Text Processing*, LNCS Springer Berlin /Heidelberg, vol. 4919/2008, pp. 241-262.
6. Jespersen, B.: 'Predication and extensionalization'. *Journal of Philosophical Logic*, vol. 37, No. 5, Springer Netherlands (2008), pp. 479 – 499.
7. Materna, P.: *Conceptual Systems.* Logos Verlag, Berlin (2004)

8. Materna, P. and Duží M.: 'The Parmenides principle', *Philosophia*, philosophical quarterly Israel, vol. 32 (2005), pp. 155-80.
9. Rescher, N.: *Epistemic Logic.* Pittsburgh: University of Pittsburgh Press (2005).
10. Sandu, G., Hintikka, J.: 'Aspects of compositionality', *Journal of Logic, Language, Information 10* (2001) 49-61.
11. Strawson, P. *Introduction to Logical Theory.* London: Methuen (1952).
12. Tichý, P.: *The Foundations of Frege's Logic*, Berlin, New York: De Gruyter (1988).
13. Tichý, P.: *Collected Papers in Logic and Philosophy*, V. Svoboda, B. Jespersen, C. Cheyne (eds.), Prague: Filosofia, Czech Academy of Sciences, and Dunedin: University of Otago Press (2004).
14. Svoboda, A. – Materna, P. (1987): Functional sentence perspective and intensional logic. In: Functionalism in Linguistics (Dirven, R. – Fried, V. eds.). Amsterdam/Philadelphia: John Benjamins, pp. 191-205. ISBN 90-272-1524-3

The JOS Language Resources: Towards Standardised and Available HLT Datasets for Slovene

Tomaž Erjavec

Department of Knowldege Technologies
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract. Linguistically annotated corpora are the basis for human language technology research but are, for a number of languages, still difficult to obtain, especially as complete datasets. Essential resources are validated part-of-speech or, better, morphosyntactically tagged corpora necessary for training taggers, themselves a basic infrastructure for more advanced HLT tasks. The paper presents the language resources for Slovene developed in the JOS project, http://nl.ijs.si/jos. We present the JOS MULTEXT-East-based morphosyntactic specifications, which define the rich tagset necessary for describing morphologically complex Slovene word-forms. The paper discusses the annotated corpora and the on-line tool available for annotation of Slovene texts. We present the steps that led to the creation of these resources, their Text Encoding Initiative compliant XML encoding, their availability under the Creative Commons licences, and sketch on-going work in the areas of syntactic and semantic annotation.

1 Introduction

Linguistically annotated corpora are the basis for human language technology research as well as corpus linguistics, but are still difficult to obtain for a number of languages, esp. as complete datasets. Annotated corpora are useful as training datasets for machine learning or statistics based tools. Here, an essential resource is a hand-validated part-of-speech or, better, morphosyntactically tagged corpus, necessary for training taggers, themselves a basic infrastructure for more advanced HLT tasks.

For Slovene the MULTEXT-East[1] resources [1] have so far contained the only available manually validated tagged corpus, the Slovene translation of the novel "1984" by G. Orwell. And while the MULTEXT-East tagset and corpus encoding practices have been adopted for a number of Slovene corpora, the "1984" corpus itself is nevertheless small (100,000 words) and contains only one translated novel, resulting in very brittle tagging models. Furthermore, the years of using the Slovene MULTEXT-East tagset have shown that it could do with several modifications.

The JOS Slovene language resources [2] attempt to bridge this gap by producing standardised and freely available linguistically annotated corpora and associated resources.

[1] http://nl.ijs.si/ME/

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 17–34, 2009.

While the syntactic and semantic annotation are still on-going, the JOS morphosyntactic specifications for Slovene and two carefully composed morphosyntactically annotated corpora are already available. These resources have been used to (re)train a tagger and lemmatiser for Slovene, giving much better accuracies than before. The on-line tool as well as all the other resources are available from the JOS homepage under Creative Commons licences.

The rest of the paper is structured as follows: Section 2 introduces the JOS morphosyntactic specifications and tagset; Section 3 describes the composition and encoding of the JOS corpora; Section 4 explains the availability of the JOS resources; and Section 5 gives conclusions and plans for further work.

2 JOS morphosyntactic specifications

The purpose of the JOS morphosyntactic specifications[2] [3,4] is to provide a well-documented and accessbile feature-based tagset, appropriate for word-level syntactic tagging of Slovene language corpora and texts.

The JOS specifications are a modification of the Slovene part of the multilingual MULTEXT-East Version 3 [1] morphosyntactic specifications, which are used in the annotation of a number of Slovene corpora, most notably the reference corpus FidaPLUS[3] [5], and, of course, the MULTEXT-East Slovene corpus. The impetus for the modifications came from linguistic grounds, but in adapting linguistic aspects of the specifications, many technical changes were also introduced, the most important one being the switch to XML. Namely, in MULTEXT-East Version 3, the morphosyntactic specifications were written in LaTeX while for the JOS specifications we converted this encoding to XML. The XML-based JOS specifications then served as a template in the development of the MULTEXT-East morphosyntactic specifications Version 4 [6]; these are, at the time of writing, still work in progress.

The core of the JOS morphosyntactic specifications is the definition of part-of-speech categories for Slovene in terms of their attributes and their values. They also define the mapping from these values into a position-based compact string encoding, the morphosyntactic descriptions (MSDs), and list all valid MSDs for Slovene, i.e., they define the MSD tagset used for tagging of Slovene corpora. So, for example, the specifications state that `Noun, Type = common, Gender = masculine, Number = singular, Case = accusative, Animate = no` maps to the MSD `Ncmsan` and that this is a valid MSD for Slovene. The specifications also provide some commentary, and corpus examples for each MSD. They are written both in Slovene and English, so that the commentary, MSDs and feature-value combinations can be expressed in either language.

The JOS morphosyntactic specifications [3] consist of the following parts:

1. Background: This part gives a short introduction
2. Definitions of Morphosyntactic Categories
   This part has 12 sections, one for each of the 12 MULTEXT-East defined categories (parts-of-speech). Each section contains a table defining the attributes and their values for the category, with notes, c.f. Section 2.2.

[2] `http://nl.ijs.si/jos/josMSD-en.html`, `http://nl.ijs.si/jos/josMSD-sl.html`
[3] `http://www.fidaplus.net/`

3. Lists: Here, the important list is the one with morphosyntactic descriptions, which contains the full MSD tagset with examples of usage, c.f. Section 2.3.
4. Differences between MULTEXT-East Version 3 and JOS specifications: This part enumerates the changes in attributes, their positions, values and their codes between the two specifications. This section is written in Slovene only.
5. Comparison with other recommendations for morphosyntactic annotation: This part contains the working notes (in Slovene only) that were produced in the course of preparing the JOS morphosyntactic specifications. The notes give a comparison between various proposals for morphosyntactic specifications / tagsets and recommendations for the JOS specifications; c.f. Section 2.5.

The specifications are written in XML, in a TEI P5 schema, c.f. Section 2.1. Several other formats are produced with XSLT stylesheets from this source, i.e., HTML in Slovene and English, tabular files giving conversions between MSDs and feature-sets, and XML libraries for use in corpora; c.f. Section 2.4.

2.1 The format of the specifications

The Text Encoding Initiative[4] is an international consortium whose primary function is to maintain the TEI Guidelines, which set out a vocabulary of elements useful for describing text for scholarly purposes. The Guidelines use XML encoding and are written as a set of XML schemas (element grammars) with accompanying documentation.

The are a number of advantages of using TEI for encoding. TEI documents are written in XML, which brings with it validation of document structure, a wealth of supporting software and related standards. The most important one is the XML transformation language, XSLT, which allows writing scripts (stylesheets) that transform XML documents into other, differently structured (XML, HTML, text) documents. The XSLT standard is nowadays generally supported, e.g., we find it implemented in most Web browsers. The JOS specifications come with a number of XSLT transforms, which help in authoring or displaying the specifications; they are further discussed in Section 2.4.

TEI is also general enough to encode the non-normative parts of the specifications, e.g., the introduction, notes, etc. The TEI also provides, amongst other software, a sophisticated set of XSLT stylesheets and associated components for converting TEI documents into HTML and PDF. These `tei-xsl` stylesheets, developed by Sebastian Rahtz and freely available via the TEI homepage, cover a large number of TEI elements, and also perform tasks such as generating the table of contents, splitting (large) TEI documents into several HTML files (while preserving cross-links), giving each HTML a project defined header and footer, etc.

Additionally, we wrote a TEI to HTML stylesheet for the TEI header (not covered by `tei-xsl`), where the TEI header contains the meta-data about a TEI document. The resulting HTML file, localised to Slovene, is illustrated in Figure 1. Such a readable view of the TEI header makes the meta-data about the corpus more understandable and accessible, especially to speakers of Slovene.

[4] `http://www.tei-c.org/`

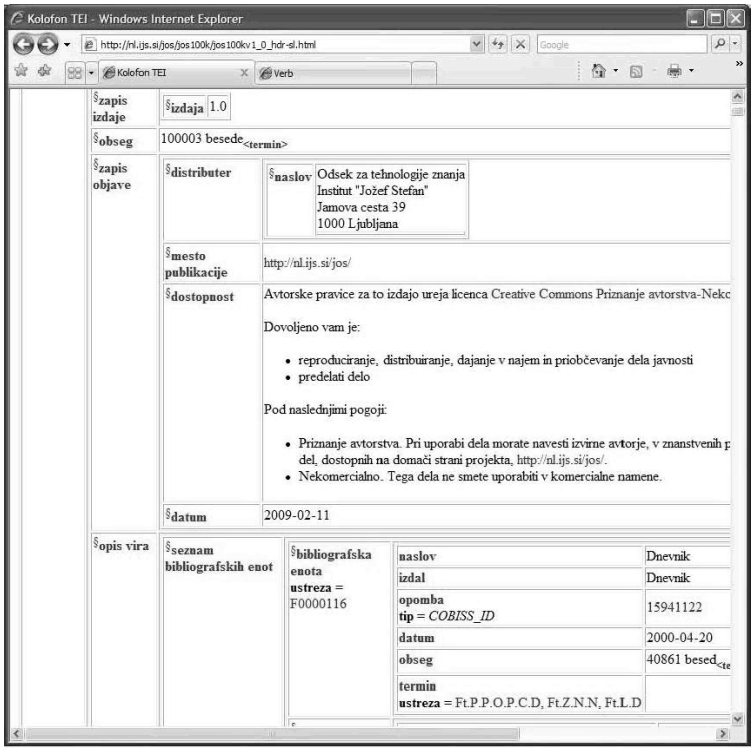


Fig. 1. Part of the jos100k corpus TEI header in HTML. Each element in the TEI header is given a gloss, in this example translated to Slovene. These glosses are also linked to their definitions in the TEI guidelines.

Another reason for using TEI is uniformity, as the JOS annotated corpora are also encoded in TEI; encoding the specifications in TEI as well gives an easy way to directly integrate the corpus with the specifications, leading to simple validation of the corpus annotations or conversion between corpus MSDs and their feature-structure representations. This can be useful for querying the corpus, as it enables, e.g., the selection of word tokens based on particular features.

The JOS specifications are written in TEI P5, the most recent version of the TEI Guidelines [7]. They use a particular TEI P5 schema called teiLite, which gives a basic element vocabulary for texts. The specifications are encoded as a text with divisions, with the formal parts encoded as tables.

2.2 Definitions of Morphosyntactic Categories

The formal core of the specifications are the tables defining attributes and their values for each of the 12 categories - c.f. Figure 2 for an example. The tables also give the mapping between the features and MSDs, by giving the position of the attribute in the MSD string, and for each attribute-value pair, a one letter code for the MSD string.

Furthermore, the attribute and value names and codes are given both in Slovene and English, and these translations (localisations) enable shifting between the two languages. For example, with the specifications and an accompanying XSLT stylesheet it is possible to translate the English `Ncmsan` to Slovene `Sometn`, and expand it either to its English or Slovene feature-structure, the latter being `samostalnik, vrsta = občno_ime, spol = moški, število = ednina, sklon = tožilnik, živost = ne`. This makes it possible for Slovene speakers to use the annotations in their native language, while also allowing English speakers to understand them.

A synopsis of the twelve JOS categories, their attributes and the number of values they can take is given in Table 1.

Table 1. Slovene JOS categories, their attributres and the number of their values.

Category	Attributes with number of values
Noun	Type(2), Gender(3), Number(3), Case(6), Animacy(2)
Verb	Type(2), Aspect(3), Form(7), Person(3), Number(3), Gender(3), Negative(2)
Adjective	Type(3) Degree(3), Gender(3), Number(3), Case(6), Definiteness(2)
Adverb	Degree(3), Participle(2)
Pronoun	Type(9), Person(3), Gender(3), Number(3), Case(6), Owner_Number(3), Owner_Gender(3), Form(2)
Numeral	Form(3), Type(4), Gender(3), Number(3), Case(6), Definiteness(2)
Preposition	Case(6)
Conjunction	Type(2)
Particle	no attributes
Interjection	no attributes
Abbreviation	no attributes
Residual	Type(3)

2.3 The MSD list

The specifications also include the list of valid MSDs for Slovene, totaling 1,902. Each MSD is given with its expansion into a feature-structure and translation to English, c.f. Figure 3 for the HTML rendering of this list. Additionally, the number of word tokens and word types tagged with this MSD in the 1 million word jos1M corpus (c.f. Section 3) is given, and examples of the usage of the MSD.

```
<div type="section" xml:id="msd.N">
 <head xml:lang="sl">Samostalnik</head>
 <head xml:lang="en">Noun</head>
 <table n="msd.cat" xml:id="msd.cat.N">
  <head xml:lang="sl">Tabela atributov in vrednosti za samostalnik</head>
  <head xml:lang="en">Attribute-value table for Noun</head>
  <row role="type">
   <cell role="position">0</cell>
   <cell role="name" xml:lang="sl">samostalnik</cell>
   <cell role="code" xml:lang="sl">S</cell>
   <cell role="name" xml:lang="en">Noun</cell>
   <cell role="code" xml:lang="en">N</cell>
  </row>
  <row role="attribute">
   <cell role="position">1</cell>
   <cell role="name" xml:lang="sl">vrsta</cell>
   <cell role="name" xml:lang="en">Type</cell>
   <cell role="values">
    <table>
     <row role="value">
      <cell role="name" xml:lang="sl">občno_ime</cell>
      <cell role="code" xml:lang="sl">o</cell>
      <cell role="name" xml:lang="en">common</cell>
      <cell role="code" xml:lang="en">c</cell>
     </row>
     <row role="value">
      <cell role="name" xml:lang="sl">lastno_ime</cell>
      <cell role="code" xml:lang="sl">l</cell>
      <cell role="name" xml:lang="en">proper</cell>
      <cell role="code" xml:lang="en">p</cell>
     </row>
    </table>
   </cell>
  </row>
  <row role="attribute">
    <cell role="position">2</cell>
    <cell role="name" xml:lang="sl">spol</cell>
    <cell role="name" xml:lang="en">Gender</cell>
    <cell role="values">
      <table>
        <row role="value">
          <cell role="name" xml:lang="sl">moški</cell>
          <cell role="code" xml:lang="sl">m</cell>
          <cell role="name" xml:lang="en">masculine</cell>
          <cell role="code" xml:lang="en">m</cell>
        </row>
        ...
```

Fig. 2. JOS morphosyntactic specifications: start of table for Noun. The first row to the table gives its type, i.e., the category, and the succeeding ones its attributes. The type row gives the name of the category and its code in Slovene and English. Each attribute row gives the position of the attribute in the MSD string and its name in Slovene and English. Furthermore, it contains the possible values of the attribute, with each given its name and one-character code for the MSD string in Slovene and English.

The examples were automatically extracted from (1) the jos1M corpus, (2) the lexicon of closed class words and (3) the lexicon derived from the FidaPLUS corpus. The examples are ordered by the number of occurrences in (1), followed by examples from (2) and (3). It should be noted that both (1) and (3) contain some tagging and lemmatisation errors so not all examples, esp. those of low-frequency words, are necessarily correct. This rather complicated scheme is necessary due to the Zipfian distribution of the MSDs. So, the jos100k corpus, even though it contains 100,000 words, uses only 1,064, or just over half, of all possible MSDs, and certain MSDs are exceedingly rare, not occuring even in the 600 million word FidaPLUS corpus. It was thus necessary to combine several sources to arrive at the complete and exemplified set of MSD.

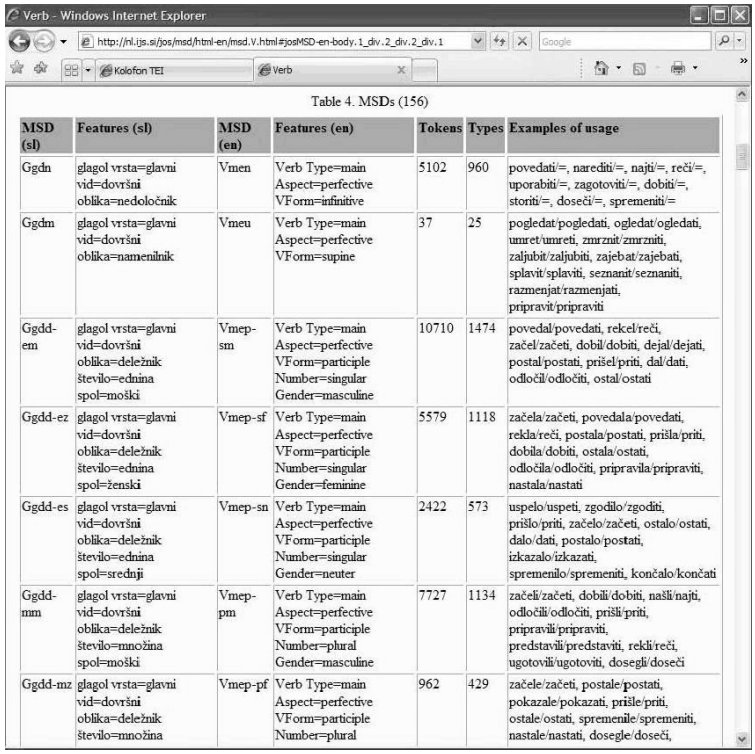


Table 4. MSDs (156)

MSD (sl)	Features (sl)	MSD (en)	Features (en)	Tokens	Types	Examples of usage
Ggdn	glagol vrsta=glavni vid=dovršni oblika=nedoločnik	Vmen	Verb Type=main Aspect=perfective VForm=infinitive	5102	960	povedati/=, narediti/=, najti/=, reči/=, uporabiti/=, zagotoviti/=, dobiti/=, storiti/=, doseči/=, spremeniti/=
Ggdm	glagol vrsta=glavni vid=dovršni oblika=namenilnik	Vmeu	Verb Type=main Aspect=perfective VForm=supine	37	25	pogledat/pogledati, ogledat/ogledati, umret/umreti, zmrznit/zmrzniti, zaljubit/zaljubiti, zajebat/zajebati, splavit/splaviti, seznanit/seznaniti, razmenjat/razmenjati, pripravit/pripraviti
Ggdd-em	glagol vrsta=glavni vid=dovršni oblika=deležnik število=ednina spol=moški	Vmep-sm	Verb Type=main Aspect=perfective VForm=participle Number=singular Gender=masculine	10710	1474	povedal/povedati, rekel/reči, začel/začeti, dobil/dobiti, dejal/dejati, postal/postati, prišel/priti, dal/dati, odločil/odločiti, ostal/ostati
Ggdd-ez	glagol vrsta=glavni vid=dovršni oblika=deležnik število=ednina spol=ženski	Vmep-sf	Verb Type=main Aspect=perfective VForm=participle Number=singular Gender=feminine	5579	1118	začela/začeti, povedala/povedati, rekla/reči, postala/postati, prišla/priti, dobila/dobiti, ostala/ostati, odločila/odločiti, pripravila/pripraviti, nastala/nastati
Ggdd-es	glagol vrsta=glavni vid=dovršni oblika=deležnik število=ednina spol=srednji	Vmep-sn	Verb Type=main Aspect=perfective VForm=participle Number=singular Gender=neuter	2422	573	uspelo/uspeti, zgodilo/zgoditi, prišlo/priti, začelo/začeti, ostalo/ostati, dalo/dati, postalo/postati, izkazalo/izkazati, spremenilo/spremeniti, končalo/končati
Ggdd-mm	glagol vrsta=glavni vid=dovršni oblika=deležnik število=množina spol=moški	Vmep-pm	Verb Type=main Aspect=perfective VForm=participle Number=plural Gender=masculine	7727	1134	začeli/začeti, dobili/dobiti, našli/najti, odločili/odločiti, prišli/priti, pripravili/pripraviti, predstavili/predstaviti, rekli/reči, ugotovili/ugotoviti, dosegli/doseči
Ggdd-mz	glagol vrsta=glavni vid=dovršni oblika=deležnik število=množina	Vmep-pf	Verb Type=main Aspect=perfective VForm=participle Number=plural	962	429	začele/začeti, postale/postati, pokazale/pokazati, prišle/priti, ostale/ostati, spremenile/spremeniti, nastale/nastati, dosegle/doseči,

Fig. 3. The start of the JOS MSD list for Verb, shown in HTML. Each row gives the MSD and its expansion to features in Slovene and English. Additionally, the number of word tokens and word types tagged with this MSD in the 1 million word jos1M corpus is given, and up to 10 examples of the usage of the MSD in the form word-form/lemma. Where the word-form and lemma are identical, lemma is written as an equal sign.

2.4 Conversions with XSLT

An important part of the XML specifications are the associated XSLT stylesheets, which allow for various transformations of the specifications. The stylesheets are written in XLST V1.0 and are documented with XSLTdoc.[5] They take the specifications as input, usually together with certain command line arguments, and produce either XML, HTML or (tabular) text output, depending on the stylesheet.

For easier reading, we produced the browser version of the specifications as a set of linked HTML pages, with a table of contents and automatically generated indexes (lists). The XSLT stylesheets to produce the HTML of the specifications heavily depend on the standard `tei-xsl` stylesheets. Conversion is currently only supported into HTML, although PDF should not be too difficult to implement via `tei-xsl`. Conversion to HTML is done in three steps:

`msd-spec2prn.xsl` generates from the specifications a display ("print") oriented teiLite document; this means making display-oriented tables and generating the indexes of attributes, values, and MSDs;

`teiHeader2html.xsl` converts the TEI header into HTML, and possibly localises the element name glosses; for expansion of element names to glosses and localisation to Slovene it uses a subsidiary file, `teiLocalise-sl.xml`;

`jos-prn2html.xsl` is a driver file, which calls `tei-xsl`; it takes as input the display-oriented document and produces a set of linked HTML pages, one for each part or section; it also links the HTML TEI header with the HTML of the specifications.

In addition to the XML and HTML the distribution also includes tabular files, which give the conversions of the MSDs into various representations, e.g. localisations, expansions to feature-structures, etc. There are two stylesheets for MSD conversion, which take a list of MSDs as a parameter and, on the basis of the JOS specifications, typically convert to some other representation:

`msd-expand.xsl` produces different types of output, depending on the values of its `mode` parameter. It also takes as parameters the required input and output localisations. The output is in plain text tabular format, with columns that can be, depending on the value of `mode`, which is a space separated list of modes, the following:

- `check` only checks the validity of the input MSDs, flagging codes that are illegal — this mode does not combine with the other ones;
- `id` identity transform (with possible localisation);
- `collate` collating sequence, with which it is possible to sort MSDs so that their order corresponds to the ordering of categories, attributes and their values in the specifications;
- `brief` expansion to values only, which is the most compact feature-expanded format and is meant for short but still readable expansions of MSD; instead of binary values (yes/no), +/-Attribute is written;
- `verbose` expansion to feature-structures (attribute=value pairs) for all attributes defined for the category of the MSD;

[5] `http://www.pnp-software.com/XSLTdoc/`

`canonical` expansion to feature-structures (attribute=value pairs) for all defined attributes, regardless of whether they are defined for a particular category or not;

`msd-fslib.xsl` transforms the MSD list into a XML/TEI feature and feature-structure libraries, suitable for inclusion into MSD annotated and TEI encoded corpora.

The above two stylesheets are not meant to be run whenever a transformation is needed but rather to run them, once the specifications are finished, over the complete set of MSDs to produce the tabular and XML files. Such conversion tables are then made available together with the specifications.

As we consider the conversion tables included in the distribution very useful for operationalising the MSDs, we introduce them in detail:

`josMSD.tbl` Full list of MSDs, with the first column giving the collating sequence, the second the MSD in Slovene, and the third the MSD in English. The table is useful for sorting MSDs and for translating between the Slovene and English MSDs.
Example:
```
02V02000100000000 Gp-n Va-n
02V02000200000000 Gp-m Va-u
02V02000300010100 Gp-d-em Va-p-sm
```

`josMSD-val-sl.tbl`, `josMSD-val-en.tbl` MSDs with short expansions to feature values. This is the shortest human readable form of MSDs.
Example -sl: `Gp-n glagol pomožni nedoločnik`
Example -en: `Va-n Verb auxiliary infinitive`

`josMSD-attval-sl.tbl`, `josMSD-attval-en.tbl` MSDs with attribute=value expansions for all attributes defined for PoS. This is a mapping of MSDs to feature-structures where categories are treated as types, i.e., all the attributes defined for a category are listed, even if undefined.
Example -sl: `Gp-n glagol vrsta=pomožni vid=0 oblika=nedoločnik oseba=0 število=0 spol=0 nikalnost=0`
Example -en: `Va-u Verb Type=auxiliary Aspect=0 VForm=supine Person=0 Number=0 Gender=0 Negative=0`

`josMSD-canon-sl.tbl`, `josMSD-canon-en.tbl` MSDs with attribute=value expansions for all defined attributes. This is the mapping of MSDs to the canonical representation of feature-structures, where all the 14 defined attributes, regardless of category are listed. This representation is used for a positional representation of features.
Example -sl: `Gp-n glagol vrsta=pomožni spol=0 število=0 sklon=0 živost=0 vid=0 oblika=nedoločnik oseba=0 nikalnost=0 stopnja=0 določnost=0 število_svojine=0 spol_svojine=0 zapis=0`

`josMSD-lib-sl.xml`, `josMSD-lib-en.xml` MSDs and features expressed as TEI XML feature-structure libraries.
Example -sl:

```
<fs xml:id="Gp-n" xml:lang="sl" feats="#G0. #G1.p #G3.n"/>
```

```
    ...
<f name="besedna_vrsta" xml:id="G0." xml:lang="sl"><symbol
    value="glagol"/></f>
<f name="vrsta" xml:id="G1.p" xml:lang="sl"><symbol
    value="pomožni"/></f>
<f name="oblika" xml:id="G3.n" xml:lang="sl"><symbol
    value="nedoločnik"/></f>
```

2.5 Comparison to other tagging schemes

In moving from MULTEXT-East V3 specifications to JOS, a number of other morphosyntactic specifications (tagging schemes) were examined, concetrating on those for Slovene, Czech, and English. The MULTEXT-East V3 specifications were, of course, looked at closely, as this was the starting point. There are two other morphosyntactic tagsets for Slovene. The SIlex tagset [8] is a rather close copy of the MULTEXT-East one, with minor differences. The tagset used in the "POS-beseda" corpus [9,10] differs from the MULTEXT-East, JOS or SIlex ones in its fundamental design, as it does not use positional attributes and is very closely tied to traditional Slovene grammars. Czech, as a language similar to Slovene, and with well developed language resources was also of great interest. The comparison included the Ajka [11] tagset[6] and the Prague tagset [12] used e.g., in the Czech National Corpus[7] and Prague Dependency Treebank.[8] Finally, the comparison included two English tagsets, as one of the first and still most widely used: the CLAWS[9] and BNC[10] tagsets.

A detailed exposition of the properties of the various tagsets and their relation to JOS organised by category is given in Appendix B of the JOS specifications [3]. Unfortunately, the comparison is written in Slovene and remains a working draft with much detail but somewhat lacking in synthesis — there are a number of interesting issues that arise when comparing this number of proposals for morphosyntactic annotation which would deserve a more rounded exposition.

3 The JOS corpora

In this section we briefly present the two corpora developed in the JOS project, jos100k, the gold-standard 100,000 word corpus and jos1M, the 1 million word JOS corpus. A more detailed exposition of is given in [2].

We first introduce the basis for the corpora, the FidaPLUS corpus and then discuss their annotation with MSDs and lemmas. FidaPLUS [5] is a reference corpus of modern-day Slovene which contains about 600 million words, is encoded in (near) SGML following the Text Encoding Initiative Guidelines, TEI P3 [13], and is annotated with automatically assigned context disambiguated Slovene MULTEXT-East Version

[6] http://nlp.fi.muni.cz/projekty/ajka/
[7] http://ucnk.ff.cuni.cz/
[8] http://ufal.mff.cuni.cz/pdt/
[9] http://ucrel.lancs.ac.uk/claws/
[10] http://www.natcorp.ox.ac.uk/

3 [1] morphosyntactic descriptions (MSDs) and lemmas. The entire text processing chain, including up-conversion from source formats, tokenisation, lexical processing and disambiguation was performed with the proprietary software by the Slovene HLT company Amebis.[11]

FidaPLUS is freely available for research via a Web concordancer, and is a very useful tool for research into Slovene language. But, outside FIDA/FidaPLUS projects partner institutions, it is not available as a dataset and so cannot serve as the basis for HLT-type research. As a training set for PoS taggers it also suffers from the drawback that it was tagged fully automatically; and while the Amebis tagger gives state-of-the-art performance for Slovene, nevertheless the corpus contains annotations errors for about 15% of the words. FidaPLUS does, however, offer an excellent basis on which to develop a corpus for HLT research. The first step to arrive at the JOS corpora was to convert FidaPLUS to TEI P5 XML [7] in order to maintain a standard format and to enable processing with XML tools, in particular XSLT; we call this XMLified version of FidaPLUS FIDA+X.

3.1 Sampling FIDA+X

The content of jos100k and jos1M corpora was obtained from the 600M word FIDA+X by a two stage filter and sampling procedure meant to help JOS corpora achieve the following characteristics:

- Are representative and balanced.
  The representativeness of JOS corpora follows from this attribute holding for FIDA+X. The balance of FIDA+X is, however, more questionable, as it contains a large percentage of newspaper texts, and a relatively small one of fiction and esp. professional writing (technical, academic prose). Simply adopting the balance of FidaPLUS would leave the much smaller JOS corpora with very small amounts of such texts.
- Consist of clean text.
  Given that the corpora will be linguistically annotated, it is worth ensuring that the corpus contains only legitimate text paragraphs and tokens: FidaPLUS contains duplicates and cases where the up-conversion produced very short texts, paragraphs or sentences, or did not remove all formatting information.
- Do not infringe copyright.
  While complete text might be preferable for certain types of analysis, this would be questionable for copyright reasons, but short samples from the texts are not problematic. A sampling procedure has the further advantage that it makes the corpus more varied.

The first sampling step randomly selected complete texts from the Fida+X corpus, but excluded anomalous texts and prefered certain text types to others. First, a filter discarded texts that are too short or too long, have too much formatting or are badly formed according to various other heuristics. Second, ponders were given to text types and other metadata, so that, in short, the bias of FidaPLUS towards newspapers is somewhat counteracted, mostly towards technical writing.

[11] `http://www.amebis.si/`

Taking these texts as the input, the second step selected random paragraphs, again subject to some constraints: the minimum and maximum size of the paragraph, and that each paragraph is unique in the corpus: duplicates were discarded by using CRC.

These two steps were run twice, with different settings. For jos100k, the first step was set to produce a 10M word corpus, and for jos1M a 100M word one, i.e., for both corpora on average 1% of paragraphs was selected from the texts.

Table 2 gives an overview of the sizes of the two JOS corpora, in terms of the number of texts, paragraphs, sentences, words and (word and punctuation) tokens.

Table 2. Tagcount of JOS corpora.

Corpus	jos100k	jos1M
<text>	249	2,565
<p>	1,599	15,758
<s>	6,151	60,291
<w>	100,003	1,000,019
<w>+<c>	118,394	1,182,945

3.2 Morphosyntactic annotation

Manual annotation, performed by a supervised team of undergraduate students, consisted of correcting the MSDs and lemmas in the two JOS corpora, where the base-line annotations were mapped to the JOS specifications from the Fida+X/MULTEXT-East MSDs and lemmas, automatically assigned by Amebis.

Technically, the manual annotation proceeded via a Web interface, which, for a given corpus and input parameters (e.g., regexp over word-form, lemma or MSD), generated Excel spreadsheets for the annotators. The spreadsheets feature a title sheet, a sheet with the text and annotations to be corrected (via drop-down menus), and guidelines. Upon correction, the spreadsheets were uploaded and the corpus updated with the new manual annotations. This overall architecture was originally developed for correcting and annotating historical texts [14], and was then successfully applied in the JOS project as well.

The annotation process was cyclical, with a mixture of manual and automatic annotation steps, depending on the corpus in question.

The annotation of the 100k corpus was carried out in parallel with developing the JOS morphosyntactic specifications, tagset and its lexical mapping, along with the guidelines for annotators. This made the process rather complicated but resulted in specifications, tagset, lexicon and corpus which are consistent and made the jos100k corpus as free of errors as possible, given project constraints. The complete corpus was validated twice by different annotators, and the words where the two manual annotations differed were validated for the third time. When further mistakes were spotted, annotations of certain token types in the corpus were unified or corrected in several subsequent steps to arrive at the gold standard jos100k manually annotated corpus.

As project resources did not allow for manual verification of the complete one million word jos1M, we manually validated only "suspicious" MSDs, as well as automatically improving the Fida+X annotations. We first trained the TnT tagger [15] on the manually annotated jos100k, and gave it, as the backup lexicon, the lexicon extracted from Fida+X with its MSD converted to the JOS specification. The word tokens where the TnT and Amebis differed in their MSD assignment (about 19%, i.e., 190,000 words) were labelled as suspicious, and were manually validated, arriving at an estimated overall 96.8% word accuracy of MSD tagging; details on this annotation are given in [2].

Figure 4 gives an example stretch from the jos1M corpus, exemplifying its XML/TEI structure.

```
<div xml:id="F0000015" n="8 35 709 814">
 <p xml:id="F0000015.13" n="4 117 133">
  <s xml:id="F0000015.13.1" n="31 36">
   <w xml:id="F0000015.13.1.1" type="auto" msd="Sosei" lemma="postopanje">
        Postopanje</w><S/>
   <w xml:id="F0000015.13.1.2" type="auto" msd="Do" lemma="pred">pred</w><S/>
   <w xml:id="F0000015.13.1.3" type="auto" msd="Ppnseo" lemma="afganistanski">
        afganistanskim</w><S/>
   <w xml:id="F0000015.13.1.4" type="auto" msd="Soseo" lemma="veleposlaništvo">
        veleposlaništvom</w>
   <c xml:id="F0000015.13.1.5">,</c><S/>
   <w xml:id="F0000015.13.1.6" type="manual" msd="Sosei" lemma="širjenje">
        širjenje</w><S/>
   <w xml:id="F0000015.13.1.7" type="auto" msd="Sozmr" lemma="govorica">
        govoric</w><S/>
   <w xml:id="F0000015.13.1.8" type="auto" msd="Dm" lemma="o">o</w><S/>
   <w xml:id="F0000015.13.1.9" type="manual" msd="Zk-sem" lemma="ta">tem</w>
   <c xml:id="F0000015.13.1.10">,</c>
   ...
```

Fig. 4. A stretch of text from the jos1M corpus. The IDs refer to FidaPLUS text identifiers, the bibliographic and taxonomic data of which is stored in the TEI header of the corpus. The n attribute summarises the number of paragraphs, sentences, words and tokens in each element, while the type attribute shows whether the annotation was automatic (where Amebis and TnT taggers agreed on the annotation) or manual. The S elements indicates whitespace between tokens.

3.3 Syntactic and semantic annotation

While the morphosyntactic annotation is finished, syntactic and semantic annotations are still work in progress, although well advanced.

Syntactic annotation is being performed in cooperation with the long-term Slovene project "Communication in Slovene"[12]. The underlying formalism is a simplification of

[12] http://www.slovenscina.eu/

the approach used in the Prague Dependency Treebank and distinguishes six types of dependencies. So far, the annotators manual, a dedicated graphical editor, and a pilot gold standard treebank comprising 500 sentences have been finalised, and work is proceeding apace on annotating the complete jos100k corpus. The annotation is being performed by a team of students, with parallel annotations. Disagreements are then checked with the whatswrong[13] program, a visualizer for Natural Language Processing problems, in particular annotation disagreements.

The third level of annotation concerns lexical semantics, and consists of manually annotating selected words in jos100k with their sense in sloWNet [16,17]. sloWNet[14] is a Slovene semantic lexicon based on Princeton WordNet which currently contains around 17,000 synsets and 20,000 literals. For semantic annotation, we chose 105 nouns that are present in sloWNet and have a frequency of between 30 and 100 in jos100k. These word tokens in the corpus are currently being annotated by students for their sense (synset id) in sloWNet. So far, the annotators' manual, an interface involving Excel spreadsheets for annotations, and the first round of sense annotation for the chosen words has been performed. In the future, these annotations will be re-checked and then incorporated into jos100k.

4 Availability of the JOS resources

The JOS resources are available from the homepage of the project[15] under the Creative Commons licences and with no required registration. The resources consist of the JOS morphosyntactic specifications, the two annotated corpora and web services.

4.1 The specifications

The JOS specifications are available for browsing in Slovene and English and for download under the Creative Commons Attribution 3.0 licence. As discussed in Section 2 they comprise the source XML, associated XSLT stylesheets, the derived HTML version, and MSD conversion tables. Additionally, the morphosyntactic annotators' manual (in Slovene) is provided in PDF.

4.2 The corpora

The jos100k and jos1M corpora, currently annotated for context disambiguated MSDs and lemmas can be downloaded under the Creative Commons Attribution-Noncommercial licence. Unfortunately, we cannot allow commercial exploitation, as that is not permitted by the agreement between FIDA and FidaPLUS and the text providers, or by the agreement between JOS and the FIDA and FidaPLUS consortium, which includes also commercial partners.

The JOS corpora are available in the source XML TEI P5 encoding, as well as several derived formats, more suitable for immediate processing and exploitation. In particular,

[13] `http://code.google.com/p/whatswrong/`
[14] `http://nl.ijs.si/slownet/`
[15] `http://nl.ijs.si/jos/`

we offer the corpora in the file format for the IMS Corpus Workbench[16] [18], where each line contains either a structural tag or a tab-separated line containing the wordform, lemma, MSD, and the MSD decomposed into canonical features. This allows for searching on particular attributes of the token, regardless of the part-of-speech. For instance, to find all tokens marked as feminine genitive, the CWB query would be: `[gender="feminine" & case="genitive"]`.

As mentioned, jos100k will in the future contain additional annotations, in particular, it will be syntactically annotated with dependency trees, and semantically with WordNet senses.

4.3 Web services

In addition to downloading, the corpora can be also queried on-line, via a Web interface to the IMS Corpus Workbench. The user can select either corpus with the Slovene or English MSDs and input a CWB query directly, or use a simple tabular interface for regular expression querying over words, lemmas and MSDs over several tokens. The display can be either as separate text snippets, as standard KWIC, or as a frequency lexicon of hits. Furthermore, the word, lemma and MSDs can be displayed in the results.

We also offer a Web service that annotates Slovene texts. The service tokenises the text, and then tags it with MSDs and lemmas. The text is either pasted into the window, or a plain text file is uploaded. The service uses the tool ToTaLe [19], which is a trainable tagger and lemmatiser trained on the jos1M corpus. The service has no pre-set limits on the size of the text to be annotated, but the practical limit is about 1 million words. The Web page of the service also provides a WordSketch grammar for Slovene, so corpora produced by the service can be used in the Sketch Engine.[17]

5 Conclusions

The paper presented the initial results of the JOS project, with the focus on the JOS morphosyntactic specifications, and MSD tagged and lemmatised jos100k and jos1M corpora. The presented corpora are the first such publicly available resources for Slovene, and should significantly advance part-of-speech tagging and lemmatisation research for the language. In addition to the resources, we also provide a Web tagging and lemmatisation service useful for students and researchers to annotate their own corpora and use them for corpus linguistic research.

In addition to providing HLT and corpus linguistics resources for Slovene, JOS also introduces some novel methods into several areas of language resource development:

Open source approach to distribution. From the plethora of accessible corpora in the world, a large number are available only via a dedicated interface (concordancer) but not for download, making it impossible to use them as a dataset for HLT research. Those that can be downloaded are often not free, via ELRA or LDC, and even free corpora insist on a registration procedure which might put stringent constraints on the use of the resources. JOS resources, in contrast, are only a click away.

[16] `http://cwb.sf.net`

[17] `http://www.sketchengine.co.uk/`

Standardisation of language resources. The JOS resources are uniformly encoded in XML, according to the widely used Text Encoding Initiative Guidelines, and the JOS morphosyntactic specification will become a part of the 13+ language Version 4 MULTEXT-East specifications. This gives a firm foundation to the resources, making them platform independent, easily processable, and suitable as a basis for multilingual and cross-lingual applications.

Localisation of linguistic features. The annotation of language resources for various languages is either in the same language as the resources, or in English. The former has the advantage of enabling native speakers of the language to peruse the annotations in their language. This positively impacts on the equality of languages and the development of (linguistic) terminology for the language in question. The disadvantage, however, is in precluding foreign researchers from understanding the specifications and corpus annotations and, in some cases even processing them, as they will most likely contain characters outside the ASCII range, which can still cause problems for taggers or other HLT software. This marginalises the produced language resources and can e.g., disqalify them from participating in open tasks. In JOS we have introduced the concept of localisation into annotations, enabling both solutions to the question of language selection for the morphosyntactic specifications and corpus annotations.

Current work on the JOS resources, as mentioned, concerns the next two levels of manual linguistic annotation of the jos100k corpus.

Acknowledgments

The work described in this paper was supported in part by Slovenian Research Agency grant J2-9180 “Linguistic annotation of Slovene language: methods and resources” and by the EU 6FP-033917 project SMART “Statistical Multilingual Analysis for Retrieval and Translation”.

References

1. Erjavec, T.: MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Fourth International Conference on Language Resources and Evaluation, LREC’04, Paris, ELRA (2004) 1535 – 1538 http://nl.ijs.si/et/Bib/LREC04/.
2. Erjavec, T., Krek, S.: The JOS morphosyntactically tagged corpus of Slovene. In: Sixth International Conference on Language Resources and Evaluation, LREC’08, Paris, ELRA (2008)
3. Erjavec, T., Krek, S., Špela Arhar, Fišer, D., Ledinek, N., Saksida, A., Sivec, B., Trebar, B.: JOS Morphosyntactic Specifications for Slovene, Version 1.0. Technical report, Jožef Stefan Institute (2009) http://nl.ijs.si/jos/msd/html-en/index.html.
4. Špela Arhar, Ledinek, N.: Oblikoskladenjske oznake JOS: revizija in nadgradnja nabora oznak za avtomatsko oblikoskladenjsko označevanje slovenščine (JOS morphosyntactic tags: revision and upgrade of the tagset for automatic morphosyntactic annotation of Slovene). In: Proceedings of the Sixth Language Technologies Conference, Ljubljana, Jožef Stefan Institute (2008) 54–59 http://nl.ijs.si/is-ltc08/IS-LTC08-Proceedings.pdf.

5. Ĺ pela Arhar, Gorjanc, V.: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa (The FidaPLUS corpus: a new generation of the Slovene reference corpus). Jezik in slovstvo **52** (2007)
6. Erjavec, T.: MULTEXT-East Morphosyntactic Specifications: Towards Version 4. In: Proceedings of the third MONDILEX workshop: "Metalanguage and encoding scheme design for digital lexicography", Bratislava, Slovakia, Slovak Academy of Sciences (2009) 59–70
7. TEI Consortium, ed.: TEI P5: Guidelines for Electronic Text Encoding and Interchange. (2007)
8. Rojc, M., Verdonik, D., Kačič, Z.: A framework for efficient development of Slovenian written language resources used in speech processing applications. International Journal of Speech Technology **10** (2007) 121–141
9. Jakopin, P., Bizjak, A.: O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična Revija **45** (1997) 513–532
10. Lönneker, B.: Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo? (Part-of-speech tagging of Slovenian texts: How far did we get?). Slavistična Revija (2005) 193–210
11. Sedlacek, R., Smrz, P.: A New Czech Morphological Analyser ajka. In: Proceedings of the 4th International Conference on Text, Speech and Dialogue. Lecture Notes In Computer Science; Vol. 2166. Springer-Verlag (2001) 100 – 107
12. Bemova, A., Hajič, J., Vidova-Hladka, B., Panevova, J.: Morphological and syntactic tagging of the prague dependency treebank. In: Proceedings of ATALA Workshop. (1999) 21–29
13. Sperberg-McQueen, C.M., Burnard, L., eds.: Guidelines for Electronic Text Encoding and Interchange, Revised Reprint. The TEI Consortium (1999)
14. Erjavec, T.: An architecture for editing complex digital documents. In: Proceedings of INFuture2007: "Digital Information and Heritage", Zagreb, Croatia, Faculty of Humanities and Social Sciences (2007) 105–114
15. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA (2000) 224–231 http://www.coli.uni-sb.de/~thorsten/tnt/.
16. Erjavec, T., Fišer, D.: Building Slovene WordNet. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, Paris, ELRA (2006)
17. Fišer, D., Erjavec, T.: Predstavitev in analiza slovenskega wordneta (Presentation and analysis of Slovene wordnet). In: Proceedings of the Sixth Language Technologies Conference, Ljubljana, Jožef Stefan Institute (2008) 37–42 http://nl.ijs.si/is-ltc08/IS-LTC08-Proceedings.pdf.
18. Christ, O.: A Modular and Flexible Architecture for an Integrated Corpus Query System. In: Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research, Budapest, Hungary (1994) 23–32 http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.
19. Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R.: Massive multilingual corpus compilation: Acquis Communautaire and totale. In: Proceedings of the 2nd Language and Technology Conference, Poznan, Poland (2005) 32–36 http://nl.ijs.si/et/Bib/LTC05-acquis.pdf.

A comparative view of noun compounds in ENGLISH and ZULU

Sonja E. Bosch[1], Christiane Fellbaum[2]

[1] University of South Africa
Pretoria, South Africa
[2] Princeton University
Princeton, USA

Introduction

A workshop in Pretoria in March, 2007 brought together the authors and Karel Pala in an effort to lay the groundwork for an African WordNet, a new member of the Global WordNet family. Among the many challenges for building and linking wordnets across such distinct languages as English, Czech and Zulu, is the analysis and representation of morphologically related word forms. We published a paper on this topic (Bosch, Fellbaum and Pala, 2008) and hope that the many remaining open questions will lead to further collaboration. In this contribution, dedicated to Karel on his 70^{th} birthday, we discuss one particular type of word formation: noun compounding. We look forward to broadening the English-Zulu-based perspective presented here with Czech data.

When building lexical resources, one is continuously confronted with the question: Where is the borderline between the lexicon and the grammar of a language? Even without the space restrictions faced by traditional paper lexicography, scientific principle and parsimony mandate that the lexicon be restricted to idiosyncratic phenomena that are not attributable to the regular processes of grammar and that need to be "looked up" rather than derived by the language user. Derivational morphology—unlike inflectional morphology—presents a large grey area between regular, compositional and idiosyncratic, non-compositional word forms.

In both English and Zulu, most lexicalized compounds are noun phrases composed of two or more nouns; this class of compounds is the focus of our paper, but we will be referring to other compositions as well.

1 Formal properties of English and Zulu noun compounds

We briefly review and contrast the structural properties of English and Zulu nominal compounds.

1.1 Headedness, endocentric and exocentric compounds

The head of a phrase constitutes its salient part and serves as the basis for the name of the phrase in many cases, where the head of a noun phrase is a noun, the head of a verb phrase is a verb, etc. However, noun phrases can also have heads that are not nouns. For

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 35–44, 2009.

example, Zulu *inqumakhanda* (a person with a fine body but ugly face) consists of a verb head *-nquma* (cut off) and and noun *ikhanda* (head). Similarly, the compound noun *insizakuzwa* (hearing aid) has both a verb head *(-siza*, help) and a verb modifier (*-zwa*, hear). (See section 1.3 for more examples.)

Within a phrase, the head can be the leftmost (initial) or the rightmost (final) element, depending on whether the language is head-initial and thus right-branching (like Zulu), or head-final and thus left-branching (like English). An example of a noun compound in Zulu is *umninindlu*, composed of the phrasal *umnini* (owner) and *indlu* (house), and corresponding to English *house owner*. Similarly, *umbiki* (reporter) and *indaba* (matter, topic, affair) combine to form *umbikindaba* (journalist).

English has very few compounds like *attorney general,* where the phrasal head is not the rightmost member.

In most compounds, the phrasal head is also the semantic head, i.e., the constituent that expresses the basic meaning of the compound (in WordNet terms, the superordinate, more general, concept). For example, the semantic heads of *windowsill* and *attorney general* are *sill* and *attorney,* respectively.

The distinction between endocentric and exocentric compounds, which goes back to Panini and Sanskrit grammar, pertains to compounds with and without a semantic head, respectively. Compounds like *pickpocket*, *chimney sweep* and *scofflaw* lack a semantic head, as they are not kinds of *pockets*, *sweeps* or *laws* (their phrasal heads). Unexpressed semantic heads in many cases are implied and easily inferable; the above examples all refer to persons. Mental or physical properties of people are often encoded in exocentric compounds referring to the person possessing the respective qualities: *dimwit, half-wit, birdbrain, hunchback, paleface.* In the case of exocentric compounds like *redhead, turtleneck* and *paperback,* the phrasal head is a meronym (part) of the implied semantic head: a person has a *head,* a sweater has a *neck* and a book has a *back*. A Zulu example is *inqumakhanda* (a person with a fine body but ugly face), composed on *-nquma* (cut off) and *ikhanda* (head).

Exocentric compounds with inferable semantic heads are somewhat productive in English for certain categories. For example, the following compounds denote items of clothing (shoes, sweater, pants, jackets): *high heels, V-neck, turtleneck, jewelneck, bellbottoms, swallowtail, cutaway. Desktop* and *laptop* are convenient neologism of the kind created in response to the need to label a new concept; such compounds arise frequently, can be short-lived, and may get calqued or straightforwardly borrowed into other languages. Exocentric compounds like *copycat, scarecrow* do not have an inferable head and thus their meanings cannot be guessed; they must be listed in the lexicon.

Examples of Zulu exocentric compounds are *ubusobubili* (two-faced person), composed of *ubuso* (face) and *bubili* (two) and *umazwimabili* (ambiguous person), constituted by *amazwi* (words) and *mabili* (two). Note that the prefixes *ubu-* (for *face*) and *ama-* (for *words*), are preserved according to the rule for the head-first member of the compound; the implied semantic head (*person*) does not supply the prefix, making these compounds semantically opaque and thus candidates for lexical entries.

1.2 Compounding and Prefixation in Zulu

As in English, any Zulu lexeme containing at least two stems or roots is considered to be a compound. But unlike in English, the process of compounding in Zulu is not merely a concatenation of independent words. Words contributing to a compound may undergo phonological and morphological changes. Parts of these words "may be elided, replaced or adapted in some way or another, so that the lexical components do not appear as full words, but as bound stems and roots." (Kosch, 2006:122).

Zulu has a rich system of prefixation, and compounding affects prefixation. A new prefix is added for every compound headed by a verb. The class prefix of a noun as head may get changed in a compound, as in

inyoni (bird) + *-nco* (speckled red-and-white) > *ubunyoninco* (craftiness, cunning, bribery)

where the prefix for class 9 (*in-*) becomes class 14 (*ubu-*) in the compound.

1.3 Noun compounds with non-noun heads

Some noun compounds have a head that is not a noun; nevertheless, the category of the phrase is nominal. Examples include the large class of English nouns derived from phrasal verbs: *cutoff, pick-up*, *set-up, look-up, push-up, knock-out*, *drop-in, stowaway, shut-down, etc.* (Note that not all phrasal verbs can be nouns: *cut up, set off, look on, push away, knock back, drop down* and *shut away* have uses as verbs only.) Noun compounds composed of verbs include *wannabe*, *must-have* and *knock-me-down.* In each case, the semantic head is unexpressed and cannot be easily inferred, as there is no obvious or regular semantic relation between the noun and the implied head.

Because of their idiosyncracies, such compounds are listed in the lexicon.

Zulu may form noun compounds with a verbal head and different kinds of modifiers. Examples of verb-noun compounds are *umhlolimigwaqo* (road inspector), composed of *-hlola* (inspect) and *imigwaqo* (roads); *isithamelalanga* (sunflower) is composed of *-thamela* (bask) and *ilanga* (sun). Note that a new noun prefix is added for every compound with a verb as head. The few existing English verb-noun compounds include *shut-eye, turnkey* and *chimney sweep.*

We give examples of Zulu noun compounds made up of a verbal head and different modifiers. In the compounds below, the verb is modified by a so-called quantifier pronoun:

-hlala (live) + *wodwa* (alone) > *umhlalawodwa* (recluse)

-vuma (agree) *zonke* (all) > *uvumazonke* (credulous person; person who assents to everything).

A compound made up of a verb and an adverb is

-duma (thunder) + *kude* (far) > *udumakude* (famous person)

Compounds composed of two verbs may receive the *in-* noun prefix:

-siza (help) + *-zwa* (hear) > *insizakuzwa* (hearing aid)

-khasa (crawl) + *kabili* (twice) > *inkasakabil*i (a very old person).

1.4 Reduplication

Zulu also forms compounds by reduplication (Ungerer 1983, Kosch 2006, Okoth Okombo 2000). Noun compounds formed in this manner take on specific nuances and can be semantically classified as follows:

(1) The reduplication adds a semantic component of goodness, kindness, pleasantness, genuineness etc.:

insizwa >*insizwa-nsi* (a real young man)
umuntu > *umuntu-ntu* (a typical person)

(2) In some cases the reduplicated nouns convey a meaning of multiplicity when in the plural:

imifula > *imifulafula* (many rivers)
izimbongolo > *izimbongombongolo* (many donkeys)

(3) The reduplicated, but not the simple forms, are independent lexical items in a few cases:

inkanankana (great difficulty; big problem) < **inkana*

Marchand (1960) notes that English nouns formed by reduplication usually have onomatopoeic or otherwise expressive character: *tick-tock* (clock), *blah-blah* (meaningless talk), *hush-hush* (secret), *choo-choo* (train), *ping-pong* (table tennis). The single morpheme has no lexical status outside the compound.

2 Compositionality

A central question is, which compound nouns are semantically idiosyncratic and carry a meaning that cannot be computed by combining the meanings of the compound's constituents? The answer to this question is important in that it tells us which compounds do and don't belong into the lexicon (and, by extension, into WordNets). As with idiomatic multi-word expressions like *kick the bucket* and *get hold of,* compositionality is distributed along a spectrum. At one end are truly non-compositional and semantically opaque compounds include *butterfly, cheapskate* and *tightwad*. Their meanings cannot be easily guessed, unlike those of *dishwasher* and *bookshelf*.

2.1 Metaphoric compounds

Some idiosyncratic compounds have metaphorical character. *Ladyslipper* and *buttercup* are flowers whose folk names derive from their similarity to the objects referred to under the literal readings. An example from Zulu is *umhlwazimamba* (species of climber), composed of the nouns *umhlwazi* (species of a rare tree) and *imamba* (type of snake). Similarly*, skyscrapers and couch potatoes* are not *scrapers* and *potatoes* under any readings of these nouns, yet the motivation for these compounds is more or less apparent.

Bellwether, though it can refer to a male sheep, most often denotes a person, who is likened to a sheep wearing a bell and leading the flock.

In semantically opaque compounds like *helicopter mother* and *forklift*, the head receives a literal interpretation while the modifying member is a metaphor. Corresponding examples in Zulu are *abantunyoni* (astronauts), composed of *abantu* (people) plus *inyoni* (bird) and *umkhumbingwenya* (submarine), which combines *umkhumbi* (boat) and *ingwenya* (crocodile).

2.2 Semantic relations between the constituents

Compounding is a generative, productive process; speakers make up compounds on the fly and hearers decode them effortlessly. This means that the majority of compounds are compositional, though their meaning may depend on the context in which they are embedded (e.g., *banana war* is best interpreted in the context of trade relations among nations that are exporting and importing the fruit). The members of a compound bear a semantic relation to one another, and the relations are the basis for productive patterns.

Levi (1978) represents an attempt to classify these relations exhaustively for English compounds. However, it seems that while many compounds can be semantically classified in terms of a small number of patterns, the patterns cannot account for all compounds.

For example, the following nouns headed by *chair* specify the LOCATION of the *chair* (where it is used): *desk chair, lounge chair, deck chair, lawn chair*, etc. The modifying noun may also specify the user: *baby chair, barber chair*.

Other compounds specify salient or characteristic PARTS of the head: *armchair, wing chair, wheelchair.* The head can be modified by a verbal participle: *stacking chairs* and *folding chairs* CAN BE stacked and folded, respectively. *Rocking chairs* rock and *massage chairs* massage (ACT ON) the sitter. But these productive patterns fail in the case of *dining chair*, which neither dines nor is dined, but is used at a dining table or in a dining room. *Easy chair* is truly idiosyncractic and does not follow a productive pattern. (Note that *easy chair* is a compound and not an adjective-noun combination: **the red and easy chair, *the very easy chair*.)

The productivity of many compounds can be attributed to patterns as well as pragmatics. Thus, if the head is a Material, the meanings of the modifier can be inferred from its function or its creation, etc.: *paperbag, plastic bag* (made from paper/plastic), *paper mill/plastic mill* (mill for producing paper/plastic). More likely candidates for inclusion in the lexicon are *paper trail* and *paper cut*, whose meanings do not follow from the fact that the head is a Material.

In English compounds formed from a verb-*ing* and a noun constituent, the noun can encode semantic roles associated with the event denoted by the verbs, such as Theme/Patient, Location, Instrument, Manner:

Baking potato/eating apple
Baking dish/sheet
Baking directions/instructions/recipes
Baking powder/soda

Poulos and Msimang (1998:86) note with respect to Zulu noun compounds that "the resultant meaning… is either a combination of the individual meanings of the two parts or an idiomatic or free rendering of the two meanings." But patterns like the above show that

a more differentiated view is required. The meaning of the whole may not be just a "free rendering" of one or both constituents; moreover, the borderline between compositionality and idiomaticity is fuzzy. A cross-linguistic perspective can show the universality of the patterns and, conversely, identify idiosyncratic, language-specific compounds.

2.3 Systematicity in Zulu noun compounds

Various types of compounds can be distinguished in terms of the relations between the semantic head and the modifying constituent. We exemplify types of s*ubordinate* and *possessive* compounds where one element modifies the other with regard to qualification, locality or possession for instance.

Examples for the group headed by the noun *umnini* (owner) include:

umnini (owner) + *amandla* (strength) > *umninimandla* (strong person, authority)
umnini (owner) + *indlu* (house) > *umninindlu* (house owner, family head)
umnini (owner) + *isitolo* (shop) > *umninisitolo* (shop owner)
umnini (owner) + *umuzi* (kraal) > *umninimuzi* (chief)

The English equivalents are formed with different heads, such as *keeper (shop keeper*), *holder (landholder, householder).*

Another group of compounds is formed with the head noun *umgcini* (guard):

umgcini (guard) + *umnyango* (door) > *umgcinimnyango* (door guard)
umgcini (guard) + *isihlalo* (chair) > *umgcinisihlalo* (chairperson; speaker)
umgcini (guard) + *isikhwama* (bag) > *umgcinisikhwama* (treasurer)

English often uses *man* in equivalent compounds: *doorman, spokesman*, *chairman*.

A third group of compounds, formed from the verbal head –*linda* (look after), includes

-linda (look after) + *inkosi* (king) > *isilindankosi* (king's body guard)
-linda (look after) + *insimu* (field) > *umlindansimu* (scare crow)
-linda (look after) + *isango* (gate) > *umlindisango* (gate guard)

Finally, the following noun compounds share the verbal head *-vala* (close):

-vala (close) + *amehl*o (eyes) > *imvalamehlo* (sun visor; eye protector)
-vala (close) + *umlomo* (mouth) > *imvalamlomo* (bribery)
-vala (close) + *umphimbo* (throat) > *imvalamphimbo* (epiglottis)
-vala (close) + *isangwana* (small gate) > *imvalasangwana* (gate guard; slow coach).

As English possesses few compounds headed by a verb, the last two groups have no comparable equivalents.

2.4 Zulu possessor noun compounds

Dimmendaal (2000:170) points out that "many African languages have a special construction expressing 'owner of' consisting of a possessor showing connections with a wide array of forms, for example 'chief', 'self', 'father', 'mother' plus the possessed item (in either order) . . . "

In some Bantu languages this specific type of compound noun expressing 'owner of' incorporates a so-called "abbreviated noun," which in the case of Zulu is *-so-* or *-no-*. Bosch and Prinsloo (2001:95) argue that these abbreviated nouns, derived from words referring to 'father' and 'mother' respectively, have lost their status as fully fledged lexical items and have developed into grammaticalized forms over time. Metaphorical

usage and a subsequent process of desemanticization, led to the reanalysis of *-so-* and *-no-* in Zulu, as grammatical units that are used productively to coin new words. In other words, they no longer function as independent nouns, but are usually compounded with a noun expressing a range of meanings relating to *owner of an object or having special skills/characteristics* or even *occupations*, as illustrated in the following examples:

usomashibhini
u-so + (a)mashibhini
cl.pref.1a-father + shebeen
'shebeen owner'

usomabhizinisi
u-so + (a)mabhizinisi
cl.pref.1a-father + businesses
'businessman/woman,
owner of businesses'

unompempe
u-no + (i)mpempe
cl.pref.1a-mother + whistle
(lit. owner of the whistle)
'referee'

unozinti
u-no + (i)zinti
cl.pref.1a-mother + sticks
(lit. owner of the sticks, i.e.
goal posts)
'goalkeeper'

Further examples are:

usokhemisi (> *ikhemisi* 'chemist shop')
'pharmacist'

usosayensi (> *isayenisi* 'science')
'scientist'

usonhlalakahle (> *inhlalakahle* 'welfare')
'social worker'

usolwazi (> *ulwazi* 'knowledge')
'professor'

unobhala (<*-bhala* 'to write')
'secretary'

A semantic continuum is postulated by Bosch and Prinsloo (2001:97) representing the range of these abbreviated nouns from the original meanings 'father/mother of', as one extreme through 'owner of' or 'to have special skills/characteristics' and 'occupation' as the other extreme.

The decision to list such abbreviated compound nouns in the lexicon would depend on their semantic transparency.

2.5 Ambiguity

In some cases, several semantic relations between the members are plausible, making the compound genuinely ambiguous. For example, the English compound *child murderer* could refer to either a child that has murdered or to the murderer of a child. Zulu by contrast forms two different compounds: *umbulali wezingane* (murderer of children, with 'murderer' as the head) and *ingane ongumbulali* (child (who is) murderer, where 'child' is the head).

2.6 Appositional compounds

In compounds such as *actor-director* and *teenage mother* there is no semantic relation among the constituents. Both members refer to different aspects of the referent. An *actor-director* is both an *actor* and a *director*, and a *teenage mother* is both a *teenager* and a *mother*. An interesting question concerns the order of the constituents: *mother teenager and *director-actor are distinctly odd, indicating that the second member functions like a semantic head. Note that compounds that are not appositional change their meanings when the order of the members is reversed with a concomitant change of heads: *pine forest-forest pine, rose hedge-hedge rose,* etc.

Zulu appositional compounds are formed as follows by means of a possessive constructon: *unesi wesilisa* (nurse of male)

3 Noun compounds in NLP applications

The morphological and semantic analysis of noun compounds is important for Natural Language Processing applications such as Information Retrieval and Machine Translation. Before it can be assigned a meaning, the compound has to be recognized as such and provided with the appropriate part-of-speech tag. This is relatively easy with a compound consisting of two nouns, such as *wing chair*, and more difficult with compounds such as *chimney sweep*, where one constituent (possibly the head) is not a noun. Two consecutive nouns are good indicators of a compound, so long as the parser has ruled out certain syntactic structures like reduced relative clauses (as in *this is the **law judges** apply regularly*). Once a compound has been recognized, it can be looked up in the lexical database. Assuming there is only one entry, the compound can be interpreted with reasonable confidence.

If a compound has more than one possible dictionary entry, this simple procedure fails. For example *angel hair* may refer to kind of pasta or to a substance claimed to emanate from Unidentified Flying Objects. Moreover, it is possible that the reading associated with the dictionary entry is in fact not the one intended in a particular context; a literal, compositional reading (hair of an angel or angel-like hair) could well be intended. In these cases, a semantic analysis of the context may be required to disambiguate the compound.

How can an automatic system interpret compound nouns that it doesn't find in the lexical database (or if it determines that the lexical entry is not the context-appropriate one)? In those cases, the semantic relation between the compound members should be determined that is expressible with phrases like X is performed on Y (*baking potato*) and X is the location for Y (*baking dish*). Work by Kim and Baldwin (2008) and Kim, Mistica

and Baldwin (2007) has recast the interpretation of noun compounds as a Semantic Role labeling task; the Semantic Roles express the relations among the members of different compounds as examined by, e.g., Levi (1978). Kim, Mistica and Baldwin took advantage of the WordNet hypernym structure to identify concepts like PRODUCTS and PRODUCER, which form a semantic subclass of noun compounds such as *honey bee* and *music clock*. Performance on automatic noun compound interpretation currently hovers near the 7 level, indicating that more work is needed to meet this challenge.

4 Conventionality and Frequency

Lexical resources often include arguably compositional compounds like *lawnmower, rice cooker, seatbelt* and *flyswatter*. Although speakers not familiar with these compounds can understand their meanings in terms of their functions (an artifact used to mow the lawn/cook rice/swat flies; a belt used when in a seat), they are not able to form an idea of their design and working. The nature of some artifacts is impossible to derive from their verbal label, even when it is arguably compositional. Dictionaries frequently treat such compounds on a par with semantically unanalyzable compounds, perhaps so as to provide a descriptive definition of the nature of commonplace artifacts. Similar examples are to be found in Zulu, e.g. *isinqamulacingo* (pliers to cut wire), composed of the verb *-nqamula* (cut) and the noun *ucingo* (wire).

5 Compounds in WordNets

A compound that is judged idiosyncratic needs to be listed in the WordNet of that language and represented in the interlingual ontology. It need not be included in the WordNets of other languages where it is compositional, and we expect this to be the case for many, if not most, noun compounds. For example, English *chimneysweep* and *chopstick* have regularly formed, fully compositional equivalents in German *(Schornsteinfeger: chimney + sweeper; Essstaebchen: eat + sticks).*

On the other hand, certain concepts may be encoded by idiosyncratic compounds crosslingually– similarly to verb phrase idioms—as they have a colloquial, derogatory, or euphemistic connotation. An example is *cheapskate*, which corresponds to German *Geizhals* (lit., 'stinginess neck').

6 Conclusion

We discussed types of noun compounds in English and Zulu. There are many commonalities as well some compound-formation processes that are specific to Zulu. We anticipate additional data from Czech that will provide a still broader perspective.

The integration of compounds into lexical resources, and WordNets in particular, remains a challenge that needs to be considered in terms of the compounds' syntactic idiosyncracy and semantic compositionality.

References

1. Bosch, Sonja, Fellbaum, Christiane and Pala, Karel. 2008. Derivational Relations in English, Czech and Zulu Wordnets. *Literator* 29(1):139-162.
2. Bosch, Sonja E and Prinsloo, DJ. 2002. 'Abbreviated nouns' in African languages: a morphological, semantic and lexicographic perspective. *South African Journal of African Languages*, 22.1:92-104.
3. Dimmendaal, Gerrit J. 2000. Morphology. In: Heine, Bernd and Nurse, Derek: *African Languages: An Introduction*. Cambridge: Cambridge University Press, 161-193.
4. Kim, Su Nam and Baldwin, Timothy. 2008. An Unsupervised Approach to Interpreting Noun Compounds. In: Proceedings of the International Conference on NLP and KE, Beijing, China.
5. Kim, Su Nam, Mistica, Meladel and Balwin, Timothy. 2007. Extending Sense Collocations in Interpreting Noun Compounds. In: Proceedings of the Australasian Language Technology Workshop, 49-56.
6. Kosch, Ingeborg M. (2006). *Topics in Morphology in the African Language Context*. Pretoria: Unisa Press.
7. Levi, Judith. 1978. *The Syntax and Semantics of Complex Nominals.* New York: Academic Press.
8. Marchand, Hans. 1960. *The Categories and Types of Present-Day English Word-Formation.* Munich: C.H. Beck.
9. Okoth Okombo, Daniel. 2000. Building techniques in African languages. In: Vic Webb and Kembo-Sure (eds.), *African Voices*. Cape Town: Oxford University Press Southern Africa, 197-219.
10. Poulos, George and Msimang, Christian Themba. 1998. *A linguistic analysis of Zulu.* Pretoria: Via Afrika Limited.
11. Ungerer, H. J. 1983. *Komposita in Zulu*. Johannesburg: Unpublished Doctoral Thesis. Johannesburg: Rand Afrikaans University.

What is a "full statistical model" of a language and are there short cuts to it?

D. Guthrie, L. Guthrie, Y. Wilks

University of Sheffield, UK
{dguthrie, louise, yorick}@dcs.shef.ac.uk

Although this article is not directly about Slavic Natural Language process (the title of this book), it describes a small investigation into a technique that may aid in representing the diversity that we find in every language. We ask what it would be like to have a *full trigram model of a language*, like English or French, where a full model would mean that given any new, unseen, test text it would be virtually certain that all the trigrams in the text (i.e. all sequences of three words) have been seen before in training data.

The notion of a trigram model is the standard one in statistical text analysis (Jelinek and Lafferty, 1991) since the limit of three represents a good compromise between a useful model of language content and the likelihood of seeing those sequences repeated; fourgrams would of course give a very accurate representation of context, but are less likely to ever be seen again, whereas bigrams, and single words, are repeated frequently, but not very useful at identifying context. A full trigram model would ;void the necessity of techniques like "smoothing" (ibid.) to compensate for the fact that so many trigrams are normally unseen in the data available. Until recently, it has seemed that the corpus of a language that would have to be gathered and analyzed to produce that result would be improbably large, larger than the world wide web, for example. The origins of such language models go back, in recent research, to Brown et al.'s (1990) efforts twenty years ago to perform machine translation using only a statistical model based on trigrams. They were only partly successful but their efforts drew attention to the notion that language consists of "rare events", and that language data always seem too sparse to build a full or complete model of a language. We introduce the notion of a skip-gram and show by experiment that this modification to tri-grams (roughly, by adding gaps in the tri-grams) can give a much fuller model of a language with little loss. We speculate on how close we are, computationally, to a full model and what that may mean for important language-based processes, such as Berners-Lee's concept of the future Semantic Web (2002). In contrast with the World Wide Web, which consists of documents humans can read but machines cannot, the Semantic Web would be a web that machines too could understand. In this paper, we discuss this future vision and what role a full model of a language might play in achieving it.

1 Introduction

The "empiricism of use" approach that has been standard in Natural Language Processing (NLP) since the work of Jelinek (Jelinek and Lafferty, 1991) has effectively driven "good old-fashioned Artificial Intelligence" style approaches based on symbolic logic to the periphery of NLP. It will be remembered that Jelinek attempted to build a machine translation system at IBM based entirely on machine learning from bilingual corpora, and

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 45–56, 2009.

with no access to linguistic constructs like grammars or lexicons. He was not ultimately successful—in the sense that his results never beat those from the leading hand-crafted system, SYSTRAN — but he changed the direction of the field of NLP as researchers tried to reconstruct, by empirical methods, the linguistic objects on which NLP had traditionally rested: the very same lexicons, grammars, etc. The barrier to further advances in NLP by these methods seems to be the "data sparsity" problem to which Jelinek originally drew attention, namely that language is "a system of rare events" so varied and complex, that a complete model for a language seems impossibly difficult to derive. Many of the trigrams (3 word sequences), for example, in any new, unseen, text corpus will never have been seen before.

Nonetheless, corpus-based trends in language processing continue to rely on the premise that language use should be modelled from training data and that sufficient data can be gathered to depict typical (or atypical) language use accurately (Young and Chase, 1998; Church, 1998; Brown, 1990). Smoothing techniques are used to account for (assign probabilities to) any data that has not been seen, yet researchers generally believe that data sparsity is causing the less than perfect results in most applications. One might argue that if smoothing were a completely satisfying method for the development of a language model, then data sparsity would not be blamed for low accuracy figures, since the probability estimates for unseen data would be very like those in a full language model. It is clear, however, that language models that make use of smoothing are much more accurate the more training data that is available and so the quest for better language models is ongoing.

It may now be possible, using the whole web — and thus reducing data sparsity — to produce much larger models of a language and to come far closer to the full language model that will be needed for tasks like complete annotation and automatically generated ontologies. A disciple of Wittgenstein (1953) will always want to look for "the use of language rather than the meaning", and nowhere is more use available than the whole web itself, even if it could not possibly be the usage of a single individual. Work will be briefly described here that seeks to quantify this idea of sparseness, and suggests the use of skip-grams as a means of reducing the sparseness of data. These results are as yet only suggestive and not complete, but they do seem to offer a way forward.

2 Background

Given a process, we use the term statistical modeling to denote defining a probability distribution that gives a reasonable explanation of the data generated by the process, and which can be used to predict future outputs of that process. Statistical Language modeling has often considered the "process" to be the selection of a three-word sequence (a trigram). The model (often called the language model) is a probability distribution on all three-word sequences, which represents the probability that a randomly selected three-word sequence will be that trigram. Another example of a statistical model used in language processing (but distinctly NOT called a "language model" in the statistical MT literature) is the early Jelinek model for translation, where a probability p(w1, w2) is assigned to pairs of words (one from the source and one from the target language), which represented the probability that word w1 will be translated as w2).

In Speech recognition, the language model is used to predict the most likely next word given a history (what has been said before up to a given point). In practice, only the last two three or four words said are taken into account to represent the history. When using a trigram language model for example, the history is always a 2-word sequence, and a probability distribution is defined for each trigram in the language. We have shown (Allison et al 2006) that even after collecting all trigrams from 30 years worth of newswire text, more than one third of all trigrams in a new text, will not have been seen. Speech recognition systems typically have attempted to exploit the available data to the maximum extent possible by *smoothing* over all unseen trigrams (making sure that no unseen trigrams have a zero probability of occurrence by assigning a small probability to all trigrams based on the bigrams and unigrams that compose them). Linguistic approaches have likewise been explored to make the best use of training text available by defining and manipulating data beyond the words in the text (part-of-speech tags, syntactic categories, parse trees etc.). While these approaches can lessen the effects of data sparsity for many tasks, they are a poor substitute for using larger corpora.

Kilgarriff and Grefenstette (2001) were among the first to point out that the web itself can now be used as a language corpus in principle, even though that corpus is far larger than any human could read in a lifetime as a basis for language learning. A rough computation shows that it would take about 60,000 years of constant reading for a person to read all the English documents on the WWW at the time of writing. But the issue here is not building a psychological model of an individual and so this fact about size need not deter us: Moore (2004) has noted that current speech learning methods would entail that a baby could only learn to speak after a hundred years of exposure to data. But this fact has been no drawback to the development of effective speech technology — in the absence of anything better. A simple and striking demonstration of the value of treating the whole web as a corpus has been shown in experiments by e.g. Grefenstette (2004) who demonstrated that the most web-frequent translation of a word pair — from among all possible translation equivalent word pairs in combination — is invariably also the correct translation.

3 A Skipgram Approach

What follows is a very brief description of the kind of results coming from the REVEAL project (Guthrie et al. 2006), which takes large corpora, such as a 1.5 billion word corpus from the web, and asks how much of a test corpus is covered by the trigrams present in that large training corpus. The project considers both regular trigrams and *skipgrams*: which are trigrams allowing a discontinuity of words. For example, if we take the sentence:

"*Chelsea celebrate Premiership success.*"

Then the two standard tri-grams in that sequence will be:

Chelsea celebrate Premiership
celebrate Premiership success

But the one-skip tri-grams will be:

Chelsea celebrate success
Chelsea Premiership success

Which seem at least as informative, intuitively, as the original trigrams.

One part of the REVEAL project investigated automatic techniques to determine anomalous phrases in text. For this problem, using a standard trigram language model is not good enough. Our estimates show that, with large amounts of training data, we can expect 30% of the trigrams to be unseen, but we do not expect 30% to be anomalous. We would like to distinguish between normal 3 word phrases that have not been seen and abnormal three word phrases (these may be a novel use of a word, they may be an unusual word exploitation in the Hanks (2009) sense, or may be a case of textual obfuscation (such as using one word to disguise another in a way that might be spotted by some keyword approach (Jibari et. al 2008). To approach this problem we focused on coverage experiments.

In an attempt to quantify just how far away we are from full language model, we looked at coverage experiments rather than perplexity experiments. In other words, we were interested in designing experiments to estimate the number of trigrams in a new text that are unseen in the training corpus. We varied the corpus size and after training computed the percentage of trigrams in the new text that are missed by the model (before smoothing). We then sought to quantify the impact skip-gram modelling has on the coverage of trigrams in real text and compare this to coverage obtained by increasing the size of the corpus used to build a traditional language model.

The work of Kaplan (1950) on disambiguation experiments showed that subjects could compare disambiguation results effectively when given 2 words on either side of an ambiguous word. He showed that they did as well with this shorter context, as with the whole sentence. Similar results were shown by Koutsoudas and Korfhage (1956), for Russian, by Gougenheim and Michéa (1961) for French, and again by Choueka and Lusignan (1985) for French. Inspired by this body of work, we limited the space complexity of our work to trigrams which allow at most 4 skips in total.

3.1 Defining skip-grams

We define k-skip-n-grams for a sentence $w_1 ... w_m$ to be the set

$$\{w_{i_1}, w_{i_2}, ...w_{in} | \sum_{j=1}^{n} i_j - i_{j-1} < k\}$$

Skip-grams reported for a certain skip distance k allow a total of k or less skips to construct the n-gram. As such, "4-skip-n-gram" results include 4 skips, 3 skips, 2 skips, 1 skip, and 0 skips (typical n-grams formed from adjacent words).

Here is an actual sentence example showing 2-skip-bi-grams and tri-grams compared to standard bi-grams and trigrams consisting of adjacent words for the sentence:

"*Insurgents killed in ongoing fighting.*"

Bi-grams = {insurgents killed, killed in, in ongoing, ongoing fighting}.

2-skip-bi-grams = {insurgents killed, insurgents in, insurgents ongoing, killed in, killed ongoing, killed fighting, in ongoing, in fighting, ongoing fighting}

Tri-grams = {insurgents killed in, killed in ongoing, in ongoing fighting}.

2-skip-tri-grams = {insurgents killed in, insurgents killed ongoing, insurgents killed fighting, insurgents in ongoing, insurgents in fighting, insurgents ongoing fighting, killed in ongoing, killed in fighting, killed ongoing fighting, in ongoing fighting}.

One can see from this example that the number of trigrams increases when more skips are allowed. A typical sentence of ten words, for example, will produce 8 trigrams, but 80 4-skip-tri-grams.

For an n word sentence, the number of trigrams with exactly k skips is given by:

$$(n - (k + 2)) (k+ 1), \text{ for } n>k+3$$

and the number of k-skip trigrams (meaning k skips or less) is given by

$$\frac{(k+1)(k+2)}{6}(3n - 2k - 6), \text{ where n> k+2.}$$

The equations above illustrate that many more tri-grams can be generated for large sentences using skip-tri-grams. This is a lot of extra contextual information that could be very beneficial provided that these skip-grams truly expand the representation of context. If a large percentage of these extra tri-grams are meaningless and skew the context model then the cost of producing and storing them could be prohibitive. Later in this paper we attempt to test whether this is the case.

Techniques that make use of skipgram models have largely been confined to speech processing and were initially applied to phonemes in human speech, but have since been applied to words. Skipgrams have been used to build language models , often in conjunction with other modelling techniques, for the goal of improving speech recognition performance (Goodman, 2001; Rosenfeld, 1994; Ney et al., 1994; Siu and Ostendorf, 2000).

4 A Simple Initial Experiment

4.1 Training Data

We constructed a range of language models from each of two different corpora using skipgrams of up to 4 skips for various corpus sizes taken from the collection below.

British National Corpus: The BNC is a 100 million word balanced corpus of British English. It is contains written and spoken text from a variety of sources and covers many domains.

English Gigaword: The Gigaword English Corpus is a large archive of text data acquired by the Linguistic Data Consortium. The corpus consists of over 1.7 billion words of English newswire from four distinct international sources.

4.2 Testing data

We used several different genres for test data in order to compare skipgram coverage on documents similar to and different from (i.e. anomalous with regard) to the training.

1. Eight Recent News Documents- From the Daily Telegraph.
2. Google Translations

Seven different Chinese newspaper articles of approximately 500 words each were chosen and run through the Google automatic translation engine to produce English texts. Web translation engines are know for their inaccuracy and ability to generating extremely odd phrases that are often very different from text written by a native speaker. The intention was to produce highly unusual texts, where meaning is approximately retained but coherence can be minimal. A short sample follows:

> *BBC Chinese net news: CIA Bureau Chief Gauss told USA the senator, the card you reaches still is attempting to avoid the American information authority, implemented the attack to the American native place goal. Gauss said, the card you will reach or if have the relation other terrorist organizations sooner or later must use the biochemistry or the nuclear weapon attack USA, this possibly only will be the time question. But he said, the card you reach only only are a holy war organization more widespread threat on the one hand.*

4.3 Method

Skipgram tests were conducted using various numbers of skips, but skips were never allowed to cross sentence boundaries. Training and test corpora were all prepared by removing all non-alphanumeric characters, converting all words to lowercase and replacing all numbers with the <NUM> tag.

We quantified the increase in coverage attained when using skipgrams on both similar and anomalous documents (with respect to the training corpus). To achieve this we computed all possible skip-grams in the training corpus and measured how many adjacent n-grams these covered in test documents. These coverage results were directly comparable with normal n-gram coverage in an unseen text results because we still measured coverage of standard adjacent bi-grams or tri-grams in the test documents and were only collecting skip-grams from the training corpus.

4.4 Results

Our first experiment (Figure 1) illustrates the improvement in coverage achieved when using skip-grams compared to standard bi-grams. We trained on the entire BNC and measured the coverage of k-skip on 300 thousand words of newswire from the Gigaword corpus. The BNC is made up of many different kinds of text other than news, but nonetheless, coverage is still improved. However, in some sense, the results are unsurprising, as there are many more bi-grams observed in training when allowing skips, but it does show that enough of these are legal bi-grams that actually occurred in a test document.

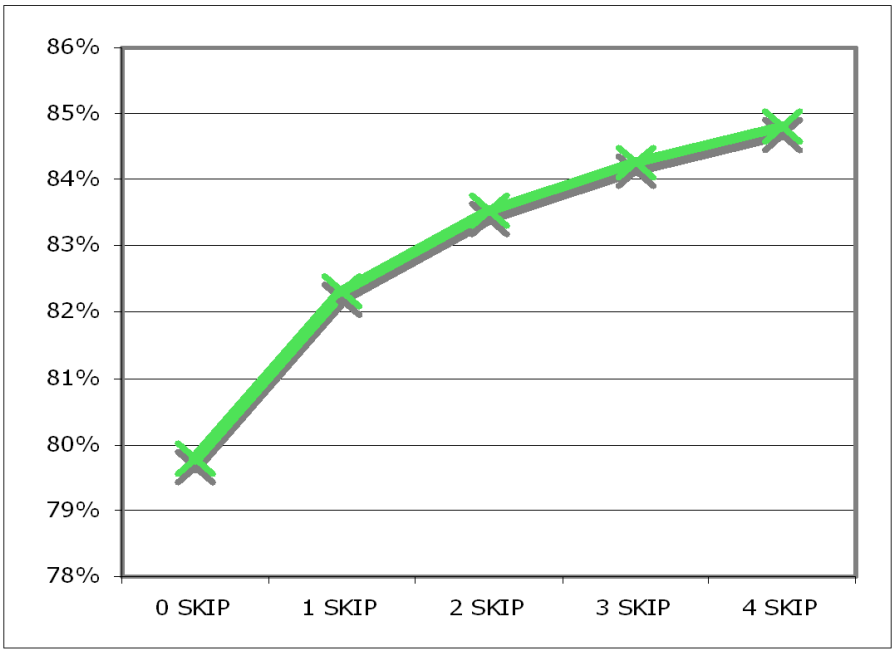


Fig. 1. Coverage of *k*-skip bi-grams on 300,000 words of news wire

The next test (Figure 2) is the same as the previous, but using tri-grams instead of bi-grams. From these results it seems that skip-grams are not improving tri-gram coverage to a very acceptable level but, as the later results show, this seems to be due to the fact that the BNC is not a specialized corpus of News text. Computing skip-grams on training documents that differ from the domain of the test document seems to add very little to the coverage. This is a promising result, in that it shows that generating random skipgrams from any corpus does not aid in capturing context.

4.5 Skipgram usefulness

Documents about different topics, or from different domains, will have less adjacent n-grams in common than documents from similar topics or domains. It is possible to use this fact to pick documents that are similar to the training corpus based on the percentage of n-grams they share with the training corpus. This is an important feature of n-gram modelling and a good indication the context is being modelled accurately. If all documents, even those on very different topics, had approximately the same percentage of n-grams in common with the training data, then we would argue that it is not clear that any context is really being modelled. *The use of skip-grams to capture context is dependent upon them increasing the coverage of n-grams in similar documents, while not increasing the n-gram coverage in different (or anomalous) documents to the extent that tri-grams can no longer be used to distinguish documents.*

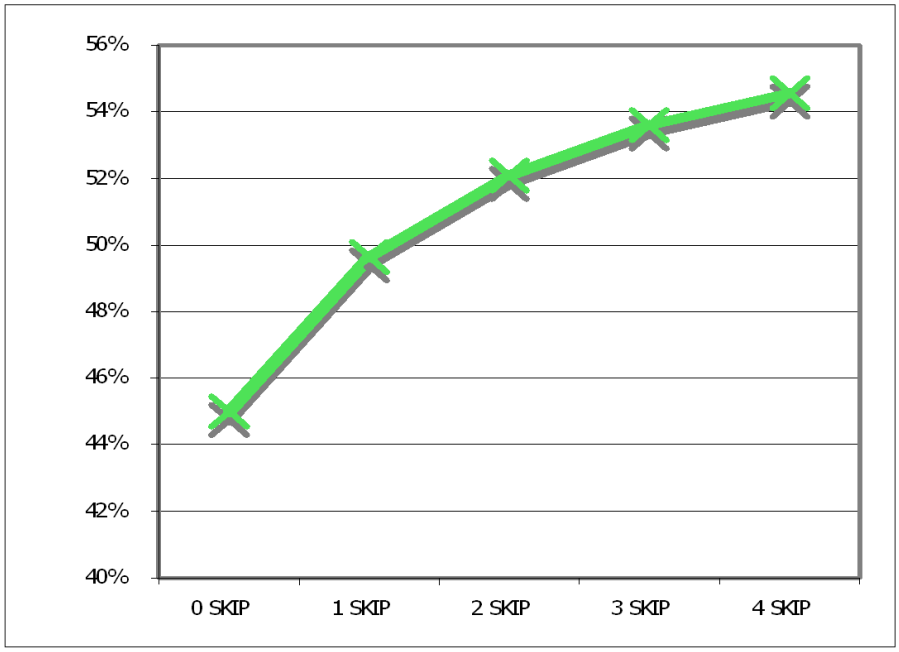


Fig. 2. Coverage of *k*-skip tri grams on 300,000 words of news wire

We tested this by training on the BNC and testing against British newspaper extracts and texts generated with Google's Chinese to English translation engine (the genre is the same, but the text is generated by an MT system).

The results (Table 1) not only illustrate the difference between machine translated text and standard English, they also show that as skip distance increases coverage increases for all documents, but it does not increase to the extent that one cannot distinguish the Google translations from the News documents. These results demonstrate that skip-grams are accurately modelling context, while not skewing the effects of tri-gram modelling. It seems that most of the skip-grams produced are either useful or they are too random to give false positives.

4.6 Skip-grams versus more training data

Normally, increasing the size of your training corpus is not an option due to lack of resources. In this section we examine skip-grams as an alternative to increasing the size of training data. The following experiments use different sized portions of the Gigaword corpus as training data and a separate, randomly chosen, 300-thousand word blind section of the Gigaword corpus for testing. We increase the amount of training and compare the results to using skip-grams for coverage. The resulting percentages are very high for trigram coverage, which is not surprising since both training and test documents come from the same domain specific corpus.

Table 1. *k*-skip tri-gram coverage on English news and machine translated Chinese news

Subject	0-skip	2-skip	3-skip	4-skip
NEWS 1	47.70%	56.69%	58.66%	62.11%
NEWS 2	55.40%	63.97%	65.67%	66.82%
NEWS 3	56.40%	60.61%	62.29%	65.24%
NEWS 4	52.68%	59.52%	62.04%	66.27%
NEWS 5	58.23%	63.80%	66.60%	71.58%
NEWS 6	54.17%	61.00%	62.95%	65.97%
NEWS 7	54.57%	61.86%	65.81%	70.48%
NEWS 8	56.23%	63.49%	65.75%	71.95%
Average	**54.42%**	**61.37%**	**63.72%**	**67.55%**

Translation 1	37.18%	45.56%	47.93%	50%
Translation 2	15.33%	23.64%	25.45%	22.61%
Translation 3	32.74%	40.22%	42.66%	45.28%
Translation 4	37.01%	33.07%	35.87%	38.05%
Translation 5	33.50%	38.09%	40.70%	42.24%
Translation 6	31.75%	39.20%	41.92%	42.71%
Translation 7	34.26%	38.54%	41.76%	42.52%
Average	**31.68%**	**36.90%**	**39.47%**	**40%**

Table 2. Corpus size vs. skip-gram coverage on a 300,000 word news document

Size of Training	base tri-gram coverage	4 skip tri-gram coverage
10 M words	44.36%	53.76%
27.5 M words	53.23%	62.59%
50 M words	60.16%	**69.04%**
100 M words	65.31%	74.18%
200 M words	**69.37%**	79.44%

Our experiments suggest that, surprisingly, skipgrams do not buy additional coverage at the expense of producing nonsense. Recent work shows the use of skip-grams can be more effective than increasing the corpus size. *In the case of a 50 million-word corpus, similar results (in terms of coverage of test texts) are achieved using skipgrams as by quadrupling corpus size.* This illustrates a possible use of skipgrams to expand contextual information to get something closer to 100% coverage with a (skip) trigram model, combining greater coverage with little degradation, and thus achieving something much closer to Jelinek's original goal for an empirical corpus linguistics.

The 1.5 billion word training corpus gives a 67%+ coverage by such trigrams of randomly chosen 1000 word test texts in English, which is to say 67% of the trigrams found in any random 1000 passage of English were already found in the gigaword corpus. But we obtained 74% coverage with 4skiptrigrams, which suggests, by extrapolation, that it would need 75x10*10 words to give 100% trigram coverage (including skipgrams

up to 4grams). Our corpus giving 74% coverage was 15x10*8 words, and Greffenstette (2003) calculated there were over 10*11 words of English on the web in 2003 (I.e. about 12 times what Google indexed at that time), so the corpus needed for complete coverage of training texts by trigrams would be about seven times the full English web in 2003, which is somewhat closer to the size of today's (2008) English web.

All this is, again, preliminary and tentative, but it suggests that empiricism of usage may now be more accessible (with corpora closer to the whole web) than Jelinek thought at the time (1990) of his major MT work at IBM.

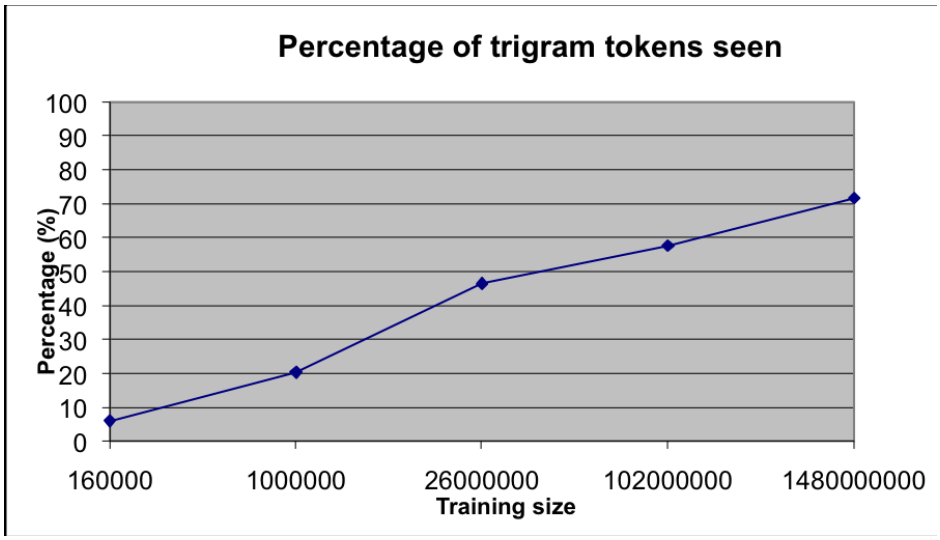


Fig. 3. Percentage of trigrams seen with training corpus size

Modern web corpora are so vast they cannot conceivably offer a model of how humans process semantics, so a cognitive semantics based on such usage remains an open question. However, one possible way forward would be to adapt skipgrams so as to make them able (perhaps with the aid of a large-scale fast surface parser of the kind already applied to large chunks of the WWW) to pick up Agent-Action-Object triples capturing proto-facts in very large numbers. This is a old dream going back at least to (Wilks, 1972) where they were seen as trivial Wittgensteinian "forms of fact", later revived by Greffenstette (Kilgarriff and Grefenstette, 2001) as a "massive lexicon" and now available as inventories of surface facts at ISI (Hovy, 2005). These objects will not be very different from standard RDF triples—the standard Agent-Action-Object data structures used in the Semantic Web (SW)(Bereners-Lee et al., 2002REF) and might offer a way to deriving massive SW content on the cheap, even simpler than that now offered by machine learning-based Information Extraction. If anything were possible along these lines, then NLP would be able to provide the base semantics of the SW more effectively than it does now, by making use of some very large portion of the WWW as its corpus. If one finds this notion unattractive, one should demonstrate in its place some other plausible technique

for deriving the massive underlying semantic (RDF) content the SW will require. Can anyone seriously believe that can be done other than by NLP techniques of some type like the one described above?

References

1. Ben Allison, David Guthrie, Louise Guthrie, Wei Liu, Yorick Wilks. 2005. Quantifying the Likelihood of Unseen Events: A further look at the data Sparsity problem. Awaiting publication.
2. Tim Berners-Lee, Jim Hendler, and Ora Lasilla. 2002. The Semantic Web. *Scientific American.*
3. Peter F. Brown, John Cocke, Stephen A. DellaPietra, Vincent J. DellaPietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2): 79–85, June.
4. Yaacov Choueka and Serge Lusignan. 1985. Disambiguation by short contexts. *Computers and the Humantities*, 19:147–158.
5. Kenneth Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143.
6. A. Koutsoudas and R. Korfhage. 1956. Mechanical Translation and the Problem of Multiple Meaning, *Mechanical Translation* 3(2):46-51, 61.
7. Joshua Goodman. 2001. A Bit of Progress in Language Modelling. *Computer Speech and Language*, October 2001, pages 403-434.
8. Patrick Hanks. Forthcoming 2009. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press
9. Frederick Jelinek and J.D. Lafferty. 1991. Computation of the Probability of Initial Substring Generation by Stochastic Context Free Grammars. *Computational Linguistics* 17:3: 315-323.
10. Georges Gougenheim and René Michéa. 1961. Sur la détermination du sens d'un mot au moyen du contexte. *La Traduction Automatique*, 2(1):16—17.
11. Abraham Kaplan. 1950. "An experimental study of ambiguity and context." Mimeographed, 18pp, November 1950. [Published as: Kaplan, Abraham (1955). "An experimental study of ambiguity and context." *Mechanical Translation, **2**(2), 39-46.*
12. Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer, Speech, and Language*, 8:1-38.
13. Sanaz Jabbari, Ben Allison, Louise Guthrie. 2008. Using a probabilistic model of context to detect word obfuscation. To appear in the *Proceedings of the sixth International Conference on Language Resources and Evaluation(LREC-2008)*
14. Ronald Rosenfeld. 1994. Adaptive Statistical Language Modelling: A Maximum Entropy Approach. Ph.D. thesis, Carnegie Mellon University, April.
15. Manhung Siu and Mari Ostendorf. 2000. Variable n-grams and extensions for conversational speech language modelling. *IEEE Transactions on Speech and Audio Processing*, 8:63–75.
16. Steven J. Young and L.L. Chase. 1998. Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes, *Computer Speech and Language*, 12, 263-279.
17. Yorick Wilks. 1972. *Grammar, Meaning and the Mcahine Analysis of Language.* Routledge, London and Boston.
18. Ludwig Wittgenstein. 1953. *Philosophical Investigations (PI).* G.E.M. Anscombe and R. Rhees (eds.), G.E.M. Anscombe (trans.), Oxford: Blackwell.

Jak dál v anotacích textových korpusů?

Jan Hajič, Eva Hajičová a Petr Sgall

ÚFAL MFF UK v Praze

1. Od začátku své odborné kariéry patří Karel Pala svým způsobem ke druhé generaci českých počítačových lingvistů: jedním z jeho učitelů (vedoucím diplomové práce a do svého zbavení školitelské funkce z politických důvodů v r. 1969 i školitelem) byl Petr Sgall, jak o tom svědčí např. společný článek i s dalšími kolegy (Sgall, Panevová, Piťha, Pala 1961). Pak nám na čas zmizel z očí, ale brzy se stal naším pojítkem s Masarykovou univerzitou, nejprve na její Fakultě filozofické, a po zřízení Fakulty informatiky na této nové fakultě. V posledních letech nás spojuje zájem o počítačové korpusy, a to nejen v oblasti softwarových nástrojů či morfologické a syntaktické analýzy, ale i v úvahách o možnostech zachycení úrovní blízkých k reprezentaci logické sémantiky (viz Pala, Materna a Zlatuška 1990).

Do této oblasti do jisté míry patří i úvahy, které nastiňujeme v tomto příspěvku. Formulujeme je zde spíš v podobě otázek, které si sami klademe, ale které podle našeho názoru směřují k dalšímu pokroku ve spolupráci teoretické, počítačové a korpusové lingvistiky, a týkají se také vztahu ligvistiky a sousedních oborů.

2. Anotační scénář Pražského závislostního korpusu (PZK; anotovaný korpus je nyní přístupný, viz Hajič a kol. 2006) vznikl na základě promyšlené koncepce tří úrovní anotace morfématické, analytické (povrchové) a tektogramatické (hloubkové) opřené o funkční generativní popis (Sgall 1967). Charakteristická a svým způsobem průkopnická je koncepce tektogramatické syntaktické roviny (TR) jako možného podkladu pro zachycení jazykového významu. I na této úrovni zůstává anotace na rovině syntaktických vztahů uvnitř věty (Mikulová a kol. 2005).

První otázka, kterou se zde chceme zabývat, se týká přechodu k anotaci mezivětných vztahů: **jde o úroveň „nadstavbovou“?** Jinými slovy, v intencích koncepce PZK: jde o úroveň – jistě ne ve smyslu jazykové roviny – „nad“ TR? Nebo o úroveň „vedle“ či „mimo“? Rozčlenění jazykového popisu do rovin, jak s ním pracuje fukční generativní popis (který je podkladem pro systém anotace v PZK), předpokládá, že toto rozčlenění se týká rozčlenění vztahu formy a funkce na vztah mezi jednotkami dvou sousedních rovin. Kdybychom tedy chápali úroveň zachycující mezivětné vztahy jako součast takové soustavy, měli bychom chápat nadvětnou jednotku jako funkci jednotky tektogramatické roviny, tedy věty. To je samozřejmě chápání zjednodušující, které pro charakteristiku nadvětného útvaru nestačí. Na druhé straně, při srovnávání možných vztahů mezi větami a mezi klauzemi (uvnitř jedné věty, ať již jde o spojení parataktické nebo hypotaktické) docházíme ke zjištění, že některé z těchto vztahů jsou velmi obdobné, viz zde bod 4. níže.

3. S předcházející otázkou (týkající se v podstatě „dimenze“ jednotek, tedy pohledu, který se liší od pohledu rozčlenění vztahu funkce a formy) souvisí i otázka vztahu mezi TR jako **rovinou jazykového významu a oblastí kognitivního obsahu**. Systematickému pohledu na tento vztah je věnováno několik statí Sgallových (srov. zejm. Sgall 1992; 1994; 2003; viz i knihu Hajičová, Partee a Sgall, 1998), jejichž závěry zde můžeme jen shrnout: Vycházíme ze stanoviska, že existuje velmi důležitý rozdíl mezi jazykovým významem a ontologickým obsahem, odrážející rozdíl mezi jazykem jako systémem

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 57–61, 2009.

a oblastí kognitivních vztahů. Pojem „význam“ lze chápat při nejmenším v šesti interpretacích: (i) význam jako jazyková strukturace („jazykový význam“, angl. *literal meaning*), (ii) význam, či lépe řečeno „smysl“ (angl. *sense*) jako jazykový význam obohacený o specifikaci reference (odkazování), tedy jako přechod od (podkladové) jazykové struktury k sémanticko(-pragmatické) interpretaci přirozeného jazyka; (iii) význam ve smyslu „strukturovaný význam“, tj. se specifikacemi jemnějšími, než jsou propozice, včetně rozlišení přenesených významů slov a jejich skupin); (iv) význam jako intenze (angl. *intension*); (v) význam jako extenze (angl. *extension*); (vi) význam jako obsah (angl. *content*), tedy s přihlédnutím ke kontextu obsahu promluvy. Tektogramatická rovina PZK odpovídá interpretaci (i); první kroky v dodatečných anotacích PZK zachycujících koreferenční vztahy (spadající do oblasti koreference gramatické i textové), srov. Nedoluzhko (2007), spadají pod interpretaci (ii), alespoň v některých jejích aspektech; jisté pokusy o uplatnění některých zřetelů zahrnutých pod interpretací (iv) lze nalézt u Nováka (2004; 2008). Zatím jsme však nedospěli k nějakému systematickému scénáři posunujícímu PZK k podrobnému členění oblasti kognitivního obsahu.

4. Pokud budeme uvažovat o PZK se zřetelem na jeho pokračování ve směru k anotační charakteristice diskurzu, jak jsme naznačili v otázce první (odst. 2), nabízí se srovnání s prvním koncepčně uceleným systémem **anotace diskurzních vztahů**, jak je postupně realizován pro angličtinu v tzv. Penn Discourse Treebank výzkumným týmem na Pennsylvánské univerzitě pod vedením prof. Aravinda Joshiho (PDTB 2007). PDTB vychází z předpokladu, že základem zachycení vztahů v diskurzu je konektor a jeho argumenty (v původní koncepci šlo vždy o dva argumenty, nyní se od tohoto omezení upouští). Konektorem se rozumí v prototypickém případě spojovací výraz, ale mohou být i konektory v textu nevyjádřené, implicitní. Přitom nemusí jít vždy o zachycení vztahů mezi dvěma oddělenými větami v textu, ale i vztahy mezi klauzemi v souvětí, ať už jde o vztahy parataktické nebo hypotaktické. Nabízí se tu tedy otázka, do jaké míry tyto vztahy už jsou reprezentovány na tektogramatické rovině PDT. Pokud jde o konektory uvnitř souvětí (popř. i o předložkové a prosté pády nominalizačních výrazů), zachycují se buď jako vztahy závislostní (v PZK se řídící sloveso zapuštěné klauze chápe jako závislé na řídícím slovesu nebo jiném členu klauze řídící) nebo vztahy klauzí v koordinaci. (Je třeba uvést, že tuto otázku nabízejí s odkazem na PZK sami američtí autoři, srov. Lee a kol. 2006). V PZK byl navíc pro vztahy mezi větami zaveden jako jeden z funktorů (tj. typů závislostního vztahu) funktor PREC; je přiřazován lexikálním uzlům, které syntakticky patří do dané věty, ale nemají vymezený významový vztah uvnitř této věty a odkazují k předcházejícímu kontextu (odtud název: PREC(eding)). V současné době se těmto otázkám věnují ve spolupráci s pennsylvánskými kolegy mladí pracovníci našeho Ústavu (Zikánová, připr.; Mladová 2008; Mladová a kol. 2009); bereme v úvahu jak možnou klasifikaci českých spojovacích výrazů (konektorů, ve smyslu PDTB) a podrobnější klasifikaci vztahů v PZK naznačených funktorem PREC, tak i vztahy, pro které není v anotovaných větách uzel ani hrana, tedy které jsou implicitní.

Zůstává ovšem otázka, nakolik závislostní vztahy uvnitř věty (jde především o vztahy volných doplnění, tedy časové, kauzální, atd.) mají stejný charakter jako vztahy mezi větami, popř. do jaké míry lze koordinační vztahy mezi klauzemi v jedné větě chápat stejně jako vztahy mezi oddělenými větami v textu. Tedy např. nakolik je časový vztah mezi řídící a závislou predikací ve větě *Když maminka přišla večer domů, hned si vzala zástěru a začala připravovat večeři.* sémanticky odlišný od časového vztahu mezi

samostatnými větami *Maminka přišla večer domů. Vzala si hned zástěru. Začala …* Stejně lze uvažovat o paralele mezi koordinací uvnitř věty (v souvětí) a mezi oddělenými větami. Podobná otázka vyvstává při zpracování úseků textu, uvnitř kterých se projevuje tzv. parcelace (v lingvistické literatuře chápaná často jako osamostatňování větných členů) jako např. v textovém segmentu *Ten pes pořád štěká. Kouše. A také škrábe.* nebo *Přišel pozdě. Asi tak kolem sedmé.* V současném anotačním schématu PZK se taková eliptická vyjádření chápou jako povrchové vypouštění uzlů, které jsou na tektogramatické rovině „rekonstruovány", tj. doplněny do podkladového závislostního stromu s příslušnými vztahy k uzlům řídícím (jako v těchto případech podmět i sloveso, a taky příslovce *pořád*, aspoň v jednom významu).

5. S těmito úvahami koncepčního charakteru souvisí i **otázka technická** – jaký formální objekt pro zachycení diskurzních vztahů zvolit. Zatím pracujeme s předpokladem jakýchsi „megastromů", tedy s propojením tektogramatických stromů do větších celků (náš anotační nástroj dovoluje až propojení 30 stromů před a za zkoumanou větou) nebo výhledově se strukturami „relačními". V každém případě však chceme reprezentaci diskurzu budovat na základě tektogramatických struktur, ať již obohacených, popř. uvnitř dále rozčleněných či propojených. A také se domníváme, že jsou dobré důvody pro rozlišení mezivětných vztahů od vztahů uvnitř vět, tedy pro zachování – ať již v jakékoliv (více či méně technické) podobě – hranice věty. Nejzávažnějším argumentem pro tento přístup je významová relevance aktuálního členění věty, např. vzhledem k dosahu kvantifikátorů a negace. V dnešní podobě PZK je aktuální členění věty anotováno na základě specifického atributu TFA nabývajícího hodnot kontextově zapojený uzel kontrastivní, kontextově zapojený uzel nekontrastivní a uzel kontextově nezapojený. Na základě přiřazení těchto hodnot je pak možné rozlišit, která část věty je jejím základem a která jejím ohniskem, a z toho pak vycházet i při stanovení dosahu kvantifikátoru a negace.

6. A další otázka, nikoli technická: potřebujeme stavět anotaci diskurzu na stromech tektogramatické roviny nebo ke stanovení diskurzních vztahů stačí **prostý text**? Jak naznačeno výše v bodě 5, je třeba vycházet ze struktury věty a nikoli jen z jejího textového zápisu. Stejně tak potřebujeme stavět na základě zachycení vztahů závislostních, jinak není možné klasifikovat vztahy zachycené funktorem PREC. Významy slov (základní i přenesené) a vztahy slov ve větách musíme rozlišovat i kvůli zachycení koreference v textu.

7. Řada specifických otázek se nabízí, rozšíříme-li okruh textového materiálu, který má většinou charakter monologu, na **dialog**. Odhlédneme-li (v této etapě) od problematiky spojené s automatickým rozpoznáváním mluvené řeči (*speech recognition*), jde především o dvě oblasti jazykových jevů: eliptické vyjadřování a s tím do jisté míry související určování koreferenčních či anaforických vztahů. V dialogu účastníci (zřejmě pod vlivem bezprostředního osobního kontaktu) často spoléhají na to, že adresát(i) ví/vědí, „o čem je řeč", a proto mohou mluvčí celé větné úseky vypustit. Při rekonstrukci výpovědí je pak třeba rozhodnout, do jaké míry je nutné vynechané části doplnit, a často také není návaznost jasná. Podobné problémy nastávají při sledování koreferenčních vztahů: v mluvené řeči (jejíž součástí prototypicky dialog je) se sice častěji užívá zájmenné reference (často ukazovacími nebo osobními zájmeny), ale někdy bez jednoznačnosti a odkazuje se i na celé segmenty promluvy, a při poměrně častém vypouštění je i koreference méně snadno určitelná.

8. Uvedený výčet otázek, které před námi na cestě od věty k mezivětným vztahům a od jazykového významu k mimojazykovému (popř. v metajazykových textech i jazykovému) obsahu stojí, zdaleka není úplný, řady dalších problémů jsme si vědomi, ale jistě je i mnoho takových, které vyvstanou až v průběhu vlastní anotace. To je ovšem jedna z důležitých motivací, proč se anotováním korpusů monolingválních i paralelních multilingválních dlouhodobě zabýváme: upozorní nás na jazykové jevy, jejichž popis se často v gramatikách ani v dosavadní lingvistické literatuře nenajde. Jejich zvládnutí je však potřeba jak k doplnění teoretických i empirických zjištění o jazyce, tak i pro budování pokročilejších aplikací počítačového porozumění přirozenému jazyku.

Reference

1. Hajič Jan a kol. (2006), Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia.
2. Hajičová Eva, Partee Barbara a Petr Sgall (1998), Topic-Focus Articulation, Tripartite Structures, and Semantic Content, Kluwer Academic Publishers: Dordrecht.
3. Lee Alan, Prasat Rashmi, Joshi Aravind, Dinesh Nikhil a Bonnie Webber (2006), Complexity of Dependencies in Discourse: Are dependencies in discourse more complex than in syntax? In: Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories, Eds. Jan Hajič a Joakim Nivre., Praha, 79-90.
4. Mikulová, M. et al. (2005), Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank: Annotation Manual. Universitas Carolina Pragensis, Prague.
5. Mladová Lucie (2008), Od hloubkové struktury věty k diskurzním vztahům. Diskurzní vztahy v češtině a jejich zachycení v anotovaném korpusu. Diplomová práce, FF UK.
6. Mladová Lucie a kol. (2009), Towards a Discourse Corpus of Czech. Příspěvek na konferenci Corpus Linguistics 2009, Liverpool.
7. Nedoluzhko Anja (2007), Anotace rozšířené anotace a asociativních vztahů (bridging) v Pražském závislostním korpusu. Technická zpráva, ÚFAL MFF UK.
8. Novák Václav (2004), Towards logical representation of language structures, The Prague Bulletin of Mathematical Linguistics 82, 5-86.
9. Novák Václav (2008), Semantic network manual annotation and its evaluation, The Prague Bulletin of Mathematical Linguistics 90, 69-82.
10. Pala Karel (1966), O nekotorych problemach aktulal'nogo členenija. In: Prague Studies in Mathematical Linguistics 1, Academia-Praha, 81-92.
11. Pala Karel, Materna Pavel a Jiří Zlatuška (1990), Logická analýza přirozeného jazyka. Praha: Academia.
12. PDTB (2007), The Penn Discourse Treebank 2.0 Annotation Manual.
13. http://www.seas.upenn.edu/∼pdtb/PDTBAPI/pdtb-annotation-manual.pdf
14. Sgall Petr (1967), Generativní popis jazyka a česká deklinace. Praha: Academia.
15. Sgall Petr (1992), Underlying structure of sentences and its relation to semantics. In: Festschrift für Viktor Rozencvejg (red. T. Reuther), Wiener Slawistischer Almanach. Sonderband 33. Vienna: Gesellschaft zur Förderung slawistischer Studien, 273-282. Přetištěno ve sborníku Petr Sgall: Language in its mulitfarious aspects, Praha: Karolinum 2006, 130-139.
16. Sgall Petr (1994), Meaning, reference and discourse patterns. In: The Prague School of Structural and Functional Linguistics. (red. P. Luelsdorff), Amsterdam – Philadelphia: John Benjamins, 277-309. Přetištěno ve sborníku Petr Sgall: Language in its multifarious aspects, Praha: Karolinum 2006, 335-361.

17. Sgall Petr (2003), From meaning via reference to content. In: Karlovy Vary Studies in Reference and Meaning (red. J. Hill a P. Koťátko), Praha: Filosofia Publications, 172-183. Přetištěno ve sborníku Petr Sgall: Language in its mulitfarious aspects, Praha: Karolinum 2006, 325-334.
18. Sgall Petr, Panevová Jarmila, Petr Piťha a Karel Pala (1961), Ze syntaktické analýzy češtiny (přípravné práce ke strojovému překladu), AUC – Philologica 3, Slavica Pragensia III, 181-196.
19. Zikánová Šárka (připr.), Discourse Annotation: Relations between Sense Tags in Penn Discourse Treebank and in Prague Dependency Treebank. Prague Bulletin of Mathematical Linguistics.

The Linguistic Double Helix: Norms and Exploitations

Patrick Hanks

Institute of Formal and Applied Linguistics, Charles University in Prague

1 Linguistic rules and linguistic data

In this paper I propose an approach to analysing the lexicon of a language that is driven by new kinds of evidence that have become available in the past two decades, primarily corpus data. In the course of developing this approach, much of which was undertaken in 2005-8 at the Faculty of Informatics, Masaryk University, Brno, it has been necessary to develop a new, lexically driven theory of language. This was inevitable because received linguistic theories proved inadequate: they were not up to the job of explaining observable facts about the way words are used to create meanings. In particular, patterns of linguistic behaviour are observable in corpus data that cannot be directly accounted for by standard linguistic rules, of the kind that govern compositionality.

There has been much confusion, misunderstanding, and even outright hostility about the relation between data and theory in linguistics, so it is necessary to be precise here. Corpus linguists object to invented examples; theoretical linguists question whether corpus data can reveal facts about language as system. I shall not delve into the theoretical linguists' objections, as these have been dealt with more than adequately elsewhere. A balanced summary can be found in Fillmore (1992). Instead, I will comment on the corpus linguists' objection. Except among a few extremists, this objection is not to the use of intuitions to interpret data, for how else could data be interpreted, other than by consulting intuitions? The objection is to the invention of data. As long ago as 1984, in a paper delivered at a conference on Meaning and Translation in Łódź, Poland (a paper that was eventually published as Hanks 1990), I argued that intuitions are a very poor source of evidence, because introspection tends to focus on less common uses, while the really common uses (e.g. the use of ***take*** with expressions of time, as in *It won't take long* and *It took three years to build*) are buried too deep in the subconscious of native speakers to be readily available for recall.

The argument to be developed here is that the so-called "mainstream" in linguistic theory in the late 20th century was out of focus, due to three factors:

1. the goal of explaining all possible well-formed utterances within a single monolithic rule system;
2. the speculative invention of evidence;
3. neglect of the lexicon and the ways in which people actually use words to make meanings.

Put together, these factors resulted in half a century of concentration on syntactic well-formedness, supported by intuitive judgements about the acceptability of invented sentences. Inventing evidence is a hard habit to break. It is insidious. It starts reasonably enough. If a language teacher wants to explain the importance of word order and

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 63–80, 2009.

prepositional phrases in English, what could be more innocent than making up an ordinary, everyday sentence such as *John asked Mary for a pen*? But when applied to speculation about the boundaries of possible usage, the practice of making up evidence has led to the invention of implausible and even misleading examples such as *The box was in the pen; The horse raced past the barn fell*; and *The gardener watered the flowers flat* (all invented by linguists speculating about possibilities rather than analysing data, and all demonstrably implausible in one way or another).

There are, of course, exceptions to this rather sweeping indictment of late 20th-century linguistic theory, notably the hard-nosed insistence of corpus linguists such as John Sinclair on the importance of collecting texts into corpora for use as evidence and the use of computational tools to analyse collocations and other phenomena statistically. This insistence has led inevitably to a demand for a bottom-up system of lexical rules that is both powerful and flexible.

An old-fashioned view of rules is that a rule is not a rule if it is flexible. But the observable facts of everyday language in texts, in corpora, and on the Internet compel us to the uncomfortable conclusion that linguistic rules are both immensely powerful and immensely flexible. Much of both the power and the flexibility of natural language is derived from the interaction between two systems of rules for using words: a primary system that governs normal, conventional usage and a secondary system that governs the exploitation of normal usage. Both these systems of rules are primarily lexical—i.e. rules for using words, rather than rules for constructing sentences. Of course syntactic rules have a role to play: there is interaction between lexis and syntax, but syntax must take second place. Why?

One reason for putting syntax in second place is presented by Wray (2002), who argues that much ordinary communication consists of familiar, conventional phrases and sentences and that, when uttering and understanding these, speakers and writers do not normally analyse them syntactically. Instead, they utter and understand the phrases (and even sentences) as a whole—ready-made, as it were. Wray shows that such ready-made phrases are far more common and widespread than was previously believed, and that, although they are formulaic, they are not "fixed". Operating on a "slot-and-filler" basis (replacing one word or phrase in a formula with another), language users can and do vary their stock of formulas to meet different requirements without necessarily building up utterances from first principles. People resort to syntactic analysis occasionally, but, according to Wray, on the basis of "needs-only analysis". When we want to say something entirely new, or when we hear or read a puzzling or complex sentence, we have the ability to analyse it syntactically. But the fact that we can do this does not entail that we do do it on all occasions. Life is too short and conversation is too quick and (mostly) too trivial to merit or need fundamental syntactic analysis, except in unusual circumstances.

There are other reasons, too, for putting syntax in second place in the analysis of meaningful language. Let us start with a simplified account of syntactic rules. The power of syntactic systems is undeniable, whether they depend on word order (as in English and French) or on inflection (as in Latin and Czech), or on some combination of the two (as in German). Simple rules can often be more powerful than complicated ones. An extremely simple classification of most of the words of many if not all languages is that some of them are nouns and others are verbs. One hugely powerful rule for making meanings is the SVO rule (subject – verb – object): you can take any two nouns, join them to a verb,

and (if you have a lot of nouns and several verbs) make a very large number of sentences in which the verb expresses a relationship between the two nouns. This is a very simple, very powerful rule. The conventional order may be SOV in some languages, or VSO or VOS, and under suitable conditions the order can be varied for rhetorical and other effects. In highly inflected languages such as Czech and Latin, inflections rather than word order determine clause roles of nouns in relation to a verb. That does not matter, as long as a language has some way or other of distinguishing subject from object. The crucial point here is the classification of words into parts of speech—verbs and nouns. Several of the possible sentences that result from the simple rule just mentioned – SVO – are actualized and used for some communicative purpose or other. Others remain no more than theoretical possibilities.

Complications begin to set in for this primitive account of linguistic rules when we observe that, for a sentence to be well formed, function words and/or inflections are needed. In English, in particular, nouns in most circumstances need a determiner, so the clause roles S and O are realized, not just by a noun but by a noun phrase (consisting basically of a noun plus a determiner, possibly with some other words such as adjectives, or even an embedded dependent clause, thrown in). Another complication is that a well-formed sentence does not always need two noun phrases—one noun phrase is sufficient with intransitive verbs. The complexities become exponentially greater as other parts of speech—in particular prepositions and adjectives—are added, each bringing with it its own set of rules. Nevertheless, all of these complexities can be expressed in a single monolithic rule system, even though such a system necessarily consists of a rather large number of rules. Most traditional grammars (including transformational and generative grammars) are monolithic rule systems of this kind.

However, a monolithic rule system such as the one just outlined, no matter how complex it may be, does not have the slightest chance of coming anywhere close to descriptive adequacy, i.e. of describing the realities of actual human linguistic behaviour. Great theorists of the past have attempted to deal with this mismatch by idealizing the language system (*langue*, competence) and distinguishing it from the everyday reality (*parol*e, performance). But these idealizations simply won't do. The exceptions to the rules are so numerous, and so obviously well motivated, that they cannot possibly be dismissed as mere 'performance errors'. Something else is going on.

J. R. Firth rejected Saussure's parole/langue distinction (as he would no doubt have dismissed the competence/performance distinction had he lived long enough to encounter it). Instead, he insisted on the close observation of actual linguistic behaviour. It is ironic, therefore, that close inspection of the textual traces of actual linguistic behaviour, looking at words in context within a neo-Firthian framework, compels us to the conclusion that the only satisfactory way of accounting for the observed facts is once again to postulate a duality. In this case, the duality is not between an idealized system and everyday reality, but rather between two interactive systems of rules governing linguistic behaviour: rules for norms and rules for exploitations. Without such a theory, perfectly well-formed, meaningful sentences such as 'I hazarded various Stuartesque destinations', 'Her eyelids yawn', 'Always vacuum your moose from the snout up', and 'Never invite two China trips to the same dinner party' – all attested in real data[1] – would either be inexplicable or

[1] These examples are discussed in more detail in Hanks (Forthcoming).

would require selectional restrictions set so wide that no meaningful study of collocations would be possible and therefore the investigation of meaning in language would be unable to proceed. The only reasonable conclusion is that "selectional restrictions" are not really restrictions at all, but rather preferences, and that preferences are rule-governed, but governed by a different set of rules from the rules that govern normal utterances. These rules yield probabilities, not determinations.

I am not the only person in recent years to have observed that collocations are preferences rather than restrictions. From a bottom-up perspective, it seems obvious. Only top-down theorists think in terms of restrictions. Equally obvious is the fact that people exploit normal usage for rhetorical and other effects. This fact has been observed in a great variety of realizations and discussed by various writers on language and meaning over the past two thousand years, dating back at least to the Roman teacher of rhetoric Quintilian (1st century AD), if not to Aristotle. What is new here is the theoretical status given to exploitations, releasing the theory and rules of well-formed normal usage from the need to account in the same breath for equally well-formed but abnormal usage. Also new—though obvious when you come to think about it—is the finding that the distinction between norm and exploitation is a matter of degree (some utterances are more normal than others). The methodology for determining the extent to which any given utterance is normal depends on statistical measurement of corpus or textual evidence. Two very common secondary rules—ellipsis and semantically anomalous arguments—may be counted as exploitations, although they are not found among the tropes and figures of speech discussed by classical rhetoricians: they are too mundane to count as rhetorical devices.

2 Genesis and summary of the theory

The theory of norms and exploitations (TNE) had its genesis in a marriage between lexicography and corpus linguistics. It is a bottom-up theory, created in response to the general question, how can we account for the ways in which people use words to make meanings? How can we account for the intuition shared by most empirical language analysts—strongly reinforced by observation of corpus data—that there are patterns and regularities lurking just below the surface of everyday usage? How does language really work, at the lexical, semantic, and pragmatic level? What are the general principles that govern word use, and what generalizations can be made about the relationship between word use and word meanings?

TNE proposes that, in natural languages, a set of rules governing the normal, conventional use of words is intertwined with a second-order set of rules governing the ways in which those norms are exploited. As its name suggest, TNE is a theory with two main components, but unlike many other theories, its two components are not sharply distinguished. Rather, they are poles at opposite ends of a cline. Some norms are more normal than others; some exploitations are more outrageous than others. And in the middle are alternations: lexical alternations, where one word can be substituted for another without change of meaning (for example, the idiom *grasping at straws* alternates with *clutching at straws*); syntactic alternations, of the kind described by Levin (1993); and semantic-type alternations, which are a device for selecting a different focus when describing what is basically the same event type (you can talk about *calming someone* or

alternatively, with a slightly different focus, about *calming someone's anxiety;* you can talk about *repairing a car* or you can focus on the presupposition and talk about *repairing the damage*).

First and foremost, TNE is a theory of prototypes and preferences, based on extensive analysis of actual traces of linguistic behaviour – what people say and what they may be supposed to mean – as recorded in large corpora. Analysing corpus data is an exercise in syntagmatics. The lexical analyst looks at large quantities of text data in various ways, using a variety of corpus-analytic tools such as a KWIC index (a concordance) and the statistical analyses of the Sketch Engine, and immediately perceives that there are patterns in the way the words are used. More thorough analysis reveals further patterns, hidden below the surface. The whole language is riddled with interconnecting patterns. But as analysis of corpus data proceeds, something very alarming happens: the patterns in a concordance which seemed so obvious and which caught the eye at first glance begin to seem more and more difficult to formalize, as more and more unusual cases are noticed. More and more exceptions show up as the data accumulates.

Different patterns are found at different levels of delicacy: discovering that what people *hazard* is usually *a guess* is a very coarse-grained discovery, easily made given a handful of corpus lines from a general corpus of English. At the other extreme, the discovery that *firing a smile at someone* is also part of a pattern (a conventional metaphor that extends also to clauses such as *She fired a shy glance at him*) is a more fine-grained, delicate discovery. At this level of delicacy, it is hard to know where to draw a line between normal and abnormal usage. More thorough examination of data leads to the conclusion that, usually, there is no line to be drawn—only a broader or narrower grey area. Likewise, it is sometimes hard to know where to draw a dividing line between any two patterns of normal usage. Nevertheless, it is usually easy to identify a few prototypical examples, around which other uses may be grouped. It seems from this that a new, prototype-based approach to linguistic formalisms needs to be developed.

When a word is associated with more than one pattern of normal use, it is usually but not always the case that different patterns activate different meanings. *Hazarding a guess* (= stating a proposition without confidence that it is true) activates a different meaning from *hazarding one's life* or *hazarding one's money at the roulette table* (= putting one's money or life at risk in the hope of some good outcome).

On the other hand, *firing a gun* and *firing at a target* have different patterns (syntactic structures) but activate the same basic meaning. The relationship between patterns and meanings is strong, but not straightforward. It takes many forms.

The other major component of TNE arises out of the observation that some uses of words are highly abnormal or unusual and do not fit into a pattern very well at all, and yet there is no reason to believe that they are mistakes. In fact, rather the reverse. Unusual expressions like *vacuuming a moose (from the snout up)* and *urging one's car through a forest* are communicatively effective and memorable precisely because they are unusual and stretch the boundaries of normal, patterned usage.

The principle governing pattern analysis in TNE is **collocation**: grouping collocates together. Different groups of word (lexical sets) have a preference for the company of certain other lexical sets, large or small. The lexical sets so grouped can in turn be mapped, as colligations, onto syntactic structures. Indeed, they must be so mapped in order to enable speakers to utter meaningful sentence at all—though not through any conscious

effort on the part of the speaker. The groupings are integral to the system that each speaker has internalized since birth (see Hoey 2005). Thus, meaning is dependent on lexical sets grouped as colligations, both according to their normal contexts and permitting exploitations of normal contexts.

The patterns associated with each word (strictly speaking, each content word) are complex because they do not merely relate to one another syntagmatically and paradigmatically; they also serve as representations of non-lexical cognitive entities, for example of beliefs about the world, of a speaker's subjective emotions, of stored recollections of reactions to past events, sensations, hopes, fears, expectations, and so on. At the same time, this complex mass of private attitudes and beliefs in an individual speaker's brain has to interact somehow with similar but not identical complex masses of private attitudes and beliefs in the brains of other users of the same language, for the whole purpose of language is communication—interaction with others—not merely the expression of private beliefs and sensations. Each content word in a language is like a huge railway station, with trains departing to and arriving from other words, other cognitive elements, and other speakers. We humans are not merely cognitive beings but also social creatures, and language is the instrument of our sociability. For this reason, the conventional patterns and uses of each content word in a language constitute a more or less complex linguistic gestalt. The gestalt for normal uses of the English verb *sentence* is very simple and straightforward, boiling down to one single pattern – *a judge sentences a convicted criminal to a punishment*. The gestalt for a verb such as *scratch* or *throw* is extremely complex, with a wide variety of syntagmatics, meanings, and pragmatic implicatures, which would take many pages to explain. The astonishing fact is that, somehow or other, all native speakers (including people with otherwise limited educational attainment) manage to internalize at least a substantial part of this gestalt for almost all common, everyday words, as well as many less common ones, depending on their particular interests and life circumstances.

The whole picture is further complicated by the necessary introduction of a diachronic perspective. Whether we know it or not, the language we use today is dependent on and shaped by the language of past generations. Most exploitations of norms are lost as soon as uttered, but every now and again one of them catches on and becomes established as a new secondary norm in its own right.

3 Theory and application

What applications can be envisioned for TNE? It is for others to judge how useful the theory is and how or whether they want to make use of it. Here I shall mention just three areas in which I believe that it has some relevance: natural language processing by computer, language teaching, and linguistic theory.

To take the first two of these, we may note that there are, broadly speaking, two main aspects to the practical application of any linguistic theory: productive applications and receptive applications. Productive applications use a theory to understand the creation of linguistic events, and even to create them: for example, to help language learners or computers to generate well-formed and relevant utterances. Receptive applications are designed to facilitate understanding: the computer or the human must understand what is being said in order to respond appropriately. (Appropriate responses include

learning—the assimilation or rejection of new information and the formation of new beliefs.) In both cases (production and reception), there is an underlying assumption that a linguistic theory serves as a basis for creating an inventory of linguistic items. The particular inventory predicted by TNE is an inventory of patterns associated with each content word in the language.

TNE can be seen as a tool for creating tools, for lexicographic resources themselves are tools for use in applications such as language learning by people, language understanding by people and machines, and language processing by computer. But of course the theory is a theory of language, not of tool building, so if it has any value, that value must be applicable directly in activities such as natural language processing by computer, language teaching, and literary studies (what Jakobson called 'poetics'). In all of these fields, it seems likely that a theory that focuses on normal language use, that has a special role for creativity, that refuses to be distracted by speculation about remote possibilities, and that insists on close empirical analysis of data has potential applications that will yield rich dividends. In addition to the applications just mentioned, the theory probably has something to contribute to cognitive science and our understanding of the way the human mind works, but that is not its main focus.

There are many areas in which the relevance of TNE could be discussed—for example grammar and grammatical theory, literary stylistics, cognitive linguistics, translation studies, and many others. In this section, I shall confine myself to sketching out a few comments in three major areas of potential application: computational linguistics, language teaching, and lexicography.

3.1 The Semantic Web, NLP, and AI

One motive for exploring new approaches to lexical analysis and develop a lexically based theory of language with a focus on normal usage is the current buzz of excitement surrounding the infinite possibilities of the so-called Semantic Web. The dream of the Semantic Web (see Berners-Lee et al. 2001) is to "enable computers to manipulate data meaningfully". Up till now (2009) work on realizing the dream has done little more than propose the construction of ontologies (see Chapter 1 above, section 1.8) and the addition of tags to documents and elements of documents, to structure them and improve their machine-tractability, without engaging with their semantic contents. It is a fair prediction that, sooner or later, if it is going to fulfil the dream of enabling computers to "manipulate data meaningfully", the Semantic Web will have to engage with natural language in all its messy imprecision. The stated aim of manipulating data meaningfully could, of course, be taken in any of a number of ways, depending on what counts as data. Current assumptions in the SW industry are that "data" means tagged data, and "manipulating data meaningfully" means little more than matching patterns and processing tags. However, Berners-Lee et al. (2001) also said:

> Web technology must not discriminate between the scribbled draft and the polished performance.

This would seem to be a clear indication that the original vision, though vague, included being able to process the meaning and implicatures of free text. But how is this to be done?

The protagonists of the Semantic Web drama, who are nothing if not canny, have avoided getting embroiled in the messy imprecision that underlies the ordinary—and sometimes precise—use of words in ordinary language. The Semantic Web's RDF (Resource Description Framework) confines itself to using and processing HTML tags and strictly defined technical terms. Insofar as ordinary words are assigned strict definitions for computational processing, scientific research, rules of games, and other purposes, they acquire the status of technical terms and are no longer part of ordinary language. Technical terms are essential for many logical, technological, and computational applications, unless they are stipulatively defined, they cannot be used to say new and unusual things or to grapple with phenomena that have previously lain outside the scope of the imagination of the definer, which is one of the most important things that can be done with ordinary language. The notion that the words of human language could all be rigorously defined was a dream that tantalized great thinkers of the European Enlightenment, in particular Wilkins and Leibniz. Their disgust with the fuzziness of word meaning was shared by philosophers up to Russell, and was indeed a factor in the latter's breach with Wittgenstein, who invited us to "look and see" what is actually going on when people use words to make meanings. But the vagueness and indeterminacy that Wilkins, Leibniz, and Russell (among others) considered to be faults in natural language may now be seen as essential design features. Sooner or later, the Semantic Web must engage with this design feature, the imprecision of natural language, if it is to fulfil its own dream. The theoretical approach to the lexicon outlined in TNE lays part of the foundation for such an engagement. This dream cannot be fulfilled without an inventory of the content words of a language, describing their normal patterns of usage and implicatures of each pattern, together with sets of rules that govern exploitations and alternations and procedures for matching usage in free text preferentially onto the patterns. This is a remote dream at the time of writing, but it does not seem unachievable in principle.

If this dream is to be fulfilled, it seems important to proceed methodically, step by step (in the right direction, of course) and to abandon—or at least suspend for the time being—the yearning for instant solution by a magic bullet that is so typical of computational linguists.

Semantic Web research is not the only computational application that stands to benefit from a long, hard look at how the lexicon actually works. In recent years, 'knowledge-poor' statistical methods in computational linguistics have achieved remarkably—some would say astonishingly—good results, at a coarse-grained level, in applications such as machine translation, message understanding, information retrieval, and idiomatic text generation. At the same time, refined methods based on syntactic and valency theory have yielded largely disappointing results. The same is true of methods based on using machine-readable versions of dictionaries that were designed for human beings. However, statistical methods, in principle, have a ceiling, while deterministic methods point to the need for a reappraisal of the relationship between lexis and syntax. TNE points a possible way forward, toward an integration of statistical and deterministic methods. Some procedures in computational linguistics and artificial intelligence—'knowledge-rich' approaches—still lean heavily on linguistic theories that are not empirically well-founded and lexical resources that are based more on speculation and intuition than analysis. Whether it acknowledges it or not, the computational linguistics community will continue to encounter fundamental difficulties, at least insofar as the serious analysis

Table 1. predictability of meaning based on Latinate morphemes.

A	B	C
In-	-script-	-ion
Pre-	-vent-	-ive
De-	-vict-	-ible
Con-	-duc-	
Pro-		

of meaning is concerned, until it starts to build and use lexical resources that are based on empirical analysis of actual use of language. Any strategy other than bypassing meaning entirely (which is what statistical methods do) will need a theoretical approach of the kind outlined by TNE.

Consider predictions of meaning in English and French words based on Latin morphology, as in Table 1. Most Romance languages contain sets of words that consist of one item taken from column A and one from column B and one from column C. However, the system is not totally productive or predictable. There can be gaps here and there, e.g. there is no word **deviction*, although, if someone chose to invent such a word, its meaning should be to some extent predictable on the basis of this table. Moreover, the meaning of lexical items can be non-compositional. There is nothing in Latin morphology to explain why ***prescription*** has something to do with doctors and drugs in English, rather than writing something in advance. And this non-compositional meaning of collocated morphemes may or may not carry over to some other Romance language.

Much the same applies to rules and phraseology. A rather trivial but telling anecdote seems relevant. Some years ago I was involved with a software company doing, among other things, information retrieval. When asked to retrieve information about "nursing mothers", the search engine retrieved vast quantities of information about care homes for the elderly.[2] This was, of course, wrong. Dictionaries do not say so, but in English the specific meaning of the collocation ***nursing mother*** is non-compositional. It means a mother who has recently had a baby and is in the phase of feeding it with milk from her breasts (rather than from a bottle). This expression yields 19 hits in BNC. Technically, syntactic analysis suggests that it could mean something else. In actuality, it does not.

A lexical resource built on the principles of TNE would show the specific normal meaning, 'mother who is breast-feeding', of this collocation and would treat it as a single lexical item, contrasting it with the normal, compositional meanings of the verb ***nurse***. This verb normally means 'tend (a sick or injured person)'. It also means 'harbour (bad feelings): *nurse a grievance, nurse a grudge*'. The sense 'feed (a baby) at the breast' is in almost all current dictionaries, but in actual usage this is not really compositional, for it is vanishingly rare (only 3 hits in BNC). What's more, although *child* is a close synonym of *baby*, the expression *nursing a child* is not used as a synonym for breast-feeding. If a mother is nursing her child, the child is sick or injured. This is not a matter of certainty

[2] It should be pointed out that this particular search engine application was aiming to "break the tyranny of text matching" (in the words of Greg Notess).

based on syntactic analysis (which gets it wrong); it is a matter of statistical probability based on collocational analysis.

Now multiply this anecdote by some number in the hundreds of thousands, and you will have some idea of the number of semantic traps that lurk in waiting for computational linguists. Clever algorithms solve many problems, but in matters of the relationship between word use and word meaning, clever algorithms create more problems than they solve.

3.2 Language learning, language teaching, and the lexicon

In broad brushstrokes, the next most important area of applied linguistics in terms of money spent and number of people affected, after applications in computational linguistics and artificial intelligence, is language learning. Literally hundreds of millions of people at any given time are currently learning one or more foreign languages. Language teaching is big business, world wide. Some learners are very proficient and seem to be able to pick up other languages with apparent ease, regardless of the teaching methods are used. Others struggle mightily. But even the proficient ones welcome well-organized help, while badly organized help can add to the struggles of the less proficient. Moreover, it seems that different learning strategies suit certain individual learners better than others. A few gifted individuals respond well to an emphasis on formal grammar, which used to be fashionable; others respond better to an emphasis on analogy and 'communicative competence'. Even apparently irrelevant factors such as the student's personal goals (short-term and long-term), the personality of the teacher, the commercial strength of different language communities, the vibrancy of different cultures, and even the beauty of the countryside, can play a part in motivating learners. TNE cannot, of course, help with any of the motivating factors just mentioned, but it does have a contribution to make in helping teachers, syllabus designers, course-book writers, lexicographers, and learners themselves to get the lexicon in perspective, make an organized selection, and to give a high priority in their teaching and learning to the most normal patterns of usage associated with particular words. In other words, it can help with a focus on lexical relevance.

This idea is not new, of course. It is what A. S. Hornby and his colleagues (Gatenby and Wakefield) tried to do in their *Idiomatic and Syntactic Dictionary* (ISED; 1942)—a remarkable work, subsequently re-published as the *Oxford Advanced Learners' Dictionary* and greatly inflated in its second and subsequent editions. For a fuller discussion, see Hanks (2008).

Language teachers have a problem with the lexicon. There is simply too much of it. Learning phonology and syntax in a classroom environment can yield valuable generalizations for learners comparatively rapidly, applicable to vast swathes of language. But as far as the lexicon is concerned, what can be taught? It is harder to make a case for "getting the words in" (Bolinger 1971) than for teaching syntax. Isn't the lexicon just a vast list of "basic irregularities", with no predictability, "an appendix of the grammar", as Bloomfield (1933) famously remarked? If so, getting the words in can, indeed must, be left to happenstance.

Even if we agree with Bolinger that the words must be 'got in', it remains to be decided precisely what should be got in. Any language learner is faced with the daunting task of learning how to use many thousands of words in a language in order to be able to

make meanings and even more if they are to understand what is said. As this book has shown, many words constitute daunting challenges of complexity in themselves.

Against this is the obvious necessity to learn at least some of the meaning potential and usage patterns of at least some words in order to be able to use a language at all. Slightly less obvious is the impossibility, even at a theoretical level, of learning *all* the words of a language. Selectivity is essential. Even less obvious, at first glance, is the impossibility of learning all possible uses of a given word. To those who use them, words seem so simple, so obvious. Surely they must be constrained by clear-cut finite boundaries of meaning and usage? But the awful fact is that such boundaries are not clearly defined. They are fuzzy and complex, and the full power of a word as a linguistic gestalt is sometimes of awesome complexity. These, in reality, are among the problems facing the unfortunate learners of a language. They not only have the immense problem of productive usage—generating idiomatically well-formed utterances in a language whose conventions are different—often subtly different, full of traps—from those of their native language; they also have to prepare themselves for receptive usage. And on the receptive side, learners never know quite what will be thrown at them. Who knows what a native speaker is going to say next? The argument of TNE is that this latter point is true, but nevertheless it is possible to predict probabilities, set up defaults, and focus attention on interpreting normal phraseology.

It has become fashionable in some places to provide learners with corpus-access tools such as WordSmith or the Sketch Engine and let them loose on a large corpus without further ado. The motivation for this practice is highly commendable: bringing learners face to face with the realities of actual usage and engaging them collaboratively in the process of learning how words are used and in solving their language problems. In every classroom I have ever visited where this is done (even—or perhaps, especially—if done chaotically), the excitement of engaging with real data and trying to solve real problems is palpable. However, without good principles of selection and organization of data, it can lead to disappointment and even confusion. Guidance is needed on such matters as what to expect from raw corpus data, how to select and sort corpus data, and how to deal with the unexpected. TNE offers a theoretical basis for developing this kind of guidance. In short, learners looking at raw data need not only to be encouraged to inspect the data thoroughly and look for patterns, but also to be informed about principles for interpreting data, to be prepared for the complexities of ordinary usage, to expect exceptions to the patterns, and to be shown how to make effective hypotheses about what the various patterns mean.

Recently, there has been a revival of interest in the lexical approach to language teaching. Pioneers of the "lexical approach" to language teaching were Sinclair, Willis, and Lewis. Following Sinclair (1988), Willis (1990) proposed a ‘lexical syllabus’ for language learning. Sinclair and Willis were writing in the very earliest days of corpus linguistics, when a corpus of 18 million tokens was regard as large and before the full enormity of the challenge posed by corpus data to established theories had even begun to be recognized. Willis’s proposal was implemented as the Cobuild English Course. It must be acknowledge that, despite its innovative approach, the Cobuild English Course was not a huge success. Why was this? No doubt part of the reason was bad marketing and off-putting presentation by the publisher: a web search reveals comments on such things as "cognitive overload" (`http://www.usingenglish.com/forum/`) and a general sense

that the pages are unpleasantly cluttered. Such problems could be easily fixed. More germane reasons may have included the non-existence of a systematic body of research into what counts as a pattern and the absence of reliable information about the relative frequency of different patterns and senses. Willis's approach to a lexical syllabus was also hampered by the absence of a thoroughly worked-out theoretical distinction between patterns in which a word can participate and patterns in which it normally participates.

In the same stream, Lewis (1993) argued that "language consists of grammaticalized lexis, not lexicalized grammar" and that the interests of language learners are seriously impaired by excessive concentration on teaching grammar rather than lexis. Lewis's lexical approach concentrates on developing learners' proficiency with lexis—i.e. words and syntagmatics—and 'chunks' of formulaic language of the kind that was subsequently to be discussed more fully in Wray (2002).

It seems a matter of obvious common sense that lexical research should contribute to syllabus design. Words and patterns of word use are far too important to be left to happenstance or the whims of individual teachers. For almost all groups of learners, it is absurd to give a high priority to teaching such terms such as ***umbrella, overcoat, hat,*** and ***cloakroom attendant***. But then, many words that deserve a high priority in a lexical syllabus, e.g. ***need, search, hope, look, find,*** are semantically complex: not only the words themselves but also the most normal uses of such words needs to be prioritized. When we examine them, the issues regarding the lexical contribution to syllabus design turn out to be rather complex. At least the following points must be taken into account:

- Integration with other approaches to syllabus design
- The distinction between function words and content words: function words should, perhaps, be considered as part of a grammatical component of a syllabus, while only content words are organized into the lexical component
- The role of pro-forms: learners have a higher-than-normal need for effective use of semantic pro-forms such as ***thing, something, anything,*** and ***do***, to help them fill lexical gaps and achieve fluency
- The relative frequency of different phraseological patterns and senses of polysemous words: selectivity is just as important at this microstructural level as at the macrostructural level of the lexical component
- Pragmatic functions of lexical items such as ***broadly, you know, I think, it seems***.

A lexical syllabus is not a magic bullet. The different interests, goals, and abilities of different individuals and different groups of learners are relevant and need to be taken into account by individual teachers working within a general framework of lexical selection. At a more general level, the best pedagogical approaches to a lexical syllabus have must necessarily understate the rich complexity of the *possible* uses of each word, while in language teaching more generally, prioritization has all too often been left to common sense and happenstance, or to the intuitions of the teacher, which are typically skewed towards boundary cases, for reasons which have been discussed throughout this book. Part of the argument here is that each word in a language has a core set of one or more prototypical uses, which can be discovered only by painstaking lexical analysis. Some prototypical uses are general; others tend to be domain-specific. Each prototypical use is associated with a prototypical meaning. Prototypical uses can be exploited in regular ways. All of these facts can and should be part of the foundations for prioritization of a

lexical syllabus, based on relevant corpus evidence. A lexical syllabus goes hand in hand with a grammatical syllabus, and both need to be empirically well founded.

3.3 Electronic lexicography

TNE was sired by lexicography upon corpus linguistics, and it would not be a runner at all if, as its owner and trainer, I did not believe that it could be entered in the Language Theory Stakes as a potential winner[3]. It has, I believe, the potential to inspire new directions in electronic lexicography. The preceding two main sections have both mentioned 'resources'. Among the new resources that need to be developed for all such applications and no doubt many others, are corpus-driven pattern dictionaries.

Pattern dictionaries TNE is an essential foundation for a new kind of dictionary which, on the basis of corpus analysis, will report the patterns of usage most associated with each word (strictly speaking, each content word) in a language. The great advantage of such a dictionary is that, for activities such as natural language processing by computer, it enables meanings to be attached to patterns, rather than to the word in isolation. This facilitates pattern matching. Thus, if a word has more than one sense, a pattern will have already identified the conditions under which each sense is activated before any attempt is made to state the meaning and do anything with it. The sense is 'anchored' to the normal phraseology with which each sense of each word is associated.

The first such dictionary is already in progress at the time of writing. It is the *Pattern Dictionary of English Verbs* (PDEV: `http://nlp.fi.muni.cz/projects/cpa/`), an on-line resource, in which each entry consists of four components:

1. The verb lemma together with a list of the phraseological patterns with which it is associated, expressed in terms of argument structure and subargumental cues.
2. The primary implicatures associated with each pattern (roughly equivalent to a dictionary definition, but 'anchored' to the arguments in the pattern, rather than floating freely, as dictionary definitions tend to do).
3. A training set of actual uses of each verb illustrating its use in each pattern, taken from the British National Corpus
4. A shallow hierarchical ontology of the semantic types of nouns (see Pustejovsky et al. 2004), populated with a lexical set of nouns to which each argument is related. As far as the data permits, nouns are related to argument patterns of verbs according to their semantic type.

Compiling a pattern dictionary—indeed, compiling any dictionary—is a long, slow process. At the time of writing (June 2009), approximately 10% of PDEV is complete, after four years work. At the current rate of progress, if there is not a substantial injection of funds to build up a professional lexicographic staff, the project will not be completed until 2040, when the author will be 100 years old. However, one of the great benefits of online publishing and internet access is that such work can be published as work in progress.

[3] To the uninitiated, it should be explained that this sentence is an extended and somewhat contrived horse-racing metaphor of a peculiarly British kind.

PDEV is an exploration of one possibility for practical implementations of TNE. A project with some similarities to PDEV is FrameNet, which is based on Fillmore's theory of Frame Semantics. Both PDEV and FrameNet are pointers to new future roles for lexical analysis. However, there are important differences. Some of them are as follows.

- FrameNet expresses the deep semantics of a number of prototypical situations (frames) associated with different lexical items. PDEV investigates syntagmatic criteria for distinguishing different meanings of polysemous words, in a 'semantically shallow' way, using semantic types as a grouping mechanism.
- FrameNet proceeds frame by frame and analyses situations in terms of frame elements. PDEV proceeds word by word and analyses patterns of use of individual words (verbs).
- FrameNet studies differences and similarities of meaning between different words in a frame, with no systematic attention to polysemy. PDEV studies differences of the relationship between usage and meaning for each polysemous verb.
- FrameNet does not analyse corpus data systematically, but goes fishing in corpora for examples in support of hypotheses. PDEV is driven by a systematic analysis of corpus data and provides statistically valid data for the comparative frequency of a verb's different meanings.
- There is considerable overlap between closely related frames in FrameNet; it does not seem to have clear criteria for distinguishing frames, and does not seem to be aiming at an inventory of all possible (or all normal) frames. The number of frames seems to be open-ended. This perhaps relates to the impossibility of postulating that the world is organized into neat and finite hierarchies of semantic frames.
- PDEV attempts to group observed lexical sets in a hierarchical ontology, according to a) their shared co-occurrence, and b) their shared semantic types. This raises interesting theoretical and practical issues, which cannot be discussed here; they could be the subject of a whole separate book.
- FrameNet does not seem to have any criterion for completeness. Exploration of FrameNet's frames reveals that many of the lexical items that ought to be members of a particular frame have been missed. Of course, once this is noticed, case by case, it is easy to bring additional lexical items into a frame, but as things stand (June 2009), there is no way of telling whether either a frame or a lexical item has been fully analysed.

Historical pattern dictionaries Literary scholars need to ask, not only in what respects does the phraseology used by a great writer of the past differ from the normal, conventional phraseology of present-day English, but also in what respects it differs from the normal, conventional phraseology of the language of his or her own time. Unfortunately, historical records of spoken language are non-existent before the 20th century invention of recording devices. However, for many periods in English, among other languages, written records of ordinary language survive in sufficient quantities to make a pattern dictionary of the language of that period a theoretical possibility. Moreover, the writings of great writers themselves are not ruled out as evidence for patterns, insofar as their usage overlaps with the usage of other writers of the same period, for proof of pattern depends on extrapolation from many sources, including sources which may include (elsewhere

in them) idiosyncrasies of usage. Translating the theoretical possibility of a historical pattern dictionary into a practical reality would, of course, depend on the usual necessary combination of scholarly interest and funding.

The masses of evidence collected and analysed over the past century and a half by the great historical dictionaries such as OED would be a valuable resource for a historical pattern dictionary of this kind, but it needs a stronger theoretical foundation. In principle, the OED evidence, which is substantial, could be reanalysed to give an account of the normal, shared phraseological patterns in use at any given period in the history of the language. In practice, it would be desirable to supplement the OED citation evidence for word use in each period with a corpus of whole texts of the same period, especially a corpus containing informal, non-literary texts (bulletins, broadsheets, journals, private letters, and suchlike). This is desirable because, inevitably and quite properly, OED has a literary bias and because it is based largely on citations collected by human citation readers, not on corpus analysis. As Murray (1878) noted, human citation readers have a natural tendency to select citations for rare words and unusual uses and to overlook common, everyday, familiar words and uses. It would be an odd citation reader who copied out all the uses of the lemmas ***give*** or ***take*** in a text being read for citations. But it is precisely the common, everyday uses of words such as ***give*** and ***take*** that a pattern dictionary concerns itself with and for which corpus technology can provide evidence 'at the press of a button'. A further issue concerns the comparative frequency of patterns. A citation can prove the existence of a word, phrase, or sense at a given period, but only a corpus (ideally, a balanced corpus of 'representative' texts) of the same period can give an approximate idea of the relative frequency of different patterns of use of the same word.

4 The broader picture

TNE is closer to Tomasello's account of human cognition (1999, 2003) than to Leibnizian primitives or Chomsky's predicate logic. In a series of studies, Tomasello and his colleagues compared the developmental behaviour of human children with that of other primates (chimpanzees, etc.). On this basis, he argues that what distinguishes humans from apes is the ability to recognize other members of the species (conspecifics) as intention-governed individuals. This makes possible shared purposeful actions (cooperative behaviour) and prediction of the likely actions and reactions of others. In other words, his conclusion is that what distinguishes humans from other primates is the ability to put oneself in others' shoes. Language plays a crucial role in this ability. It is, therefore, a biological and cultural phenomenon rather than a mathematical one. He says:

> The understanding of conspecifics as intentional beings like the self is a uniquely human cognitive competency that accounts, either directly on its own or indirectly through cultural processes, for many of the unique features of human cognition.
> —Tomasello (1999: p. 56).

Tomasello argues that there simply has not been enough time, in evolutionary terms, for these unique features to have developed by genetic evolution. There must be another explanation – and there is, namely cultural transmission.

When an intelligent ape or other mammal makes an important discovery—for example, how to use a stick as a tool—the discovery is useful to that individual, it may be

remembered and repeated by that individual. It may even be imitated by other members of the species in the same clan, pack, or social group. However, individuals of nonhuman species have no means of sharing, recording, and transmitting their discoveries, so that sooner or later each dscovery is lost.[4] When a human makes a discovery, on the other hand, it is (or can be) disseminated throughout the community, not lost, because humans have a mechanism for sharing and storing the knowledge gained. This mechanism is language. It operates what Tomasello calls "the ratchet effect': faithful dissemination and storage of knowledge acts as a rachet, preventing backward slippage that would cause knowledge to be lost. This has enabled *Homo sapiens* to evolve at an astonishing speed compared with the genetically bound evolution of other species.

Thus, human linguistic behaviour is cooperative social behaviour. It involves, among other things, the sharing of knowledge. The relevance of all this to TNE lies in the Gricean mechanism. In order to communicate, a human relies on the ability of other members of her species (her conspecific interlocutors) recognizing her intention to communicate, together with an underlying body of shared communicative conventions to encode the message. These shared conventions are words and phrases and their meanings. TNE shows how these conventions work and provides a theoretical framework for compiling an inventory of the conventions in any given culture on which successful communication depends.

Thus, TNE provides a basis for explaining, within Tomasello's Darwinian model and Grice's theory of conversational cooperation, what the shared conventions of linguistic behaviour in any given community are and how they are flexible enough to encompass and develop novel ideas and novel situations s well as repetition of the norm.

Tomasello goes on to argue that the diversity of human language is too great to be accounted for by the Innate Universal Grammar hypothesis: there just are not enough linguistic universals to explain anything of any great interest about the rich, culture-specific complexities of human linguistic behaviour. Again, the explanation lies in cultural transmission.

5 Conclusion

This chapter has proposed a contribution to the empirical foundations for a lexical theory of language, pointing towards new ways of exploring meaning in text and conversation, the development of new research methodologies, and new insights into relationships between lexis and grammar, lexis and cognition, and lexis and the world. It will encourage a reappraisal of all pre-corpus theories of language and abandonment of the sloppy habit of inventing evidence to support research.

At the same time it points to the need for a fresh approach to studying the relationship between logic and analogy. A human language is a curious mixture of logical and analogical processes. Tidy-minded thinkers such as Wilkins (1668), Leibniz, and Russell tended to regard the analogical aspect as a fault, but more careful observers such as Wittgenstein and Rosch have laid foundations for an approach that regards the analogical

[4] Such discoveries may, of course, be rediscovered independently by other members of the species at other times and other places.

aspect as an essential design feature. Corpus-driven lexicology can build on these foundations.

The title of this chapter is intended to mark a new departure. Corpus linguistics—approaching language through the analysis of patterns of lexis observable in corpus and textual data—is in its infancy, for the simple reason that corpora large enough for this purpose did not exist until about 20 years ago. The first astonishing finding of corpus linguistics has been an apparent contradiction: the regularities are much more regular than most pre-corpus linguists expected, while the irregularities are much more irregular. The theory of norms and exploitations outlined here shows how this apparent contradiction can be reconciled.

Systematic corpus analysis of the whole lexicon of all languages is called for, leading to new lexicons and new grammars. Some aspects of existing linguistic theories will receive confirmation from such an exercise; others will have to be jettisoned.

The Theory of Norms and Exploitations is a response to the challenges posed by corpus data, offering a contribution to the study of meaning in language. It can be seen, if you like, as a step towards making explicit the nature of the conventions on which interlocutors rely in expecting to be understood and to understand—i.e. the Gricean mechanism of conversational cooperation. It is to be hoped that much future linguistic research will be bottom-up, driven by empirical analysis of lexis, and will focus on exploring the nature both of conventions and of alternating probabilities. In this way, new light can be expected to be shed on the nature of human behaviour and human linguistic creativity.

References

1. Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. 'The Semantic Web' In: *The Scientific American*, May 2001.
2. Bloomfield, Leonard. 1933. *Language.* Holt, Rinehart, Winston.
3. Bolinger, Dwight. 1970. 'Getting the words in'. In *American Speech*, 45. Reprinted in Raven I. McDavid and Audrey R. Duckert (eds., 1973), *Lexicography in English,* New York Academy of Sciences.
4. Fillmore, Charles J. 1992. '"Corpus linguistics" or "Computer-aided linguistics"?' In: J. Svartvik (ed.): *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82.* Stockholm, August 1991.
5. Hanks, Patrick. 1984 [1990]. 'Evidence and intuition in lexicography'. In: Jerzy Tomaszczyk and Barbara Lewandowska-Tomaszczyk (eds.), *Meaning and Lexicography*. Benjamins.
6. Hanks, Patrick. 2008. 'Lexical Patterns: from Hornby to Hunston and beyond' (the Hornby Lecture). In: E. Bernal and J. de Cesaris (eds.) P*roceedings of the XIII Euralex International Congress*. 9 Sèrie Activitats 20. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.
7. Hanks, Patrick. Forthcoming. *Analyzing the Lexicon: Norms and Exploitations*. MIT Press.
8. Hoey, Michael. 2005. *Lexical Priming: a New Theory of Words and Language*. Routledge.
9. Jackendoff, Ray. 2002. *Foundations of Language*. Oxford University Press.
10. Levin, Beth. 1993. *English Verb Classes and Alternations.* University of Chicago Press.
11. Lewis, Michael. 1993. *The Lexical approach*. Hove: Language Teaching Publications.
12. Murray, James. 1878. Presidential Address to the Philological Society.
13. Pustejovsky, James, Anna Rumshisky, and Patrick Hanks. 2004. 'Automated induction of sense in context'. In *COLING 2004 Proceedings*. Geneva

14. Sampson, Geoffrey. 2001. *Empirical Linguistics.* Continuum.
15. Sinclair, John. 1988. 'A lexical syllabus for language learning'. In M. J. McCarthy and R. A. Carter (eds.) *Vocabulary in Language Teaching*. Longman.
16. Tomasello, Michael. 1999. *The Cultural Origins of Human Cognition*. Harvard University Press.
17. Tomasello, Michael. 2008. *Origins of Human Communication.* MIT Press.
18. Wilkins, John. 1668. *Essay towards a Real Character, and a Philosophical Language*. The Royal Society, London. Excepts reprinted in Hanks (ed., 2008): *Lexicology: Critical Concepts in Linguistics.* Routledge.
19. Willis, Dave. 1990. *The Lexical Syllabus*. HarperCollins.
20. Wray, Alison. 2002. *Formulaic Language and the Lexicon.* Cambridge University Press.

Přínos bohemistického pracoviště filozofické fakulty k rozvoji počítačové lingvistiky na brněnské univerzitě

Zdeňka Hladká

Filozofická fakulta Masarykovy univerzity, Brno

Abstrakt Příspěvek v přehledu zaznamenává účast bohemistů z Filozofické fakulty Masarykovy univerzity na formování matematicky a počítačově zaměřené lingvistiky v rámci brněnské univerzity, a to od počátků této orientace v 60.-70. létech 20. století, kdy sehrálo významnou iniciační roli jubilantovo působení na Katedře českého jazyka, slovanské a obecné jazykovědy FF UJEP, přes 80. léta, kdy se do počítačového zpracování přirozeného jazyka zapojuje Klára Osolsobě, a 90. léta, kdy Zdeňka Hladká začíná v ÚČJ FF MU vytvářet jazykové korpusy, až do současnosti. Zdůrazněna je spolupráce Ústavu českého jazyka FF MU s Centrem zpracování přirozeného jazyka FI MU. Pozornost je věnována i aspektům výukovým.

K formování matematicky, počítačově a korpusově orientované lingvistiky na brněnské univerzitě od samých počátků přispívali – ne sice masivně, zato však soustavně – i bohemisté z filozofické fakulty. Zásadní úlohu v utváření této tradice sehrála jubilantova osobnost. Karel Pala se od svého raného působení na Katedře českého jazyka, slovanské a obecné jazykovědy na Filozofické fakultě Univerzity Jana Evangelisty Purkyně[1] v 60. letech 20. století iniciačně zasloužil o šíření matematické lingvistiky nejen na tomto, ale i na dalších fakultních pracovištích, později jako člen fakultní a univerzitní komise pro výpočetní techniku prosazoval využívání počítačů ve výuce i výzkumu (první dva počítače pro katedru českého jazyka získal v r. 1988, tedy v době, kdy byla výpočetní technika ještě považována za potenciální nástroj nepřátelské diverze), před svým odchodem na Fakultu informatiky Masarykovy univerzity v r. 1995 vychoval v duchu nastoupené orientace bohemistku Kláru Osolsobě a také v dalších letech všestranně podporoval kontakty brněnských bohemistů a informatiků při řešení úkolů spjatých s počítačovým zpracováním češtiny.

60.-80. léta 20. století

Počátky zájmu o matematickou lingvistiku na katedře českého jazyka FF UJEP jsou jednoznačně spjaty s příchodem Karla Paly, který zde v roce 1964 po absolvování dvouletého studijního pobytu v oddělení matematické a aplikované lingvistiky Ústavu pro jazyk český ČSAV v Praze nastoupil na asistentské místo. Badatelsky se Pala soustředil na analýzu jazyka v oblasti syntaxe, sémantiky a pragmatiky. (Ze spolupráce s logikem Pavlem Maternou, anglistou Alešem Svobodou a informatikem Jiřím Zlatuškou vzešly

[1] Dnes Ústav českého jazyka Filozofické fakulty Masarykovy univerzity.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 81–84, 2009.

např. studie Materna, P. - Pala, K. - Svoboda, A.: *An Ordered-Triple Theory of Language.* Brno Studies in English 12, Brno 1976, s. 159-186; Materna, P. - Pala, K. - Svoboda, A.: *An Ordered-Triple Theory of Language Continued*. Brno Studies in English 13, Brno 1979, s. 119-165; Materna, P. - Pala, K. - Svoboda, A.: *Externí a interní pragmatika*. In: Otázky slovanské syntaxe IV/1, Brno 1979, s. 53-60; v následujících letech i monografie Materna, P. - Pala, K. - Zlatuška, J.: *Logická analýza přirozeného jazyka*, Praha 1989.) Od druhé poloviny 70. let se Pala v součinnosti s programátory Vysokého učení technického v Brně a později Ústavu výpočetní techniky při Přírodovědecké fakultě UJEP podílel na prvních experimentech v oblasti automatického porozumění přirozenému jazyku, především na pokusech testujících možnosti automatické syntaktické analýzy češtiny (programový systém Wander).

Také ve výuce Pala seznamoval studenty bohemistiky a dalších oborů s matematickou a počítačovou lingvistikou a moderními lingvistickými směry opírajícími se o exaktní metodologii (v l. 1964-1971 vedl např. přednášku Úvod do matematické lingvistiky, v l. 1974-1989 přednášku a seminář Úvod do počítačové a matematické lingvistiky), i když si později postěžoval, že „snažit se učit studenty a studentky filozofické fakulty nějakým matematickým popisům jazyka, nebo je dokonce učit programovat, je nevděčná práce“.[2]

Od konce 80. let, po získání základní výpočetní techniky pro katedru českého jazyka, se Pala výrazněji zaměřil na výzkum a popis češtiny s využitím počítačů, mj. s cílem vytvořit algoritmický popis české morfologie, strojový slovník češtiny a automatický pravopisný strojový korektor. Na těchto úkolech se na bohemistickém pracovišti začala významně podílet Palova žačka Klára Halasová (Osolsobě), která se soustředila na formální analýzu české morfologie (např. Osolsobě, K.: *Algoritmický popis české formální morfologie substantiv a adjektiv*, SPFFBU, A 37-38, 1989-1990, s. 83-97; později Osolsobě na toto téma obhájila i doktorskou práci *Algoritmický popis české formální morfologie a strojový slovník češtiny*, Brno 1996). Společně s programátorem Stanislavem Francem pracovali na integrovaném morfologickém analyzátoru *klara,* který využíval jazyk Prolog a aparát DC gramatik (např. Pala, K. - Halasová, K. - Franc, S.: *Česká morfologie a syntax v PROLOGu*. In: Sborník semináře SOFSEM 1987, Bratislava 1987, s. 38-42). Klára Osolsobě s Karlem Palou spolupracovala také při přípravě několika učebních textů (*Základy výpočetní techniky pro filology,* Brno 1989; *Základy počítačové lingvistiky,* Brno 1992). Výrazem pokračující výukové orientace na počítačovou analýzu jazyka bylo v r. 1990 i zřízení semináře počítačové lingvistiky (v rámci Ústavu českého jazyka FF MU), jehož se stal Karel Pala vedoucím.

Od 90. let do současnosti

Lingvisté bohemistického pracoviště brněnské filozofické fakulty (Karel Pala, Klára Osolsobě, Mirek Čejka a později Zdeňka Hladká) se od samého počátku podíleli také na aktivitách směřujících k vytvoření korpusových zdrojů češtiny a etablování korpusové lingvistiky v českém prostředí. V r. 1988 spolu s pražskými lingvisty a matematiky vytvořili Iniciativní skupinu pro přípravu počítačových korpusů a slovníků, která dala v začátku 90. let impuls k vybudování „Počítačového fondu češtiny“ a v r. 1994 stála

[2] Citováno z interview Davida Povolného s Karlem Palou *Snažím se lépe poznat, jak funguje přirozený jazyk*, muni.cz / únor 2009, s. 7.

u založení Ústavu Českého národního korpusu, jehož úkolem bylo a je koordinovat tvorbu reprezentativního korpusu českých textů. Bohemisté Filozofické fakulty Masarykovy univerzity se na tvorbě Českého národního korpusu podíleli jednak vytvořením dvou jeho částí (viz např. Hladká, Z.: *Zkušenosti s tvorbou korpusů češtiny v ÚČJ FF MU v Brně*, SPFFBU, A 53, 2005, s. 115-124), jednak účastí na přípravě nástrojů umožňujících automatickou morfologickou analýzu spisovných textů i textů s mluvenostními rysy.

Korpusy češtiny vytvořené v Ústavu českého jazyka Filozofické fakulty MU ve spolupráci s programátory Fakulty informatiky MU

Pracovníci a studenti Ústavu českého jazyka FF MU se budování Českého národního korpusu zúčastnili nejprve přípravou **Brněnského mluveného korpusu** (BMK, součást ČNK od r. 2002; garant Zdeňka Hladká), který obsahuje elektronický přepis 250 magnetofonových nahrávek z let 1994-1999 zachycujících v neformálních i řízených rozhovorech 294 mluvčích. Velikost korpusu je cca 600 tisíc pozic. BMK byl ve snaze o kompatibilitu pořizován v souladu se zásadami Pražského mluveného korpusu, částečně diferenční postupy týkající se přepisu nahrávek (nahrazení tradiční interpunkce interpunkcí „pauzovou“ a zachycení simultánnosti dialogických promluv) byly později v ÚČNK akceptovány i pro tvorbu dalších korpusů mluvené češtiny. V rámci ČNK je BMK uložen jen v čisté neoznačkované podobě, na FI MU však byla vytvořena pracovní morfologicky označkovaná verze. Při její přípravě byly dosavadní nástroje morfologické analýzy, určené primárně k rozpoznávání spisovné češtiny, upravovány pro automatické rozpoznávání substandardních forem běžné mluvy, což lze v českém prostředí označit za významný průkopnický čin. Konkrétně byl upravován morfologický analyzátor ajka, který vznikl na FI MU v r. 1999 v rámci diplomové práce Radka Sedláčka (program se opírá o algoritmický popis české formální morfologie vypracovaný Klárou Osolsobě). Úpravy analyzátoru se snažily vyrovnat jednak s velkou hláskovou, tvarovou i lexikální variabilitou brněnské mluvy (prolínáním dialektických, interdialektických, obecněčeských i spisovných podob), jednak s některými obecně mluvenostními rysy (zjednodušováním souhláskových skupin, nedokončováním slov apod.). Na tomto úkolu se kromě odborníků z FI MU podíleli i lingvisté z Ústavu českého jazyka FF MU, zejména Klára Osolsobě a její žačka, později též pracovnice FI MU, Dana Hlaváčková (viz např. Hlaváčková, D.: *Korpus mluvené češtiny*, diplomová práce na FF MU, Brno 1998; Hlaváčková, D.: *Korpus mluvené češtiny z brněnského prostředí a jeho morfologické značkování*, SaS 62, 2001, s. 62-70; Hlaváčková, D. – Osolsobě, K.: *Morfologické značkování mluvených korpusů, zkušenosti a otevřené otázky.* In: Čeština v mluveném korpusu 1, Praha 2008, s. 105-114). Zkušenosti získané při úpravách ajky byly později využity a rozšířeny o nové poznatky při morfologickém značkování textů Korpusu soukromé korespondence.

Korpus soukromé korespondence (KSK) vznikal v Ústavu českého jazyka FF MU od 90. let 20. století (pod garancí Zdeňky Hladké). Obsahuje 2 tisíce ručně psaných dopisů a 1 tisíc e-mailů z let 1990-2005; shromážděné texty reprezentují 3 tisíce idiolektů. Velikost korpusu v části ručně psaných dopisů je cca 940 tisíc pozic, v části e-mailů cca 220 tisíc pozic. Korpus je detailně sociolingvisticky a poprvé v historii české korpusové lingvistiky i nářečně označkován. V rámci ČNK je od r. 2006 veřejně přístupná pouze část obsahující ručně psané dopisy (KSKdopisy). V komplexnosti byly shromážděné dopisy i e-maily zveřejněny v práci Hladká, Z. a kol.: *Čeština v současné soukromé korespondenci. Dopisy, e-maily, SMS*, Brno 2005, která kromě korpusového zpracování korespondence přináší i plné textové podoby dopisů a e-mailů a digitální fotokopie originálů. Jako doplnění navíc zaznamenává v textovém zpracování cca 2 tisíce SMS. V publikaci je zveřejněna i pracovní verze morfologicky označkované první půlky KSKdopisy. Automatické morfologické značkování probíhalo opět ve spolupráci Ústavu českého jazyka FF MU a příslušných pracovišť FI MU (především Centra zpracování přirozeného jazyka). Automatickou morfologickou

analýzu v tomto případě komplikovala vedle prolínání spisovného kódu s kódy nespisovné češtiny též vysoká expresivita a také pravopisná nestandardnost textů (viz např. Hlaváčková, D. – Sedláček, R.: *Morfologické značkování korpusu soukromé korespondence*. In: Varia XIV, Bratislava 2006, s. 371-379).

V polovině 90. let odešel Karel Pala z Ústavu českého jazyka Filozofické fakulty MU na Fakultu informatiky MU. (Ještě předtím na FF MU v r. 1993 obhájil habilitační práci na téma *Počítačové zpracování češtiny* a v souvislosti s vývojem jazykové podpory textových editorů mj. připravil s Janem Všianským *Slovník českých synonym*, Praha 1994[3]). Palův odchod však neznamenal konec počítačově zaměřené lingvistiky v ÚČJ FF MU. Naopak jím zprostředkovávané kontakty mezi bohemistickým pracovištěm FF MU a specializovanými pracovišti FI MU tento směr spíše posilují. Probíhá spolupráce jak v rovině výzkumné (z bohemistů se jí účastní zejména Klára Osolsobě, která se v poslední době věnuje převážně slovotvorné problematice, např. Pala, K. – Osolsobě, K. – Hlaváčková, D. – Šmerk, P.: *Formální vztahy české morfologie a slovotvorby*, předneseno na konferenci Čeština ve formální gramatice, Brno 2009), tak v rovině praktické tvorby korpusů češtiny a jejich značkování (viz výše; v nedávné době vznikl ve spolupráci bohemistek Zdeňky Hladké a Lucie Rychnovské a informatiků Pavla Rychlého a Pavla Šmerka např. korpus korespondence Bedřicha Smetany) a v neposlední řadě i v oblasti pedagogické. Už v 1. ročníku se studenti češtiny v rámci Úvodu do studia českého jazyka aktivně účastní sběru a zpracování materiálu pro přípravu korpusových zdrojů (aktuálně pro korpus ORAL), ve výběrových kurzech (např. Úvod do korpusové lingvistiky, Počítače a přirozený jazyk, Lingvistický software, Praktická cvičení z korpusové lingvistiky) se seznamují se základy počítačového zpracování přirozeného jazyka, učí se pracovat s korpusy, využívat lingvistický software vyvíjený na Fakultě informatiky MU apod. Na výuce uvedených disciplín se vedle bohemistů mateřského pracoviště podílejí i učitelé z FI MU (Dana Hlaváčková) a naopak bohemisté z FF MU se v roli konzultantů nebo oponentů doktorských prací (Klára Osolsobě) účastní výukových aktivit na FI MU. V současné době se také uvažuje o možnosti otevření mezifakultního oboru, v němž by se tradiční studium češtiny spojilo se základy počítačového zpracování přirozeného jazyka.

Zájem bohemistů z Ústavu českého jazyka FF MU o matematickou a počítačovou lingvistiku našel výraz i ve zpracování přehledové studie o rozvoji těchto disciplín v českém prostředí (Osolsobě, K.: *Matematická lingvistika.* In: *Kapitoly z dějin české jazykovědné bohemistiky*, Praha 2007, s. 447-466).

[3] Zkušenosti získané při práci na tomto slovníku Pala později uplatnil při práci na projektu EuroWordnet.

Stupňování sloves

Jarka Hlaváčová

ÚFAL, Matematicko-fyzikální fakulta, Univerzita Karlova v Praze

Prolog

V zimě jsem jela do Itálie lyžovat. Na sjezdovky. Vytáhla mě Marie, byl to její nápad, že si společně zalyžujeme. Na sjezdovkách jsem nelyžovala už několik let, tak jsem z toho měla strach. Ale nakonec to nebyl takový problém. Rozlyžovala jsem se poměrně rychle. První den jsme si jen tak zvolna polyžovaly. Další dny jsme se už nalyžovaly víc. Poslední den jsem už měla pocit, že jsme se vylyžovaly hodně. Už jsme musely odpočívat, abychom se neulyžovaly.

Předpona + zvratná částice

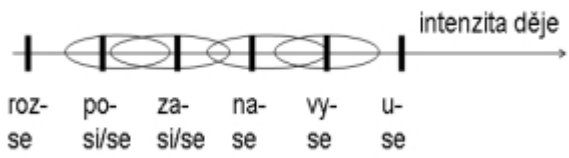
Mnoho nedokonavých (avšak ne iterativních) sloves má schopnost spojovat se s některými speciálními předponami a se zvratnou částicí *se* nebo *si*, a tím vytvářejí celé paradigma nových slovních tvarů s poměrně přesně definovaným významem.

Předpony, k nim příslušející zvratné částice a význam celého prefigovaného slovesa ukazuje tabulka:

Předpona	Sloveso	Zvratná částice	Význam
roz-	X	se	začít X
po-		si/se *	X v klidu, většinou příjemně
za-		si/se *	X po delší dobu a užít si to
na-		se	hodně X
vy-		se	hodně X, s víceméně kladným výsledkem
u-		se	X až do vyčerpání

Hvězdička (*) v tabulce znamená, že pokud jde o reflexivum tantum, zůstává i po přidání těchto prefixů zvratná částice *se*. Dosadíme-li v tabulce místo X např. sloveso *lyžovat,* dostaneme sadu nových paradigmat se zvratnou částicí.

Jednotlivá prefigovaná zvratná slovesa je možno s určitou tolerancí uspořádat podle intenzity děje, jak ukazuje následující obrázek:



Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 85–90, 2009.

Krajní body tvoří předpony *roz-* a *u-*, uprostřed je podle intenzity posloupnost *po-*, *za-*, *na-* a *vy-* s vágním až překrývajícím se rozsahem.

Z tohoto důvodu nazývám tento způsob tvoření s jistou nadsázkou pracovně „stupňování“ intenzity slovesného děje.

Příklady z korpusu SYN

1 roz-

= začít

Někde v tom mlází se znova rozťukal datel
Promnul si prsty, samým vzrušením se mu rozbrněly.

2 po-

= v klidu, většinou příjemně, s potěšením

místečko ve stínu, kde by si každý druhý pejsek rád pěkně pochrupkal
pohrál si se startovacími klapkami na křídlech

3 za-

= po delší dobu a užít si to

o jejich tématech si budou moci rovnou i zachatovat.
Poté zašel do tělocvičny, aby si podle svého rituálu ještě zaposiloval.
Na druhou stranu : trochu si čas od času zašílet v takovéto nevinné záležitosti patří v této nespravedlností a stresem napěchované době skoro k léčebným procedurám.

4 na-

= hodně (často s intenzifikátorem)

Člověk se hrozně naběhá.
matinka zatim doma vykládala, co se nastará a naběhá,...
Co jsem se jen natančila to léto

5 vy-

= hodně, s víceméně kladným výsledkem

vyplačte se do sytosti
V pohodě se tam celé dny můžeme do sytosti vyjezdit

6 u-

= až do vyčerpání

To by se asi norští fanoušci uslavili k smrti.
jak bylo zjištěno, unudit se nikdo nemůže
málem se uštěkal

Starší zmínky

Šmilauer (1971) se v této souvislosti zmiňuje o „míře děje“ (viz str. 181-182). Tímto termínem však nepopisuje přesně to, co já. Mezi předpony vyjadřující míru děje zahrnuje i jiné (např. *pře-*, *o-, pod-*, *pro-*). Tyto předpony skutečně ve spojení s určitými slovesy vyjadřují stupeň intenzity děje, ovšem ne zcela pravidelně, a tedy nespolehlivě. Stejná předpona může mít (a vždy má) mnoho významů, v závislosti na slovese, ke kterému se připojí.

Šmilauer tedy popisuje čistě sémantickou stránku. Jeho prefigovaná slovesa navíc nemusí být zvratná, kromě těchto tří výjimek: *na- se*, *po- si* a *za- si*. Význam těchto předpon ve spojení se zvratnou částicí hodnotí podobně jako já (v závorce jsou originální Šmilauerovy příklady):

po- velká míra („úplně“) s významem „hodně, dlouho, dosyta“
(*to jsme si pohověli, popovídali, pokouřili*)
za- velká míra s významem „s chutí se věnovat“
(*zacvičil si, zatančili si, zakouřili si*)
na- velká míra, často s určením míry (tolik, něco)
(*Tolik jsme se nasmáli, navtipkovali, naskotačili a navztekali. ... tam jsem se něco nachodil, nalelkoval se až hanba*)

Teprve zvratná částice totiž dodá slovesu konkrétní a velmi přesně vymezený význam, který jsem uvedla v tabulce. Tvrdím, že tento význam je u každé předpony neměnný, tedy připojením k libovolnému slovesu se jeho význam modifikuje stejným způsobem.

Havránek a Jedlička se též na str. 242-243 zmiňují o zvratných předponových slovesech, která označují „intenzitu děje“ s příklady (*najíst se, napít se, vyspat se, prospat se, vyskákat se, dokopat se*) nebo „různou fázi průběhu děje“ (*roznemoci se, rozejít se, sejít se, rozprchnout se*).

Některé jejich příklady považuji za sporné: např. *sejít se*. Jiné jsou podle mého názoru opět závislé na konkrétním připojeném slovese, předpony zde nemají všeobecně spolehlivý význam (*prospat se, dokopat se*).

Porůznu najdeme zmínky i u **Trávníčka**, v oddíle o slovesném vidu.

Nejlépe popisuje předpony z obrázku a z tabulky **Kopečný**.

První zmínka je v §18 na str. 23 až 24, kde popisuje sémantický rys „dějové míry“. Uvádí zde „augmentativní typy“ *nadělat se, nasedět se*, dále *posedět si, poležet si, počíst si*, které přebírá ze starší práce Šmilauera (1946).

Slovesa s předponou *roz-* charakterizuje sémantickým rysem „fázovost“. Mimo jiné praví: „Pro fázi začáteční se udává jako hlavní morfologický prostředek předpona *roz-*, spojená se změnou základního slovesa ve sloveso reflexivní: *rozepsat se, rozležet se, rozplakat se*.“

Dále se o uvedených předponách zmiňuje na str. 110 v kapitole nazvané Perfektiva tantum. Zařazuje je mezi tzv. „afektivní typy“ spolu s několika dalšími slovesnými předponami.

Nejpodrobnější komentáře jsou v kapitole České slovesné předpony, kde podává stručný přehled významů jednotlivých českých slovesných předpon. Vybírám ty, kterých se týká „stupňování“:

roz-

O jednom z významů předpony *roz-* Kopečný píše: „rozproudění činnosti, obyčejně až po dosažení náležité míry."

Jestliže k tomu přidáme požadavek zvratnosti, dostáváme spolehlivě význam jediný, uvedený výše.

po-

Krátce si Kopečný všímá i uvedeného významu předpony *po-*, o kterém píše, že může znamenat i „velkou míru děje", což je jen částečně ve shodě s mým pozorováním, viz příklady výše.

za-

Kopečný se zmiňuje i o předponě *za-*, jíž přiřazuje význam „vzplanutí děje, jeho začetí" a spojuje ji „s pocitem malé míry", což nám nepřipadá zcela přesné. To platí jen o první části jím uvedených příkladů, kde se předpona *za-* připojuje k dokonavému slovesu (*zablesknout se, zastesknout si* a další). Příklady uvedené tam pod písmenem b) už takto charakterizovat nelze (*zabásnit si, zalyžařit si, zalhat si*). Ani moje příklady z korpusu tomuto hodnocení nenasvědčují.

na-

Výrazům s předponou *na-* (*namodlit se, nasmát se* apod.) říká Kopečný augmentativnost, případně také intenzitivnost.

vy-

U předpony *vy-* zařazuje Kopečný zvratná slovesa do skupiny s významovým odstínem „vyčerpání děje", což je charakteristika podobná té mojí. Uvádí ještě případ zdvojené předpony *vyna-*, která však nezapadá zcela do mojí pomyslné škály.

u-

Předponu *u-* hodnotím shodně jako Kopečný, když říká: „Reflexivní typ upracovat se je téměř paradigmatický". Já považuji za paradigmatické všechny právě vyjmenované typy.

V kapitolce „Předponové typy paradigmatické" na str. 133 Kopečný uvádí 6 předpon, které považuje za paradigmatické. Z toho 4 (*na- se, po- si, u- se* a *za- si*) se shodují s mým hodnocením.

Další paradigmatické předpony jsou podle Kopečného *do-* („vyjádření absolutního zakončení") a nezvratné *vy-* („vyčerpání děje"). Vzhledem k tomu, že tyto předpony mají ještě jiné významy, nezařazuji je do své (paradigmatické) stupňovací škály.

Kopečný tedy zahrnul výše uvedené charakteristiky mezi významy, ale ne vždy specifikoval požadavek zvratnosti, který uvedené předpony vymezuje velmi přesně.

Jinak řečeno, jestliže se k nedokonavému slovesu připojí jedna z uvedených předpon a přidá se zvratná částice podle tabulky, alespoň jeden z významů výsledného slovesa bude ten, který je uvedený v tabulce.

Je zřejmé, že žádná předpona nemá jediný význam. Spolu se zvratnou částicí je však význam uvedených předpon velmi pravidelný.

Lemma „stupňovaného slovesa"

Přestože se pomocí prefixu a reflexiva vytvoří nové sloveso s celým paradigmatem, umisťujeme tato slova do paradigmatu neprefigovaného (základního) slovesa. Pokládáme zde prefixaci za tvoření slovesného tvaru, nikoli za slovotvorbu. Jinými slovy: tvrdíme,

že všechny takto utvořené tvary mají společné lemma. Konkrétně v příkladě krátkého vyprávění z Prologu je lemmatem všech podtržených slovních tvarů sloveso *lyžovat*.

Důvodů je hned několik.

Předně je to velká produktivita. Není sice pravda, že takto lze vytvářet celou sadu od každého nedokonavého slovesa (protipříkladem budiž třeba tvar **zadotýkat se*), přesto lze takto vytvořit velké množství nových slov. Navíc předpony v těchto slovech mají VŽDY stejný význam. Tento význam je naznačen v tabulce a vyplývá též z příkladu v Prologu.

Předpony se mohou přidávat i k předponovým slovesům. Nedávno jsem v nějakém televizním rozhovoru zaslechla větu:

Přijeli jsme si do města zanakupovat.

Kdybychom tyto slovní tvary lemmatizovali jako samostatné předponové sloveso (případně i s reflexivní částicí), museli bychom pro každé takové slovo zavést nové lemma, někdy homonymní s lemmatem již existujícím. Jestliže ho místo toho zařadíme do paradigmatu příslušného lemmatu bez předpony, můžeme se spolehnout na jeho přesnou interpretaci, včetně nutné přítomnosti reflexivní částice. Toho lze využít jak při morfologické analýze, tak při syntéze.

Některá takto vytvořená slova již ve slovní zásobě existují, ale mají jiný význam. Je to např. tvar *zamávat*, ovšem bez zvratné částice, z dalšího vymyšleného krátkého příběhu:

Včera se na nádraží pořádal kompars na nový film. Měli jsme za úkol mávat na odjíždějící vlak. Když dal režisér pokyn, rozmávali jsme se. Nejdřív to vypadalo, že si pomáváme a půjdeme domů. Scéna s máváním se však mnohokrát opakovala, takže jsme si zamávali víc, než se nám líbilo. Namávali jsme se opravdu hodně, vymávali jsme se do sytosti. Měli jsme strach, že se umáváme k smrti.

Příkladem úplné homonymie, včetně zvratné částice, je *vysmát se*. Běžný význam je zřejmý z příkladu:

Budu na něj hodná a on se mi pak vysměje.

Ale vyskytují se i příklady ve významu, který popisuji zde:

Stavil jsem se tady jen proto, že se tady člověk může v klidu vysmát.

I z těchto případů je zřejmý rozdíl významu. V první větě jde o výsměch, zatímco ve druhé o smích. Další rozdíl spočívá ve valenci. Zatímco v prvním příkladě jde o sloveso s dativní valencí, druhý příklad je intranzitivní. To však není pravidlem v jiných případech.

Podobné je *usmát se*, které může vypovídat buď o úsměvu, nebo opět o smíchu. I zde je rozdíl ve valenci (*usmát se* na koho – *usmát se* (bez valence)).

Výskyt stupňované intenzity není zpravidla vysoký, ale najdou se výjimky. Některá takto vytvořená slova jsou naopak velmi běžná, i s uvedeným významem, např. *rozesmát se*. V těchto případech je ovšem rozumné předponovou odvozeninu přímo zahrnout do slovníku.

Rozhodnutí, která prefigovaná (stupňovaná) slovesa jsou běžná, lexikalizovaná, a kde jde jen o okazionalismy, je samozřejmě značně obtížné. Velmi pravděpodobně v tom nebude panovat shoda, navíc se názory budou měnit v čase. Pro současné morfologické slovníky je nejspíš nejlepší konzervativní řešení, tedy ponechat ve slovníku ta paradigmata, která tam jsou, včetně zavedené lemmatizace, ale nepřidávat globálně nová. Vycházíme z toho, že současné slovníky již naprostou většinu běžných slov obsahují.

Rozhodně do morfologického slovníku patří ta slovesa, která nejsou zvratná nebo jsou tranzitivní, a dále potom tzv. odvozená reflexiva. Např. *utancovat se* je odvozené

reflexivum od nezvratného tranzitivního slovesa *utancovat* (koho). Lemma *utancovat* by se tedy do morfologického slovníku zahrnout mělo.

Reference

1. Havránek Bohuslav, Jedlička Alois: Česká mluvnice. Státní pedagogické nakladatelství. Praha 1981.
2. Kopečný František: Slovesný vid v češtině. Nakladatelství Československé akademie věd. Praha 1962
3. Šmilauer Vladimír: Novočeské tvoření slov. Státní pedagogické nakladatelství. Praha 1971.
4. Šmilauer Vladimír: Slovesný vid a způsob slovesného děje (První hovory o českém jazyce. Praha 1946.)
5. Trávníček František: Mluvnice spisovné češtiny. Část II. Slovanské nakladatelství, Praha 1951.
6. Český národní korpus – SYN. Ústav Českého národního korpusu FF UK, Praha 2005. Dostupný z WWW: <http://www.korpus.cz>

Doporučení místo závěru

Tak, a teď můžeš jít slavit. Rozeslav se zvolna. Velmi pravděpodobně sis už trochu poslavil, ale určitě sis ještě nezaslavil s každým. Užij si to slavení, naslav se podle libosti, vyslav se do sytosti. Jenom si dej pozor, aby ses neuslavil!

Počet lemmat v synsetech VerbaLexu

Dana Hlaváčková

Centrum zpracování přirozeného jazyka
Fakulta informatiky Masarykovy univerzity
hlavack@fi.muni.cz

1 Úvod

V Centru zpracování přirozeného jazyka na Fakultě informatiky Masarykovy univerzity je od roku 2005 budována databáze českých slovesných valenčních rámců VerbaLex. [1] Data jsou organizována ve valenčních rámcích *základních* – vlastní rámec se syntaktickou a sémantickou rovinou zápisu a *komplexních* – strukturovaný celek s dalšími informacemi o slovesech. Na rozdíl od jiných valenčních slovníků a databází, kde je standardně základním organizačním prvkem slovesné lemma[1], jsou ve VerbaLexu valenční rámce přiřazovány k celým slovesným synonymickým řadám – synsetům (z angl. *sets of synonyms*). Synonymické řady jako řídící prvek struktury databáze byly zvoleny z důvodu blízké návaznosti VerbaLexu na sémantickou síť WordNet (WN)[2], která má obdobnou strukturu. Konkrétní spojitost spočívá ve třech bodech:

- v první fázi práce na databázi byly valenční rámce zapisovány k českým synonymickým řadám přímo do české podoby WordNetu (Czech WN – CzWN);
- sémantická rovina anotace valenčních rámců ve VerbaLexu vychází z vrcholové ontologie vytvořené v rámci projektu EuroWordNet (EWN) [3] a využívá přímé odkazy na literály (lemma s číslem významu) v Princetonském WordNetu (PWN);
- většina českých synsetů byla v minulém roce nalinkována na anglické ekvivalenty v PWN, tzn. českým synsetům ve VerbaLexu bylo přiřazeno identifikační číslo (WNID) anglické synonymické řady. [4]

Kromě valenčních rámců s morfologicko-syntaktickou i sémantickou rovinou zápisu zachycuje valenční databáze řadu dalších informací o českých slovesech – definice jejich významů, záznam schopnosti sloves tvořit opisné pasivum, zachycení některých typů reflexivity, značení slovesného vidu, informaci o užití slovesa v přirozeném kontextu a jeho zařazení do slovesné sémantické třídy. Aktuálně VerbaLex zachycuje 10 478 slovesných lemmat, 19 360 valenčních rámců, 21 123 různých slovesných významů a 6 287 synsetů. Výchozím zdrojem jazykových dat je pro valenční databázi Slovník povrchových rámců BRIEF [5] obsahující celkem 15 tisíc sloves a téměř 50 tisíc valenčních rámců. Databáze je dostupná v textovém formátu, v převodu do jazyka XML, v pdf verzi a pod webovým prohlížečem (html)[2].

[1] Např. Slovník povrchových valenčních rámců BRIEF (FI MU), Vallex (MFF UK) [6], Slovesa pro praxi. Slovník nejčastějších českých sloves. [7]

[2] Formáty byly převzaty z valenčního slovníku Vallex (MFF UK) a upraveny pro potřeby VerbaLexu.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 91–99, 2009.

2 Vysoký počet lemmat v synsetech

V současné době je valenční databáze editována s cílem odstranit některé duplicitní záznamy a sjednotit nekonzistence, obojí vzniklé v důsledku manuální anotace VerbaLexu. V souvislosti s úpravami vyvstávají některé otázky a problémy, které dosud nebyly systematicky řešeny. Jednou z nich je problematika příliš dlouhých synsetů, tj. synonymických řad s velkým počtem slovesných lemmat. Jedná se o problém čistě uživatelský, na strojové zpracování jazykových dat v databázi nemá výraznější vliv. Dlouhé synsety jsou problémem hlavně pro orientaci uživatele v databázi (špatná přehlednost velkého množství dat) a také při jejich zobrazování ve webovém prohlížeči (rozdělení synsetu na dva řádky nebo nutnost používat posuvník pro zobrazení celého řádku). V tomto příspěvku se pokusíme odpovědět na otázky, čím je délka synsetu ve VerbaLexu způsobena (stylem zápisu, chybou anotátorů, pestrostí češtiny) a jaká řešení problému je možné nabídnout (synset odstranit, rozdělit, zvolit jiný způsob zápisu). Pro potřeby tohoto příspěvku jsme použili vzorek 50 nejdelších synsetů ve VerbaLexu, jejichž délka se pohybuje od 11 do 26 lemmat.

3 Organizace synsetu

Při volbě nejvhodnějšího způsobu notace jazykových dat jsme usilovali o možnost podchytit všechny varianty tvarů slovesných lemmat. Současně však bylo nutné stanovit jednotná a jasná kritéria pro rozhodování anotátorů při manuální přípravě a úpravě synonymických řad. Výsledná podoba zápisu je tedy ovlivněna jak charakterem jazykového materiálu, tak potřebou jeho jasné formální a grafické reprezentace. Pro stanovení pravidel záznamu dat byly v maximální míře využívány mluvnice, jazykové příručky a slovníky ve snaze omezit vliv individuálního rozhodování na základě jazykové intuice anotátorů. Způsob, jakým jsou organizována jazyková data v synsetech VerbaLexu, podstatně ovlivňuje jejich grafickou podobu a mimo jiné také jejich délku. Synonymické řady jsou ve valenční databázi sestaveny ze synonymních lemmat doplněných číslem slovesného významu (v souladu s notací ve WN). Synonyma byla vybírána na základě Slovníku českých synonym [8], údajů ve Slovníku spisovného jazyka českého (SSJČ) [9], Slovníku spisovné češtiny (SSČ) [10] a jazykových dat v CzWN. Pojem synonymity ve VerbaLexu (a WN) není vymezen jen striktním požadavkem možnosti záměny slov v jednom kontextu, slovesa jsou charakterizována především stejnými sémantickými rysy, což umožnuje zahrnout je do jednoho synsetu se společnou definicí významu. Např.:

lézt$_1^{impf}$ **plazit se**$_1^{impf}$ **plížit se**$_1^{impf}$

Def: pomalými pohyby těla se pohybovat plazivě nebo šplhavě s celým tělem přiblíženým k podkladu, po němž se pohyb děje

3.1 Vid

Slovesné lemma a číslo významu nejsou jedinou charakteristikou českých sloves, která je ve VerbaLexu uvedena. Přímo v synsetu jsou zachyceny také informace o slovesném vidu. Graficky v textové podobě valenční databáze pomocí závorek, kdy je na prvním místě

uvedeno dokonavé sloveso a za ním jeho nedokonavý protějšek, dvojice má v těchto případech společné číslo významu. Např.:

ZAMLČET(ZAMLČOVAT):1

Ve formátu webového prohlížeče je potom u sloves v horním indexu uvedena značka pro vid (*pf. - perfektivní, impf. - imperfektivní, biasp. - obouvidá*) a číslo významu je zaznamenáno dolním indexem.

zamlčet $^{pf}_{1}$
zamlčovat $^{impf}_{1}$

V případě, že má dokonavá nebo nedokonavá podoba samostatný význam, je perfektivum nebo imperfektivum uvedeno zvlášť se samostatným číslem významu. Např.:

podniknout$^{pf}_{1}$
podnikat$^{impf}_{1}$
Def: uskutečnit nějaký plán, akci

podnikat$^{impf}_{2}$
Def: provozovat hospodářskou činnost

Dokonavá slovesa vytvořená prefixací od sloves nedokonavých zachycujeme ve VerbaLexu jako vidové dvojice jen v případech prefixace pomocí čistě vidové předpony, které jsou explicitně uvedeny ve slovnících SSČ a SSJČ (např. **učesat** – dok. *k česat$1$2*). Např.:

ucítit$^{pf}_{1}$
cítit$^{impf}_{1}$

Ostatní prefigovaná slovesa, u kterých nejde o prefixaci pomocí čistě vidové předpony, zapisujeme ve dvojici se sekundárním imperfektivem, nebo samostatně, v závislosti na jejich významu:

vytvořit$^{pf}_{2}$ **tvořit**$^{impf}_{1}$
vytvářet$^{impf}_{2}$

Tento způsob zachycení slovesného vidu ve VerbaLexu je ovšem jedním z faktorů, které ovlivňují délku synsetu.

3.2 Variantní lemmata

Způsob zápisu synsetu umožňuje zachytit i další informaci o slovesných tvarech, a to záznam variantních lemmat. Jejich způsob grafického zápisu má také značný vliv na délku synonymické řady. Řada českých sloves má dva (někdy i více) infinitivních tvarů,

u kterých nejde o změnu významu ani vidu, tvary se od sebe liší pouze malou hláskovou změnou. Ve VerbaLexu jsou tyto případy graficky zachyceny pomocí lomítek. Např.:

myslit/myslet$_1^{impf}$

Jako variantní lemmata jsou ve valenční databázi označeny – hláskové varianty (*škrábat/škrabat, spráskat/zpráskat, sepisovat/spisovat*), morfologické varianty (*moct/moci, tlouct/tlouci*) a pravopisné dublety (*representovat/reprezentovat*). Stejným formálním způsobem jsou zapsány také další případy, kdy však hlásková změna souvisí s mírně odlišným sémantickým příznakem. Jde o tvary dokonavých sloves s významem jednorázový/opakovaný děj (*zapísknout/zapískat*), tvary nedokonavých sloves se vztahem imperfektivum/sekundární imperfektivum utvořené od prefigovaného perfektiva (*umýt - mýt/umývat*), slovesa vyjadřující děj determinovaný/nedeterminovaný (*běžet/běhat*) a několik reflexivních sloves, která jsou kromě přidaného morfému *-se, -si* zcela homonymní (*koukat/koukat se*).

Často se v jednom synsetu setkávají různé kombinace záznamu slovesného vidu a variantních lemmat. Např.:

dolézt$_2^{pf}$
dolézat/dolízat$_2^{impf}$

pomoct/pomoci$_4^{pf}$
pomáhat$_4^{impf}$

kouknout/kouknout se$_1^{pf}$
koukat/koukat se$_1^{impf}$

obléct/obléci/obléknout$_1^{pf}$
oblékat$_1^{impf}$

Tato seskupení lemmat tvoří v podstatě vždy jednu významovou jednotku (mají stejné číslo významu), uvnitř skupiny se lemmata liší pouze drobnými hláskovými změnami, videm, případně malou mírou sémantického příznaku.

Stejným způsobem grafického záznamu jsou tedy zachycena jednak skutečná variantní lemmata a jednak případy dvojic (skupin) sloves, u kterých nejde pouze o hláskové varianty infinitivu, ale netvoří ani vidové opozice nebo nejsou slovesy s výrazně odlišným významem. Na první pohled jde o poněkud nekonzistentní způsob řešení, je však nutné si uvědomit, že formální (grafické) zachycení jazykového materiálu pro potřeby počítačového zpracování vyžaduje pokud možno jasná a jednoduchá pravidla bez velkého množství výjimek. Při tomto způsobu zpracování jazyka není efektivní explicitně vyznačovat jednotlivosti a zvláštnosti, je však nutné umět formálně zapsaná data správně interpretovat.

3.3 Možná řešení

Fakta uvedená v předchozím textu jsou jedním z důvodů, proč značně narůstá počet lemmat v jednom synsetu a stává se z něj pro uživatele nepřehledná řada sloves. Tento problém by mohl být řešen novým způsobem zápisu, který by ovšem nezpůsobil ztrátu důležitých informací. Variantní infinitivy a vidové protějšky by mohly být zaznamenány na jiném místě v textu se zachovanou vazbou ke slovesnému lemmatu, které bude reprezentantem významové skupiny. Při jejich vynechání ve formálním zápisu se může počet lemmat v synsetu snížit až o polovinu (a více), zápis se tak stává daleko čitelnějším. Jako příklad uvádíme synset, který s vidovými opozicemi a variantními infinitivy obsahuje 19 lemmat (pro svoji délku je zde rozdělen na 2 řádky):

naduřet$_1^{pf}$	**nalít se**$_1^{pf}$	**napuchnout**$_1^{pf}$	**natéct/natéci**$_1^{pf}$
	nalévat se/nalívat se$_1^{impf}$	**napuchat**$_1^{impf}$	**natékat**$_1^{impf}$

opuchnout$_1^{pf}$	**otéct/otéci**$_1^{pf}$	**podlít se**$_1^{pf}$	**zduřet**$_1^{pf}$	**naběhnout**$_2^{pf}$
opuchat$_1^{impf}$	**otékat**$_1^{impf}$	**podlévat se**$_1^{impf}$		**nabíhat**$_2^{impf}$

Def: zvětšit svůj objem, stát se objemnějším, než je normální; většinou o částech těla

Synset by bylo možné zkrátit uváděním pouze jednoho zástupce významové jednotky (infinitiv zapsaný na prvním místě), čímž by se např. výše uvedená synonymická řada zkrátila na 9 lemmat. Informace o vidových protějšcích a dalších variantách infinitivů by bylo možné uvádět na jiném místě v záznamu, případně je skrýt a ponechat na uživateli možnost vyvolat jejich zobrazení. Ve zkráceném synsetu by byly zastoupeny převážně dokonavé podoby slovesa, z variantních lemmat ta, která stojí na prvním místě (jejich pořadí je řízeno přesně stanovenými pravidly), a nedokonavá slovesa, která v daném významu nevstupují do vidové opozice. Způsob zápisu by mohl vypadat např. takto:

naduřet$_1^{pf}$ **nalít se**$_1^{pf}$ **napuchnout**$_1^{pf}$ **natéct**$_1^{pf}$ **opuchnout**$_1^{pf}$ **otéct**$_1^{pf}$ **podlít se**$_1^{pf}$ **zduřet**$_1^{pf}$ **naběhnout**$_2^{pf}$

pf.: /natéci$_1^{pf}$ /otéci$_1^{pf}$

impf.: nalévat se/nalívat se$_1^{impf}$ napuchat$_1^{impf}$ natékat$_1^{impf}$ opuchat$_1^{impf}$ podlévat se$_1^{impf}$ nabíhat$_2^{impf}$

4 Prefixace

Dalším faktorem, který má vliv na délku synonymické řady, je záznam prefigovaných sloves, jejichž významy lze zahrnout pod jednu společnou definici. Díky bohatosti prefixů

sloves v češtině počet lemmat v synsetu často značně narůstá, bez jejich zachycení by však byla databáze značně ochuzena. Např. v synsetu s definicí: *dosáhnout dohody jednáním, uzavřít smlouvu* najdeme 5 lemmat, která vznikla prefixací neprefigovaného imperfektiva *jednat* a 5 jejich sekundárních imperfektiv. Celý synset obsahuje celkem 20 lemmat, kromě prefigovaných podob slovesa *jednat* také lemmata *dohodnout, dohodovat, domluvit, domlouvat, smluvit, smlouvat, umluvit, umlouvat*.

sjednat$^{pf}_{1}$	**ujednat**$^{pf}_{1}$	**dojednat**$^{pf}_{1}$	**projednat**$^{pf}_{2}$	**vyjednat** $^{pf}_{1}$
sjednávat$^{impf}_{1}$	**ujednávat**$^{impf}_{1}$	**dojednávat**$^{impf}_{1}$	**projednávat**$^{impf}_{2}$	**vyjednávat**$^{impf}_{1}$

Omezit počet prefigovaných lemmat v zápisu synsetu a přitom neztratit žádnou z uvedených informací je poněkud obtížné. Řešením by však mohlo být například uvedení pouze základního neprefigovaného imperfektiva a sekundárního imperfektiva s poznámkou o možnosti prefixace vymezenou množinou prefixů. Zápis by se měl řídit přesnými pravidly, zachovat všechny uvedené charakteristiky a umožnit rekonstruovat původní (dlouhou) podobu synsetu. Např.:

pref1-**jednat**$^{impf}_{1}$ *pref2*-**jednat**$^{impf}_{2}$
pref1-**jednávat**$^{impf}_{1}$ *pref2*-**jednávat**$^{impf}_{2}$

*pref1 {**s-, u-, do-, vy-**}*
*pref2 {**pro-**}*

Je ovšem otázkou, zda je tento způsob zápisu pro uživatele přijatelnější než příliš dlouhý synset.

5 Stylistický příznak

Ve VerbaLexu je kromě sloves stylově neutrálních (*bít, mýt, sedět*, atd.) zachycena také řada sloves s různým stylistickým příznakem, především expresivním (*marodit, baštit, šmejdit, atd.*). Kritériem pro rozlišení sylistických příznaků u sloves jsou poznámky a značky v SSČ a SSJČ, které uvádějí, zda jde o slovo hovorové, knižní, básnické, nářeční, slangové, vulgární, užívané zřídka atd. U většiny stylistických příznaků (knižní, zastaralé, básnické, užívané zřídka, vulgární) je tendencí vůbec tato slovesa v databázi neuvádět. Výběr stylově zabarvených sloves do valenční databáze se omezuje převážně na slovníková hesla s poznámkou *expr.* a *hovor.* (expresivní, hovorová), navíc je jejich použití omezeno frekvencí výskytů v korpusech. Ve valenční databázi se slovesa neutrální i expresivní často nacházejí v jedné synonymické řadě, jejich synonymní významy lze totiž zahrnout pod jednu společnou definici. Tento způsob záznamu ovšem opět způsobuje výskyt příliš dlouhých synsetů. Např. (pro svou délku je synset rozdělen na 2 řádky):

spráskat/zpráskat$^{pf}_{1}$	**zbít**$^{pf}_{2}$ **bít**$^{impf}_{2}$	**zmlátit**$^{pf}_{1}$ **mlátit**$^{impf}_{1}$	**zmydlit**$^{pf}_{1}$ **mydlit**$^{impf}_{1}$	**ztřískat**$^{pf}_{1}$ **třískat**$^{impf}_{1}$	
vypráskat$^{pf}_{2}$	**napráskat**$^{pf}_{1}$	**domlátit**$^{pf}_{2}$	**namlátit**$^{pf}_{1}$	**zvochlovat**$^{pf}_{1}$ **vochlovat**$^{impf}_{1}$	**ztlouct/ztlouci**$^{pf}_{1}$ **tlouct/tlouci**$^{impf}_{1}$

Def: tělesně ztrestat

Synset v této podobě obsahuje celkem 20 lemmat. Řešením problému nadměrné délky synsetu způsobené výskytem expresivních sloves by mohlo být (mělo být) rozdělení synonymické řady na dvě části – slovesa neutrální a slovesa se stylistickým příznakem a explicitním uvedením poznámky o expresivitě. Uvedený synset by tedy mohl být rozdělen do dvou:

- *zbít(bít), ztlouct/ztlouci(tlouct/tlouci)*
- expr.: *spráskat/zpráskat, zmlátit(mlátit), zmydlit(mydlit), ztřískat(třískat), vypráskat, napráskat, domlátit, namlátit, zvochlovat(vochlovat)*

Expresivní synset zůstává stále příliš dlouhý, čeština je v těchto významech velmi bohatá na synonyma (bylo by možné doplnit ještě některá další). Jeho zkrácení bychom mohli dosáhnout použitím výše navrhovaného zápisu vidu, variantních lemmat a prefixace.

Rozdělování synsetů na stylově neutrální a expresivní může však přinést potíže v jiné oblasti valenční databáze. V úvodu bylo již zmíněno, že k českým synsetům ve VerbaLexu byla doplňována čísla významů (wnid) anglických ekvivalentů v PWN. Výsledkem práce bylo 85 % úspěšně nalinkovaných synsetů a 15 % synonymických řad, které nebylo možné spojit s anglickým ekvivalentem. Šlo o případy, kdy dané sloveso nelze vůbec do angličtiny přeložit, případně lze jeho význam vyjádřit pouze opisem nebo víceslovnou frází, většinou tedy nebylo možné v PWN najít vhodný ekvivalent. Tento problém se týkal konkrétně dokonavých sloves označujících dokončení činnosti (*dočesat – to finish combing, dokrmit – to finish feeding*), reflexivních sloves (*maskovat se – to disguise oneself*) a právě sloves se stylistickým příznakem expresivity. Pokud rozdělíme synsety ve VerbaLexu na stylově neutrální a expresivní, dojde patrně k nárůstu počtu nenalinkovaných synsetů.

V některých významech je však i angličtina bohatá na synonyma a expresivní výrazy jsou zachyceny také v PWN. V těchto případech můžeme synsety i po rozdělení provázat s jejich anglickými ekvivalenty. Například následující synset v původní (dlouhé) podobě obsahoval 18 lemmat. Po rozdělení na slovesa neutrální a expresivní získáme dva synsety (10 a 8 synonym) a také jejich anglické protějšky z PWN:

1. umřít$^{pf}_{1}$ **zemřít**$^{pf}_{1}$ **skonat**$^{pf}_{1}$ **zesnout**$^{pf}_{1}$ **zahynout**$^{pf}_{1}$ **zhynout**$^{pf}_{1}$ **dodýchat**$^{pf}_{1}$ **dokonat**$^{pf}_{2}$ **odejít**$^{pf}_{3}$

Def: přestat žít, pozbýt života (převážně o lidech)

wnid: ENG20-00347202-v

PWN: die:1, decease:1, perish:1, go:17, exit:3, pass away:1, expire:2, pass:25
Def: pass from physical life and lose all bodily attributes and functions necessary to sustain life

2. dodělat$^{pf}_{1}$ **chcípnout**$^{pf}_{1}$ **pochcípat**$^{impf}_{1}$ **zdechnout**$^{pf}_{1}$ **pojít**$^{pf}_{1}$ **uhynout**$^{pf}_{1}$ **zcepenět**$^{pf}_{1}$

Def: přestat žít, pozbýt života (převážně o zvířatech, expr. i o lidech)

wnid: ENG20-00349254-v

PWN: kick the bucket:1, cash in one's chips:1, buy the farm:1, conk:3, give-up the ghost:1, drop dead:1, pop off:2, choke:12, croak:1, snuff it:1
Def.: die

6 Závěr

Na základě rozboru vzorku 50 nejdelších synonymických řad ve VerbaLexu byly stanoveny faktory, které ovlivňují jejich délku. Jde především o způsob záznamu různých tvarů slovesných lemmat – zachycení vidových opozic, variantních lemmat a prefigovaných sloves a uvádění sloves s rozličnými stylovými příznaky v jednom synsetu. Řešením první uvedené problematiky by mohlo být přenesení informací o vidu, variantních lemmatech a prefixaci do jiných částí zápisu se zachováním všech relevantních údajů a možností zpětné rekonstrukce původního synsetu. Dlouhé synsety, ve kterých jsou uvedena slovesa s různým stylistickým příznakem, by měly být rozděleny do dvou (či více) synonymických řad. V mnoha případech však ani těmito navrhovanými změnami nedosáhneme výrazného zkrácení synsetu, neboť čeština je jazyk bohatý na synonyma. Dlouhé synsety, které by ve VerbaLexu zůstaly i po zavedení všech navrhovaných změn, by nebyly jakousi anomálií, ale dokladem pestrosti českého jazyka.

Pozitivním výsledkem uvedeného rozboru je, že délka synsetů není způsobena chybami anotátorů při výběru synonymních lemmat. Negativním zjištěním potom fakt, že zobrazení dlouhých synonymních řad z VerbaLexu způsobuje značné potíže i při přípravě textu na toto téma.

6.1 Úplný závěr

Na vyvážení některých morbidních příkladů v článku uvádíme ještě jeden veselý:

blahopřát$^{impf}_{1}$ **gratulovat**$^{impf}_{1}$ **poblahopřát**$^{pf}_{1}$ **pogratulovat**$^{pf}_{1}$ **popřát**$^{pf}_{1}$ **přát**$^{impf}_{2}$

Def: projevovat společenskou formou touhu, aby se někomu dostalo něčeho dobrého

-frame: AG<person:1>$^{obl}_{kdo1}$ *VERB* **PAT**<person:1>$^{obl}_{komu3}$ **ABS**<jubilee:1>$^{opt}_{k\ čemu3}$

-example: *Dana blahopřeje Karlovi k významnému jubileu*

References

1. Hlaváčková, D. *Databáze slovesných valenčních rámců VerbaLex*. Disertační práce. Brno: FF MU, 2007.
2. Fellbaum, Ch. (ed.) *WordNet. An Electronic Lexical Database.* Cambridge: MIT Press, 1998.
3. Vossen, P. - Bloksma, L. et al. *The EuroWordNet Base Concepts and Top Ontology.* Technical Report Deliverable D017, D034, D036, WP5 EuroWordNet, LE2-4003. University of Amsterdam, 1998.
4. Němčík, V. - Pala, K. - Hlaváčková, D. Semi-automatic Linking of New Czech Synsets Using Princeton WordNet. In *Intelligent Information Systems XVI, Proceedings of the International IIS'08 Conference*. Warszawa: Academic Publishing House EXIT, 2008, s. 369–374.
5. Pala, K. - Ševeček, P. Valence českých sloves. In *Sborník prací Filosofické fakulty Masarykovy university*, A45. Brno, 1997, s. 41–54.
6. Žabokrtský, Z. *Valency Lexicon of Czech Verbs.* Disertační práce. Praha: MFF UK, 2005.
7. Svozilová, N. - Prouzová, H. - Jirsová, A. *Slovesa pro praxi. Valenční slovník nejčastějších českých sloves.* Praha: Academia, 1997.
8. Pala, K. - Všianský, J. *Slovník českých synonym.* Praha: Lidové noviny, 1996.
9. *Slovník spisovného jazyka českého.* Praha: Academia, 1989.
10. *Slovník spisovné češtiny pro školu a veřejnost.* Praha: Academia, 2000.

Czech Word Sketch Relations with Full Syntax Parser

Aleš Horák[1], Pavel Rychlý[1], and Adam Kilgarriff[2]

[1] Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{hales,pary}@fi.muni.cz
[2] Lexical Computing Ltd.
Brighton, UK
adam@lexmasterclass.com

Abstract. This paper describes the exploitation of dependency relations obtained from syntactic parsing of Czech for building new Czech Word Sketch tables. Standard Word Sketch construction process usually uses so called Sketch grammars – a simplified process of identifying dependency relations based on regular expressions. This may, of course, lead to errors, which should however not influence (so much) the overall numbers computed on a very big corpus.
The paper presents an experiment of using relations resulting from full syntactic parsing – will they perform better than the standard Sketch grammar or not?

1 Introduction

Dictionary making involves finding the distinctive patterns of usage of words in texts. State-of-the-art corpus query systems can help the lexicographer with this task. They offer great flexibility to search for phrases, collocates, grammatical patterns, to sort concordances to a wide range of criteria and to identify subcorpora for searching only in texts of a particular genre or type. The Sketch Engine [1] is such a corpus query system.

In this paper we discuss the work involved in setting up the Sketch Engine for the new Czech corpus named CZES using two different systems for the dependency relations discovery – the standard Sketch Grammar approach based on regular expressions, and dependency relations obtained by means of full syntax parsing of Czech. We give a detailed description of the various features of the Sketch Engine in relation to the Czech language. The structure of this paper is as follows. First we give some background information on the new CZES corpus and its setup within the Sketch Engine. Then we discuss some general features of the Sketch Engine in Section 3 followed by a detailed description of the work involved in setting up the Sketch Engine for the two sources of dependency relations. We conclude with a short evaluation in the last section.

2 The New Corpus CZES

The Institute of Czech National Corpus has prepared several large Czech corpora. The data of these corpora are provided for research only through web access, it is not possible to add new annotation and process texts by specific batch tools. This is the main reason

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 101–112, 2009.

why a new Czech Corpus CZES was built in the Masaryk University NLP Centre. CZES was built purely from electronic sources by mostly automated scripts and systems. The corpus name is an acronym of **CZ**ech **E**lectronic **S**ources.

Texts in the CZES corpus come from three different sources:

1. automated harvesting of newspapers (either electronic version of paper ones or electronic only), with annotation of publishing dates, authors and domain; these information is usually hard to find automatically from other sources;
2. customized processing of electronic versions of Czech books available online; and
3. general crawling of the Web.

The whole corpus should contain Czech texts only. There are small parts (paragraphs) in Slovak or English because they are parts of the Czech texts. Some Czech newspapers regularly publish Slovak articles, but we have used an automatic method to identify such articles and remove them from the corpus.

There was no restriction on the publication date of texts. There are both latest articles from current newspapers and 80 year old books present in the corpus.

We are adding more texts to the corpus, the current full corpus size is about 600 million word forms. To speed-up processing and research of different annotations, the work described in this paper uses only a sample of about 85 million tokens from the whole CZES corpus.

In order to support lexicographic searches such as searches by lemma, by part of speech and the extraction of words belonging to a specific word class, the corpus has been annotated with lemma and morphological tags. We have used the Czech tagger DESAMB developed at the NLP Centre [2]. The tagger is based on morphological analyzer AJKA [3] and uses so called "Brno" tag-set for morphological tags.

2.1 Preparing the Corpus

The Sketch Engine input format, often called "vertical" or "word-per-line", is as defined at the University of Stuttgart in the 1990s and widely used in the corpus linguistics community. Each token (e.g. word or punctuation mark) is on a separate line and where there are associated fields of information, typically the lemma and a POS-tag, they are included in tab-separated fields. Structural information, such as document beginnings and ends, sentence and paragraph markup, and meta-information such as the author, title and date of the document, and its text type, are presented in XML-like form on separate lines – see an example from CZES in Figure 1.

A special tag, <g>, was added before punctuation marks: it has the effect of suppressing the space character which is otherwise output between one token and the next. (G is for 'glue' as the <g> tag 'glues' the punctuation onto the preceding word.)

The <s> tag is used to annotate sentence boundaries and it was added by the tagger.

3 The Sketch Engine

The Sketch Engine is a sophisticated corpus query system. In addition to the standard corpus query functions such as concordances, sorting, filtering, it provides *word sketches*,

```
<doc id="autodesk/1995/05/7" t_main="sci1" medium="cdrom"
    t_orig="Software" lang="cs" title="Autodesk WorkCenter"
    auth_n="Petr Kumprecht" source="CD Modrých stránek"
    d_publ="1995-10" t_sub="inf">
<head>
<s>
Autodesk        Autodesk        kA
WorkCenter      WorkCentra      k1gFnPc2
</s>
</head>
<p>
<s>
Document        Document        k1gInSc1
Management      management      k1gInSc1
a               a               k8xC
Workflow        Workflow        k1gInSc1
Management      management      k1gInSc1
System          System          k1gInSc1
Začátkem        začátkem        k7c2
letošního       letošní         k2eAgInSc2d1
roku            rok             k1gInSc2
uvedla          uvést           k5eAaPmAgFnS
společnost      společnost      k1gFnSc1
Autodesk        Autodesk        kA
na              na              k7c4
trh             trh             k1gInSc4
zcela           zcela           k6eAd1
nový            nový            k2eAgInSc4d1
systém          systém          k1gInSc4
pro             pro             k7c4
správu          správa          k1gFnSc4
dokumentace     dokumentace     k1gFnSc2
<g/>
,               ,               kIx,
Autodesk        Autodesk        kA
WorkCenter      WorkCenter      k1gInSc1
<g/>
.       .       kIx.
</s>
```

Fig. 1. An example of the corpus vertical format with document meta-data.

one page summaries of a word's grammatical and collocational behaviour by integrating grammatical analysis.[3]

Based on the grammatical analysis, the Sketch Engine also produces a distributional *thesaurus* for the language, in which words occurring in similar settings, sharing the same collocates, are put together, and *sketch differences*, which specify similarities and

[3] The Sketch Engine prefers input which has already been lemmatized and POS tagged. If no lemmatized input is available it is possible to apply the Sketch Engine to word forms which, while not optimal, will still be a useful lexicographic tool.

differences between near-synonyms. The system is implemented in C++ and Python and designed for use over the web.

Once the corpus is loaded into the Sketch Engine, the concordance functions are available. The lexicographer can immediately use the search boxes provided, searching, for example, for a lemma specifying its part of speech. This search is case-sensitive as generally lemmas starting with uppercase need to be distinguished from those starting with lower case.

We must note here that the quality of the output of the system depends heavily on the input, i.e. the quality of tagging and lemmatization, which as mentioned in Section 2 is not always entirely satisfactory. According to the sources of some parts of the CZES corpus, the texts can contain misspelled words and neologism, which are tagged by the *guesser* module of the tagger.

On the results page the concordances are shown using KWIC view. With VIEW options it is possible to change the concordance view to a number of alternative views. One is to view additional attributes such as POS tags or lemma alongside each word. This can be useful for finding out why an unexpected corpus line has matched a query, as the cause could be an incorrect POS-tag or lemmatization.

It is central to the process of corpus lexicography that lexicographers often want to insert example sentences from the corpus into the dictionary. Some corpus sentences make good dictionary examples, but others do not. Perhaps they are too long, or too short, or are not well-formed sentences, or contain obscure words or spelling mistakes or abbreviations or strange characters. To find a good dictionary example is a high-level lexicographic skill. But to rule out lots of bad sentences is easy, and the computer can help by doing this groundwork. A new function, GDEX (Good Dictionary Example eXtractor [4]) was added to the Sketch Engine in 2008. This takes the first 200 (by default) sentences matching a query, scores them according to how good a dictionary example the computer thinks they will make, and returns them in order, best first. The scoring is done with a series of simple rules addressing the considerations listed above: how long is the sentence; does it contain words outside the core language vocabulary; does it begin with a capital letter and end with a full stop, exclamation mark or question mark; does it contain an excessive number of characters other than lower-case `a`-to-`z`? The goal is that the average number of corpus lines that a lexicographer has to read, before finding one suitable to use or adapt for the dictionary entry, is substantially reduced, so they rarely have to look beyond the first ten whereas without GDEX, they may often have had to look through thirty or forty.

4 Word Sketches and the CZES corpus

Word sketches are the distinctive feature of the Sketch Engine. Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. Word sketches improve on standard collocation lists by using a grammar and parser to find collocates in specific grammatical relations, and then producing one list of subjects, one of objects, etc. rather than a single grammatically blind list.

In order to identify a word's grammatical and collocational behaviour, the Sketch Engine needs to know how to find words connected by a grammatical relation. For this to work, the input corpus needs to be parsed or at least POS tagged.

If the corpus is parsed, the information about grammatical relations between words is already embedded in the corpus and the Sketch Engine can use this information directly. A modification of this method was used to handle output of a syntactic parser. If the corpus is POS-tagged but not parsed, grammatical relations can be defined by the developer within the Sketch Engine using a Sketch Grammar.

An example of the word sketch is in Figure 2. The user can set various preferences for the display of the word sketches. Collocates can be ranked according to the frequency of the collocation, or according to its salience score (see [5] for the formula used to compute the salience). The user can set a frequency threshold so low-frequency collocations are not shown. On the results screen the user can go to the related concordance by clicking on the number next to the lemma.

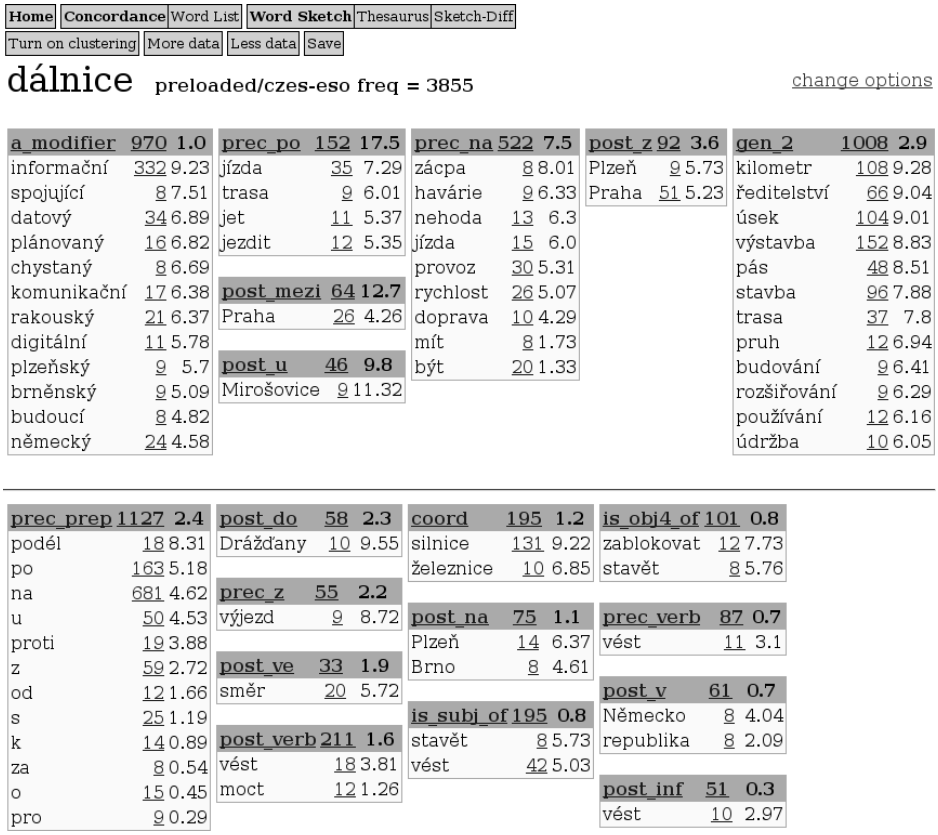


Fig. 2. Word sketch for the word “dálnice” (highway).

4.1 Czech Sketch Grammar

In this model, grammatical relations are defined as regular expressions over POS-tags. For example, a grammatical relation specifying the relation between a noun and a pre-modifying adjective looks like this.

```
=modifier
2:"A.*" 1:"N.*"
```

The first line, following the `=`, gives the name of this grammatical relation. The `1:` and `2:` mark the words to be extracted as first argument (the keyword) and second argument (the collocate).

The result is a regular expression grammar which we call a Sketch Grammar. It allows the system to automatically identify possible relations of words to the keyword. These grammars are of course less than perfect, but given the errors in the POS-tagging, this is inevitable however good the grammar. The problem of noise is mitigated by the statistical filtering which is central to the preparation of word sketches.

The first version of the Czech Sketch Grammar was created in the early stage of the Sketch Engine development [1]. It was prepared for the "Prague" tag-set used in the Czech National Corpus. We have adopted the grammar to match the Brno annotation.

When the corpus is parsed with the grammar, the output is a set of tuples, one for each case where each pattern matched. The tuples comprise (for the two-argument case), the grammatical relation, the headword, and the collocate, as in the third column in the table. This work is all done on lemmas, not word forms, so headword and collocate are lemmas.

As can be seen from Table 1, grammatical relations in the Czech Sketch Grammar are of four types, i.e. regular (one way dependency relation), symmetric (between two items with equal status), dual (between two items with dependent relations), trinary (between three dependent items). The sketch engine also supports unary relations but these are not used in the Czech Sketch Grammar. Unary relations are used to extract certain complementation patterns. For instance, a lexicographer would like to know that a verb is frequently followed by a relative clause starting with *že* (that) or that a noun is preceded by an article or not.

Dual relations are the most common. They work similarly to symmetric relations but inversing a dual relation results in a different grammatical relation. A typical dual is the pair, "verb and its object" and "noun and the verb it is object of".

Figure 2 shows the resulting word sketch for word *dálnice* (highway).[4] We can see that the discovered collocations can say a lot about the document sources – here, the most frequent adjective modifier of *dálnice* is *informační* (information highway). An interesting evidence of the state of Czech highways is the list corresponding to the preposition *na* (at), which contains *zácpa* (traffic jam), *havárie* (crash) and *nehoda* (accident) as its top entries.

The Czech Sketch Grammar generates about 46 million triples (dependences) from the 85 million token corpus.

[4] The word sketch is about two times bigger with the default options.

Table 1. The Czech Sketch Grammar grammatical relations

Relation	**Example**	**Triple(s)**
symmetric relations		
COORD	silnice a dálnice *roads and highways*	⟨coord,silnice,dálnice⟩ ⟨coord,dálnice,silnice⟩
regular relations		
PREC_VERB	v blízkosti vede dálnice *a highway is nearby*	⟨prec_verb,dálnice,vést⟩
POST_VERB	dálnice většinou vede obcemi *highway usually goes through cities*	⟨post_verb,dálnice,vést⟩
POST_INF	kudy měla nová dálnice vést *where should the new highway go*	⟨post_inf,dálnice,vést⟩
PREC_PREP	telefony podél dálnic *phones along highways*	⟨prec_prep,dálnice,podél⟩
POST_PREP	dálnice před Prahou *the highways in front of Prague*	⟨post_prep,dálnice,před⟩
dual relations		
IS_SUBJ_OF/HAS_SUBJ	kudy dálnice povede *where will the highway go*	⟨is_subj_of,dálnice,vést⟩ ⟨has_subj,vést,dálnice⟩
IS_OBJ2_OF/HAS_OBJ2	co se týká dálnice *what applies to highway*	⟨is_obj2_of,dálnice,týkat se⟩ ⟨has_obj2,týkat se,dálnice⟩
IS_OBJ3_OF/HAS_OBJ3	situace přinese dálnici ... *the situation brings new possibilities to the highway*	⟨is_obj3_of,dálnice,přinést⟩ ⟨has_obj3,přinést,dálnice⟩
IS_OBJ4_OF/HAS_OBJ4	kamion zablokoval dálnici *truck blocked the highway*	⟨is_obj4_of,dálnice,zablokovat⟩ ⟨has_obj4,zablokovat,dálnice⟩
IS_OBJ7_OF/HAS_OBJ7	vládá se zabývá dálnicemi *the government deals with highways*	⟨is_obj7_of,dálnice,zabývat se⟩ ⟨has_obj7,zabývat se,dálnice⟩
GEN_1/GEN_2	dálnice budoucnosti *highway of the future*	⟨gen_1,dálnice,budoucnost⟩ ⟨gen_2,budoucnost,dálnice⟩
PASSIVE/SUBJ_OF_PASSIVE	přeplněná dálnice *crowded highway*	⟨passive,přeplnit,dálnice⟩ ⟨subj_of_passive,dálnice,přeplnit⟩
CATEG1/CATEG2	dálnice je typ silnice *highway is a type of a road*	⟨categ1,dálnice,silnice⟩ ⟨categ2,silnice,dálnice⟩
AJINE1/AJINE2	dálnice a jiné projekty *highways and other projects*	⟨ajine1,dálnice,projekt⟩ ⟨ajine2,projekt,dálnice⟩
BYT_ADJ/SUBJ_BYT	dálnice byla namrzlá *the highway was frosty*	⟨byt_adj,namrzlý,dálnice⟩ ⟨subj_byt,dálnice,namrzlý⟩
A_MODIFIER/MODIFIES	informační dálnici *information highway*	⟨a_modifier,dálnice,informační⟩ ⟨modifies,informační,dálnice⟩
trinary relations		
POST_*	na dálnici v Německu, *at the highway in Germany*	⟨post_*,dálnice,Německo,v⟩ ⟨post_v,dálnice,Německo⟩
PREC_*	u výjezdu z dálnice *at the highway exit*	⟨prec_*,dálnice,výjezd,z⟩ ⟨prec_z,dálnice,výjezd⟩

4.2 Dependency Relations from Syntactic Parser

The Czech syntactic parser `synt` [6,7] is developed in the Natural Language Processing Centre at Masaryk University. The parsing system uses an efficient variant of the head driven chart parsing algorithm [8] together with the meta-grammar formalism for the language model specification. The advantage of the meta-grammar concept is that the grammar is transparent and easily maintainable by human linguistic experts. The meta-grammar includes about 200 rules covering both the context-free part as well as context relations. Contextual phenomena (such as case-number-gender agreement) are covered using the per-rule defined contextual actions. The meta-grammar serves as a basis for a machine-parsable grammar format used by the actual parsing algorithm – this grammar form contains almost 4,000 rules.

Currently, the `synt` system offers a coverage of more than 92 percent of (common) Czech sentences[5] while keeping the analysis time on the average of 0.07s/sentence.

Besides the standard results of the chart parsing algorithm, `synt` offers additional functions such as partial analysis (shallow parsing) [10], effective selection of *n*-best output trees [8], chart and trees linguistic simplification [11], or extraction of syntactic structures [12]. All these functions use the internal chart structure which allows to process potentially exponential number of standard derivation trees still in polynomial time.

Apart from the common generative constructs, the metagrammar includes feature tagging actions that specify certain local aspects of the denoted (non-)terminal. One of these actions is the specification of the head-dependent relations in the rule — the `depends()` construct:

```
/* černá kočka (black cat) */
np → left_modif np
    depends($2,$1)
/* třeba (perhaps) */
part → PART
    depends(root,$1)
```

In the first rule, `depends($2,$1)` says that (the head of) the group under the `left_modif` non-terminal depends on (the head of) the `np` group on the right hand side. In the second example, `depends(root,$1)` links the `PART` terminal to the root of the resulting dependency tree. The meta-grammar allows to assign *labels* to parts of derivation tree, which can be used to specify dependencies "crossing" the phrasal boundaries. The `synt` system thus allows to process even *non-projective phenomena*, which would otherwise be problematic within a purely phrasal approach.

The relational `depends` actions sequentially build a graph of dependency links between surface tokens. Each call of the action adds a new edge to the graph with the following information about the *dependent* group:

1. the non-terminal at the top of the group (`left_modif` or `np` in the example above),
2. the pre-terminal (word/token category) of the *head* of the group, i.e. the single token representing the group, and
3. the grammatical case of the head/group, if applicable.

[5] measured on 10,000 sentences from the DESAM corpus [9].

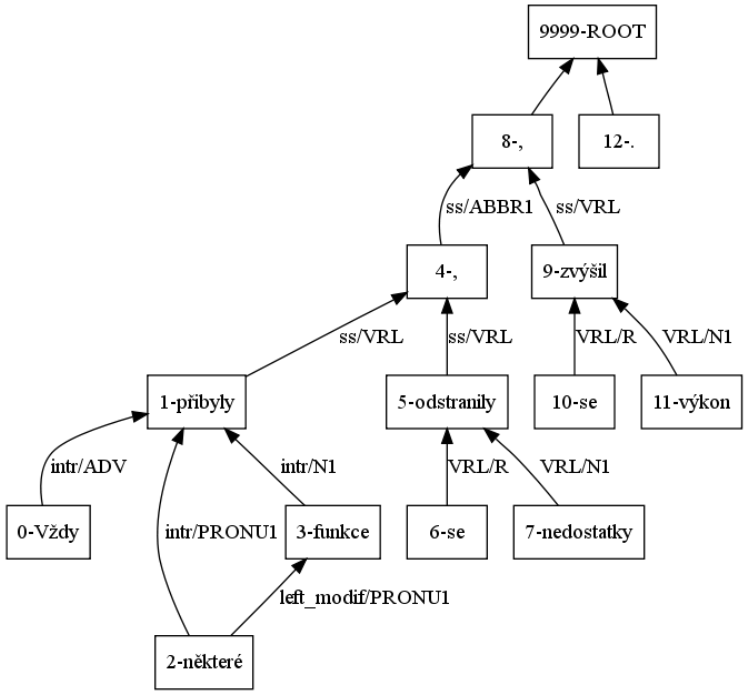


Fig. 3. An example of `synt` dependency graph output for the sentence "*Vždy přibyly některé funkce, odstranily se nedostatky, zvýšil se výkon.*" (Each time new functions were added, drawbacks were removed, the power increased).

An example list of such dependency relations for a corpus sentence "*Vždy přibyly některé funkce, odstranily se nedostatky, zvýšil se výkon.*" (Each time new functions were added, drawbacks were removed, the power increased) may look like this:

from	*label*	*to*	*from*	*label*	*to*
0	`intr/ADV`	1	5	`ss/VRL`	4
1	`ss/VRL`	4	6	`VRL/R`	5
2	`intr/PRONU1`	1	7	`VRL/N1`	5
2	`left_modif/PRONU1`	3	9	`ss/VRL`	8
3	`intr/N1`	1	10	`VRL/R`	9
4	`ss/ABBR1`	8	11	`VRL/N1`	9

The corresponding dependency graph of this sentence is depicted in Figure 3.

We can see that the information in these relations contains more details that come from the parsing process. However, not all details bring the same amount of linguistic adequacy – e.g. distinguishing `left_modif/ADJ1` and `left_modif/ADJ2` does not bring any new

information,[6] whereas `intr/N1` links to verbs where the dependent group is a subject and `intr/N4` lists objects in accusative.

Within the experiment of parsing the CZES corpus (about 4 million sentences), we have obtained more than 52 millions of dependency relations.

4.3 Thesaurus

Once the corpus has been parsed and the tuples extracted, we have a very rich database that can be used in a variety of ways.

We can ask "which words share most tuples", in the sense that, if the database includes both $\langle gramrel, w_1, w \rangle$ and $\langle gramrel, w_2, w \rangle$ (for example $\langle$prec_na, dálnice, provoz$\rangle$ and $\langle$prec_na, silnice, provoz$\rangle$), then we can say that w_1 and w_2 share a triple. A shared triple is a small piece of evidence that two words are similar. Now, if we go through the whole lexicon, asking, for each pair of words, how many triples do they share, we can build a 'distributional thesauruses', which, for each word, lists the words most similar to it (in an approach pioneered in [13,14]). The Sketch Engine computes such a thesaurus. A thesaurus entry for *dálnice* obtained from the standard Sketch Grammar starts with:[7]

- silnice (road)
- železnice (railway)
- trasa (path), trať (route), most (bridge)
- elektrárna (power station), komunikace (communication), vozovka (pavement), ropovod (pipeline)
- infrastruktura (infrastructure)

The same thesaurus entry computed with the dependency relations obtained from syntactic parsing looks like:

- silnice (road)
- ropovod (pipeline), tunel (tunnel), trasa (path)
- vozovka (pavement), infrastruktura (infrastructure), most (bridge), železnice (railway), trať (route), komunikace (communication)
- dráha (line)
- elektrárna (power station)

The main synonym *silnice* stays the same, but other similar words are grouped in different order. Evaluation of these two approaches, however, needs further studies from both grammarian and lexicographer's point of view.

5 Conclusion

We have loaded the CZES corpus into the Sketch Engine. The process was designed to support various lexicographic tasks at the Masaryk University NLP Centre.

The distinctive feature of the Sketch Engine are its word sketches. The standard way to set them up for Czech involved writing a Sketch Grammar to define the set of Czech

[6] It just says that the collocation *adjective+noun* was in nominative or genitive.

[7] The words are grouped according to the thesaurus score.

Grammatical relations. Each grammatical relation is defined using a regular-expression grammar over part-of-speech tags. The paper documents the grammatical relations for Czech.

Another way of defining word sketches, that was experimentally tested, lies in using dependency relations obtained from full syntax parsing of Czech. The resulting dependency relations provide further levels of details coming from the parsing process at the place of the relation label. What remains to be done is to prepare a linguistically justified translation of these labels to provide the most adequate information based on the parsing results.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project 407/07/0679.

References

1. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. Proceedings of Euralex (2004) 105–116. http://www.sketchengine.co.uk
2. Šmerk, P.: Towards czech morphological guesser. In: Proceedings of Recent Advances in Slavonic Natural Language Processing 2008, Brno, Czech Republic, Masaryk University (2008) 1–4
3. Sedláček, R.: Morphemic Analyser for Czech. PhD thesis, Masaryk University (2005)
4. Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P.: GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In: Proceedings of the XIIIth EURALEX International Congress. Barcelona: Universitat Pompeu Fabra. (2008) 425–432
5. Rychlý, P.: A Lexicographer-Friendly Association Score. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, Brno, Czech Republic, Masaryk University (2008) 6–9
6. Horák, A.: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. PhD thesis, Masaryk University (2002)
7. Kadlec, V., Horák, A.: New Meta-grammar Constructs in Czech Language Parser synt. In: Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2005)
8. Horák, A., Kadlec, V., Smrž, P.: Enhancing Best Analysis Selection and Parser Comparison. In: Lecture Notes in Artificial Intelligence, Proceedings of TSD 2002, Brno, Czech Republic, Springer Verlag (2002) 461–467
9. Pala, K., Rychlý, P., Smrž, P.: DESAM – Annotated Corpus for Czech. In: Proceedings of SOFSEM '97, Springer-Verlag (1997) 523–530
10. Ailomaa, M., Kadlec, V., Rajman, M., Chappelier, J.C.: Robust stochastic parsing: Comparing and combining two approaches for processing extra-grammatical sentences. In Werner, S., ed.: Proceedings of the 15th NODALIDA Conference, Joensuu 2005, Joensuu, Ling@JoY (2005) 1–7
11. Kovář, V., Horák, A.: Reducing the Number of Resulting Parsing Trees for the Czech Language Using the Beautified Chart Method. In: Proceedings of the 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland (2007) 433–437

12. Jakubíček, M., Horák, A., Kovář, V.: Mining Phrases from Syntactic Analysis. In: Proceedings of TSD 2009, Springer-Verlag (2009)
13. Grefenstette, G.: Explorations in automatic thesaurus discovery. Springer (1994)
14. Lin, D.: Automatic Retrieval and Clustering of Similar Words. In: Conference on Computational Linguistics (COLING-ACL). (1998) 768–774

Syntaktická struktura *Petr byl boxovat*: české specifikum, nebo evropské univerzále?*

Petr Karlík

Filozofická fakulta, Masarykova univerzita, Brno
pkarlik@phil.muni.cz

(věnováno Karlovi, který mně kdysi ukázal, že lingvisticky myslet se dá i s matematickou exaktností, a taky
městům V. a V., které mně letos umožnily prožít v nich esteticky krásný zážitek)

Abstract. The syntactic structure *Petr byl boxovat*(Petr be-3.sg.past box-Inf): a Czech specific or European universal?
The goal of the study is to give an answer to the question whether Czech constructions BE InfP of the type *Petr byl boxoval* and constructions BE InfP in eight other European languages (e.g. Italian *Gianni è a boxare*, German *Jan ist boxen*, Swedish *John är och boxar* etc.) represent a linguistically specific variant of one structure generated by principles of Universal Grammar. On the basis of different syntactic behavior of the constructions BE InfP it is concluded that in Czech it is a construction with a deep structure other than that in the rest of the languages, which explains why the Czech construction is semantically interpreted as a resultative whereas the construction in the other languages is interpreted as an absentive (in the sense of de Groot). A different syntax and semantics of a construction with an identical surface structure provide at the same time a support for the assertion that identical surface structure in various European languages (Indo-European and non-Indo-European) is not associated with what is known as languages in integration and integration is languages.

V této studii mám tento úkol: podat deskriptivní analýzu jednoho typu českých struktur obsahujících ESSE (tj. *být*), totiž takových, v nichž ESSE bere za svůj komplement Inf(initiv)P(hrase). Nechám-li teď stranou konstrukci s dvěma infinitivy typu *Koupiti není koupiti*, protože je to struktura syntakticky méně zajímavá, zbývají nám v moderní češtině čtyři typy konstrukcí s ESSE InfP: (2), jejíž syntax se patrně čeští native speakers musí naučit empiricky (včetně distribuce prézentních tvarů ESSE a InfF ukazované v příkladu), a (1), (3) a (4), jejichž syntax čeští native speakers umějí na základě nabytí parametrizace univerzálněgramatických principů, a tedy umějí tyto struktury z jednotek uložených v lexikonu (ať už je organizován jakkoli) tvořit, a taky sémanticky interpretovat:

(1) Petr bude fotografovat Ilonu z loďky
(2) Všem lidem jest/*je umříti/?umřít
(3) Petr byl lovit ryby ze břehu
(4) Petr byl vidět ze břehu

* Studie vznikla v rámci řešení projektu *Integrace v jazycích – jazyky v integraci*, podporovaného GA ČR grantem č. 405/07/0652.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 113–124, 2009.

Mým úkolem bude rozebrat českou konstrukci ESSE InfP v (3)[1], zjistit, zda konstrukce s identickou (povrchovou) syntaxí a se stejnou sémantickou interpretací existuje i jinde, a pokud existuje ve více evropských jazycích, pak odpovědět na otázku, zda (3) přesto představuje české specifikum, nebo je výsledkem nějakého integračního procesu řízeného univerzálněgramatickými principy a nastartovaného jazykovým kontaktem.

Vyjdu z truismu: sloveso ESSE, viděno napříč jazyky, ukazuje maximum ideosynkratických rysů, a to morfologických (např. vyznačuje se maximální supletivitou – a defektivnost paradigmatu je typická pro funkční kategorie) i – a to bude teď důležité – syntaktických: vyznačuje se minimálně restriktivní selekcí komplementu: NP/DP, AP/DegP, AdvP, PP, VP. Budu-li komplementy ESSE charakterizovat základními kategoriálními rysy N(omen) a V(erbum) jejich lexikálních hlav, můžeme na tomto základě dospět k nějaké kontrolovatelné syntaktické klasifikaci ESSE:

ESSE_1 selektující $_{kompl}$[+N, ±V], tj. NP/DP (*Já jsem profesor/-em fyziky*) a AP/DegP (*Já jsem starší než ty o 24 let*)

ESSE_2 selektující $_{kompl}$[–N, +V], tj. VP: l-ParticipiumP (*Já jsem přišel*), InfP (*Já jsem zase vidět, Já budu zase vidět*)

ESSE_3 selektující $_{kompl}$[bez kategoriálních rysů N a V], tj. AdvP (*Já jsem doma*)

ESSE selektující $_{kompl}$[PP], tj. strukturu, jejíž lexikální hlavou je (za předpokladu, že P je funkční kategorie) prototypicky N, tj. komplex rysů [+N, –V], musí mít stejné vlastnosti buď jako ESSE_1 selektující $_{kompl}$[+N, ±V] (*Já jsem učitel / na blondýnky*), nebo jako ESSE_3 selektující $_{kompl}$[bez kategoriálních rysů N a V], protože P jakožto funkční kategorie rysy N a V nemá (*Já jsem doma / na fakultě*).

Tyto pro lexikální sloveso netypické vlastnosti ESSE jsou samozřejmě notoricky známé už z klasických gramatik, v nichž se podle nich rozlišují taky tři ESSE jako „pomocná slovesa“: ESSE_1, známé pod označením *kopula* (ESSE_{cop}), ESSE_2, označované v kookurenci s l-participiem jako *auxiliár* (ESSE_{aux}), v kookurenci s infinitivem různě, a ESSE_3 (existenční a lokalizační, event. i posesivní, jako v ruštině), které se chápe buď jako varianta ESSE_{cop}, nebo jako *lexikální sloveso*.[2]

Podíváme-li se na ESSE_1 – ESSE_3, snadno zjišťujeme, že mezi kategoriálními požadavky ESSE na selekci komplementu a morfofonologickými a morfosyntaktickými vlastnostmi takto syntakticky rozdílných ESSE existuje zjevná korelace.

První vlastností, kterou budu při sledování zmíněné korelace reflektovat, je to, že ESSE je (analyzováno jako) sloveso, které má s finitní morfologií dvě supletivní subparadigmata, která se liší různými hodnotami rysu [aspekt]; srov. Migdalski (2006) a mnoho jiných (i z okruhu diachronie) před ním.

Nejprve budu pozorovat ESSE *budu*. Za standardní se dnes pokládá analýza, podle níž ve struktuře [bud-e-u > budu, bud-e-š, bud-e-Ø][3] vyjadřuje sada flexémů {-u, -š, -Ø, -me, -te, -ou} rys [–minulost], stejně jako např. u lexikálních sloves: hnět-e-u

[1] Čtenář zatím může hádat, proč jubilantovi věnuji jako dárek popis právě této konstrukce. Dozvíte se to, budete-li číst moji studii až do konce.

[2] Největším kandidátem na to, aby ESSE bylo pokládáno za lexikální kategorii, je ESSE existenční (tzv. there is-konstrukce); jsou nicméně dobré důvody pro to, aby i toto ESSE bylo analyzováno jako spona (viz Blaszczak(ová), 2008, mezi jinými taky Karlík, 2009).

[3] Thematický vokál (v tomto případě -e-) je vypouštěn, jestliže se dostane do přímého kontaktu s personálně-numerální koncovkou začínající vokálem: bud-e-u > budu; tato regularita je známa jako Jakobsonův zákon, viz Halle – Matushansky(ová) (2006).

Tabulka 1. Formy 1.os.sg. slovesa ESSE v přítomném čase

	imperfektivum	perfektivum
prézens	jsem	budu

$>$ hnětu, hnět-e-š, hnět-e-Ø…, a kořen $\sqrt{\text{BUD}}$ je nositelem rysu [+perfektivnost].[4] Kombinace rysu [–minulost], vyjadřovaného flexémem, a rysu [+perfektivnost] derivuje, jak známo, normálně tzv. futurální interpretaci morfologické struktury obsahující tyto rysy. U lexikálních sloves platí, že je-li rys [+perfektivnost] vlastností kořenu/kmenu, pak je přístupná interpretace, že událost začne a dosáhne završení po okamžiku promluvy, srov. *dá-m*, *bodn-u*, *koupí-m* (*Až mu dám ten dopis, zavolám ti*). Je-li rys [+perfektivnost] do struktury slova dodán mergováním lexikálního prefixu, je přístupná interpretace, že událost dosáhne završení po okamžiku promluvy, a nic se neříká o tom, zda probíhá v okamžiku promluvy, nebo neprobíhá: *do-píš-u* (*Až dopíšu ten dopis, zavolám ti*). Z toho lze predikovat interpretaci *bud-u*: protože *bud-u* vyjadřuje rysy [–minulost] [+perfektivnost] (a teď nedůležité phi-rysy) a rys [+perfektivnost] vyjadřuje kořenem/kmenem, mělo by poskytovat interpretaci futurální takovou jako *dám*, tj. že událost začne po okamžiku promluvy, tj. neprobíhá v okamžiku promluvy. Data to potvrzují: *Až budu psát ten dopis, zavolám ti.*

Infinitivní komplement, který *bud-u* bere, musí mít ovšem rys [–perfektivnost] (5), což znamená, že je-li tento aspektový požadavek splněn rysem jiného než lexikálního slovesa, např. modálního slovesa jako v (6), lexikální sloveso může už být přirozeně vidu nedokonavého i dokonavého:

(5) Petr bude $_{impf}$psát / $_{pf}$*napsat dopis

(6) Petr bude $_{impf}$muset $_{impf}$psát / $_{pf}$napsat dopis

Kontrast v (7), využívající durativní adjektiva, pak ukazuje, že komplex [bud-u MOD VP] může být interpretován jako imperfektivum (a), nebo jako perfektivum (b), v závislosti na aspektu lexikálního V:

(7) (a) Petr bude muset psát dopis hodinu / *za hodinu

(b) Petr bude muset napsat dopis *hodinu / za hodinu

V analytických modelech s rozšířenými projekcemi (minimálně lišící doménu lexikální a doménu/y funkční) se na základě těchto (a dalších) empirických dat standardně předpokládá, že ESSE$_{budu}$ je do struktury – jakožto funkční kategorie – mergováno

[4] Tato analýza je v souladu s diachronií. Kořen, který je vidět ve struktuře [$\sqrt{\text{BUD}}$-e-flexém] měl, jak známo, ve stsl. podobu $\sqrt{\text{BǪD}}$, a podle Whaley(ové), 2000, je nosovka o stsl. „následníkem“ nazálního konsonantu v pozdní protoindevropštině, tedy: $_{ide}$*$\sqrt{\text{BHŰ-N-D}}$ $>$ $_{\text{stsl}}\sqrt{\text{BǪD}}$ $>$ $_{\text{stč/mč}}\sqrt{\text{BUD}}$. Závažné je ovšem právě autorčino zjištění, že ide. nazální konsonant byl infix s rysem [+perfektivnost]. Ve stsl. je ostatně viditelný u několika sloves vyjadřujících inchoativnost (typ: *sędo*). Z toho vyplývá, že stsl. kořen $\sqrt{\text{BǪD}}$ byl nositelem deskriptivního významu („stát se“) a aspektového rysu [+perfektivnost] a v syntaktickém procesu (kdo chce, může mu říkat gramatikalizace) docházelo k „vyblednutí“ jeho deskriptivního významu, takže struktura [$\sqrt{\text{BUD}}$-e-flexém] je – po završení procesu gramatikalizace – nositelem jen rysů gramatických: [–minulost] [+perfektivnost] (a phi-rysů).

v doméně I, a to v hlavě funkční projekce F° pod I°, přičemž F°je hlavou hostící rys [aspekt: impf] a je pod hlavou hostící polaritu (negaci): *Petr nebude spát*.

Nyní můžu přistoupit k rozboru druhého členu subparadigmatu z Tab. 1, k ESSE *jsem*. Shoda panuje v tom, že ve struktuře [js-e-flexém] vyjadřuje sada flexémů rys [–minulost] a kořen $\sqrt{\text{JS}}$[5] je nositelem rysu [+imperfektivnost]. ESSE *jsem* je k dispozici ve dvou kontextech: (a) v kookurenci s l-participiem, tj. jako ESSE$_{aux}$, je interpretováno jako nositel phi-rysů a celá konstrukce má (dnes) ideosynkretický rys [+minulost][6], (b) v kookurenci s NP/DP, AP, PP a AdvP, tj. jako ESSE$_{cop}$, je interpretováno kompozicionálně na základě rysu flexému a kořenu jako nositel rysu [–minulost] (a phi-rysů), tedy stejně jako *toč-í-m*, srov. kontrast v (8):

(8) (a) Já jsem rozbil vázu včera / dnes / *zítra
(b) Já jsem učitel *včera / dnes / zítra

Připomenu, co každý ví, že *jsem…* v kontextu (a), tj. ESSE$_{aux}$, má jiné morfologické tvary než *jsem…* v kontextu (b), tj. ESSE$_{cop}$:

(9) (a) 1.os.sg.: já jsem / $*^{dial}$su chválil × (b) já jsem / ^{dial}su chválen / učitel / doma
^{substandard}já ~~jsem~~ chválil × *já ~~jsem~~ chválen / učitel / doma
2.os.sg. koho jsi / kohos chválil? × kým jsi / *kýms chválen?
chválil *(j)seš × ^{substandard}chválen (j)seš
3.os.sg. chválil Ø / *je × chválen / učitel /doma *Ø / je

A doplním ještě další notoricky známý fakt, že *jsem* v kontextu (a) je klitika, kdežto v kontextu (b) nikoli, viz kontrast v (10), a že v kontextu (a) nemůže být hostitelem negační částice *ne-*, zatímco v kontextu (b) jím být může, srov. kontrast v (11):

(10) (a) *chválil jsem – *jsem chválil*
chválen jsem – jsem chválen, vodníci jsou – jsou vodníci?
(11) (a) *chválil *nejsem – nechválil jsem*
(b) *chválen nejsem / nejsem učitel/em*

Co tato data prozrazují o konstrukci (3), kterou chci poznat? Připomenu, že z hlediska selekčních požadavků ESSE by konstrukce (3) – a taky (1), (2) a (4) – měly obsahovat ESSE funkční, tedy ukazující vlastnosti, jaké jsou typické pro ESSE$_{aux}$. ESSE v konstrukci (1), jak jsme viděli, tuto předpověď splňuje. Než přistoupím k rozboru (3), porovnám všechny konstrukce (1) – (4).

(a) Už první intuice, kterou má Čech o konstrukcích ESSE InfP v (1) – (4), napovídá, že (1) stojí v opozici k (2) – (4): v povrchové (= slyšitelné) struktuře vět (2) – (4) je přítomno méně elementů, než kolik jich při sémantické interpretaci těchto vět vyrozumíváme, zatímco v (1) tomu tak není. Zcela neformálně lze postihnout významy vět (2) – (4) parafrázemi:

(2) Všem lidem jest umříti

rozumíme tak, že „povrchový dativní subjekt věty musí vykonat to, co označuje infinitivní VP, aniž by byl vyjádřen původce modality nutnosti“

(3) Petr byl lovit ryby ze břehu

rozumíme tak, že „povrchový nominativní subjekt věty se vzdálil z místa X, ve větě neoznačeného a ani neoznačitelného, za účelem vykonání toho, co označuje infinitivní

[5] Kořen uvádím v této podobě, protože nemám důvod uvádět ho v podobě ide. nebo prasl.

[6] Diachronní výklad viz např. Kopečný (1955), nověji Migdalski (2006).

VP, na místě Y, a toto místo Y opustil, aniž by bylo vyjádřeno, zda to, co označuje VP, vykonal“

(4) Petr byl vidět ze břehu

rozumíme tak, že „povrchový nominativní subjekt věty bylo možné vizuálně vnímat, aniž by byl vyjádřen nositel možnosti vnímat a původce modality možnosti vnímat“

(b) Z hlediska povrchové syntaxe ukazuje unikátní vlastnosti konstrukce (2), majíc povrchový dativní subjekt, protože zbývající konstrukce mají nominativní subjekt.

(c) Mezi konstrukcemi s nominativním subjektem (1), (3), (4) má syntakticky specifické vlastnosti konstrukce (4), protože její povrchový nominativní subjekt je interním argumentem infinitivního slovesa, a jeho realizace v pozici subjektu slovesa ESSE je jen alternativou k primární realizaci v pozici akuzativního objektu; srov. analýzu v Caha & Karlík (2005). V (1) a (3) je naproti tomu nominativní subjekt interpretován jako subjektový argument infinitivního slovesa (který se posunul do [Spec IP]),[7] takže nahrazen akuzativním objektem být nemůže:

(4a) Ilona / Ilonu bude vidět z břehu

(1a) Ilona / *Ilonu bude lovit z břehu

(3a) Ilona / *Ilonu byla/*o lovit z břehu

(d) Podezření, že (1) a (3) by mohly být syntaktické struktury se společnými vlastnostmi, kterými se odlišují od (2) a (4), je podepřeno i tím, že sdílejí další vlastnost, kterou (2) a (4) nemají, totiž maximálně restriktivní časovou interpretaci: v konstrukci typu (1) je možné jen *budu*, zajišťující futurální interpretaci verbálního komplexu, v (3) je naopak dovolena určitě interpretace minulá, zajišťovaná tím, že je v ní konstrukce ESSE_{aux} *jsem* l-participium[8], a možná je dovolena interpretace přítomná, určitě je ale nepřístupná interpretace futurální:

(1b) Petr *byl / *je / bude lovit

(3b) Petr byl / ?je / *bude lovit

(2b) Všem lidem / bylo / je / bude umřít

(4b) Petr / Petra byl/-o / je / bude vidět

Hypotézu, že by (1) a (3) mohly být syntaktické struktury se společnými vlastnostmi, vyvracejí ovšem přesvědčivě data ukazující rozdíly mezi (1) a (3):

(e) V (1) má ESSE jen tvar *bud-u/-eš* ... (viz (1c)), kdežto v (3) má k dispozici nejen formy interpretované jako minulý čas slovesa ESSE, ale taky infinitiv a kondicionál, viz (3c):

(1c) Budu lovit sumce

(3c) Byl jsem lovit sumce / Musel jsem být lovit sumce / Byl bych lovit sumce

[7] Srov. taky různou interpretaci (3) *Ilona byla lovit* (= „Ilona je ta, kdo lovil“) × (4) *Ilona byla vidět* (≠ „Ilona je ta, kdo viděl“, nýbrž = „Ilona byla ta, koho bylo možno vidět“). S tím souvisí i kontrast ukazující komplementární distribuce infinitivních sloves: sloveso, které je možné v typu (4), je v typu (3) vyloučeno, a vice versa: (4) *Petr byl vidět / slyšet / cítit / *lovit / *pracovat / *nakupovat* × (3) *Petr byl *vidět / *slyšet / *cítit / lovit / pracovat / nakupovat.*

[8] Správnost požadavku, aby konstrukce měla význam minulosti, je podporována příkladem, který ukazuje, že v kombinaci s modálním slovesem musí mít modální sloveso tvar minulého času a pro konstrukci je přístupná pouze epistémická, nikoli deontická interpretace; srov. (3) *Petr musel být lovit / Petr *musí být lovit* × (4) *Petr musel být vidět / Petr musí být vidět*. Vysvětlení pro to zatím nemám.

(f) V (1) je ESSE mergováno nad modálním slovesem, takže modální sloveso má tvar infinitivu, zatímco v (3) je modální sloveso mergováno nad ESSE, takže tvar infinitivu má ESSE:

(1d) Ty budeš muset nakoupit

(3d) Ty jsi musel být nakoupit

(g) V (3) existuje omezení na komplementaci ESSE syntaktické: dovoleny jsou infinitivy sloves s externím argumentem, který má děj pod svou kontrolou (srov. (3e) × (3e')),[9] zatímco v (1) je jedinou podmínkou komplementace, aby infinitiv byl od nedokonavého slovesa, srov. (1e) × (1e'):

(3e) Petr byl chytat / nachytat ryby (e') Petr byl *závidět Marii
Petr byl *být pochválen ředitelem
Petr byl *jet do Prahy / *jít do kina
*Ono$_{expl}$ bylo pršet

(1e) Petr bude chytat / *nachytat ryby (e') Petr bude závidět Marii
Ono$_{Expl}$bude pršet

Zatím jsme dospěli k tomu, že každá z konstrukcí ESSE InfP z katalogu (1) – (4) má unikátní syntaktické vlastnosti, a tedy i unikátní syntaktickou strukturu a z ní derivovatelnou sémantickou interpretaci.

Pro další analýzu konstrukce typu (3) lze vyjít z kontrastu, který už jsme zaregistrovali (viz pozn. 8), a teď opakuji v (12): pouze (a) je s interpretací, o kterou nám jde, gramaticky správné (zatímco (a') je gramaticky správné s interpretací modální):

(12) (a) byl nakoupit × (a') *byl vidět

Nově využijeme pozorování, že pouze v (a), tj. v konstrukci typu (3), nikoli v (a'), tj. v konstrukci typu (4), je možné místo *byl* užít slovesa *šel* (s významovou změnou, viz dále):

(13) (a) byl nakoupit – (b) šel nakoupit
(a') *byl vidět – (b') *šel vidět

Lze tedy vyslovit hypotézu, že konstrukce ESSE InfP *Byl nakoupit* je v nějakém vztahu k MOVERE InfP *Šel nakoupit*[10] a deskriptivně formulovat tento vztah jako jednosměrnou implikaci:

Jestliže je možná konstrukce MOVERE InfP, tak je možná konstrukce ESSE InfP.

Pokud jde o sémantický vztah mezi (a) MOVERE InfP a (b) ESSE InfP, lze pozorovat toto:

Shoda mezi (a) MOVERE InfP a (b) ESSE InfP je v tom, že obě konstrukce interpretujeme tak, že povrchový nominativní subjekt věty se vzdálil z místa X, ve větě neoznačeného, za účelem vykonání toho, co označuje infinitivní VP, na místě Y, aniž

[9] S neakuzativem je – podle mnoha Čechů – konstrukce možná, pokud lze interní argument v pozici subjektu interpretovat tak, že má nad dějem kontrolu: *Teta byla hubnout v Lázních Lipová.*

[10] V tomto okamžiku taky prozrazuji, proč studie, kterou věnuji jubilantovi, zkoumá konstrukce typu *Petr byl lovit ryby*. Karel Pala stál totiž u kolébky české formální gramatiky a jeden z jeho dodnes vlivných a citovaných článků se zabývá v povrchové syntaxi viditelnými slovesy pohybu (Pala, 1972). Já teď přidávám rozbor struktur s v povrchové syntaxi sice neviditelným slovesem pohybu, ale obsahujících rys MOVERE ve své podkladové struktuře.

by bylo řečeno, zda událost popisovaná InfP nastala (to infinitivní morfologie slovesa umožňuje).[11]

(14) (a) Ilona šla/jela ráno nakoupit, ale protože měli zavřeno, nic nekoupila
(b) Ilona byla ráno nakoupit, ale protože měli zavřeno, nic nekoupila

S tím souvisí i to, že subjekt věty interpretujeme jako úmyslně jednající osobu, tedy mající děj pod kontrolou, a to nejen v případě dynamického slovesa MOVERE, ale i v případě statického slovesa ESSE. To ale nutně znamená, že v syntaktické struktuře ESSE-konstrukce musí být přítomna projekce, v níž je MOVERE vkládáno do struktury jako rys.

Sémantický rozdíl mezi ESSE InfP (a) a MOVERE InfP (b) je v tom, že (a) nutně interpretujeme tak, že subjekt dosáhl místa Y, kde se má uskutečnit událost denotovaná InfP, kdežto (b) nikoli.

(15) (a) *Ilona byla ráno nakoupit, ale potkala kamarádku, takže do obchodu se vůbec nedostala
(b) Ilona šla/jela ráno nakoupit, ale potkala kamarádku, takže do obchodu se vůbec nedostala

Navíc ESSE InfP (a) interpretujeme nejen tak, že subjekt dosáhl místa Y, kde se má uskutečnit událost popisovaná InfP, ale že místo Y opustil:

(15) (a') *Ilona byla ráno nakoupit a ještě je tam.

Pokud jde o morfosyntaktické vlastnosti MOVERE InfP (a) a ESSE InfP (b), shoda je v tom, že sloveso v InfP může být v (a) i (b) dokonavé i nedokonavé:

(16) (a) Ilona šla ráno nakoupit / nakupovat
(b) Ilona byla ráno nakoupit / nakupovat

První rozdíl je v tom, že MOVERE je možné ve všech časech, kdežto ESSE asi jen v préteritu (což už víme):

(17) (a) Ilona šla / jde / půjde nakoupit / nakupovat
(b) Ilona byla / ?je / *bude nakoupit / nakupovat

Druhý rozdíl je v tom, že MOVERE je možné ve všech modech, kdežto ESSE není možné v imperativu:

(18) (a) Ty běž nakoupit / nakupovat
(b) *Ty buď nakoupit / nakupovat

Třetí rozdíl je v tom, že MOVERE je možné v infinitivu bez omezení (19a), která jsou pro infinitiv u ESSE: ESSE-konstrukce je v infinitivu možná jen jako komplement modálního slovesa, které umí vyjádřit rys [+minulost] (19b), viz už pozn. 9:

(19) (a) Musel jít nakoupit / Rozhodl se jít nakoupit / Je příjemné jít nakoupit
(b) Musel být nakoupit / *Rozhodl se být nakoupit / *Je příjemné být nakoupit

Čtvrtý rozdíl je v tom, že v MOVERE InfP (asi) mohou být ve struktuře dvě časová adverbia, kdežto v ESSE InfP (určitě) jen jedno:

(20) (a) Jeli jsme hodinu jet na lodičce hodinu
(b) *Byli jsme hodinu jet na lodičce hodinu

Pro další analýzu využiju pozorování, že ESSE kookuruje jen s lokálními adjunkty nedirekcionálními (*Petr byl doma* × **Petr byl domů*), kdežto MOVERE zase jen s adjunkty direkcionálními (**Petr šel doma* × *Petr šel domů*). Tato data můžeme použít

[11] K syntaxi infinitivu viz Karlík & Veselovská (v tisku); tam taky nabízíme vysvětlení, v jakých kontextech musí mít sloveso infinitivní morfologii a čím je to motivováno.

k tomu, abychom zjistili, kolik lokálních adjunktů a jaké jsou k dispozici v MOVERE InfP (a) a ESSE InfP (b):

(21) (a) Jeli jsme obědvat do Valtic / *ve Valticích[12]

(b) Byli jsme obědvat do *Valtic / ve Valticích

Na základě příkladů s lokálními adjunkty lze přidat další dílek k poznání syntaktické struktury ESSE InfP. Z nezávislých dat totiž víme, jak se lokální adjunkty chovají. Jak ukazují příklady (22), v konstrukci ESSE -n-/-t-PassivPart, která má defaultovou interpretaci „dějovou“, je možný adjunkt lokální i direkcionální (rozdíl sémantický teď není důležitý) (a), zatímco v konstrukci ESSE -n-/-t-DeverbAdj, která má defaultovou interpretaci „rezultativní“, je možný jen adjunkt lokální (b).

(22) (a) Petr byl schován ve sklepě / do sklepa

(b) Petr byl schovaný ve sklepě / *do sklepa

Srovnáme-li data v (22) s daty v (23),

(23) Ilona byla nakoupit v Tescu / *do Tesca

máme dobrý důvod předpokládat, že (23) má taky interperataci rezultativní.

Mezi (22a) s dějovou interpretací a (22b) s rezultativní interpretací je ještě jeden důležitý rozdíl. Jak ukazuje (24), v (a) struktuře je možné jen časové adverbium typu *hodinu*, kdežto v (b) jen *za hodinu*:

(24) (a) Petr byl schován do sklepa *hodinu / za hodinu

(b) Petr byl schovaný ve sklepě hodinu / *za hodinu

Jsou-li tato adverbia (spolehlivým) testovacím prostředkem k rozlišení telicity (*za hodinu*) od atelicity (*hodinu*), pak v (24a) adverbium modifikuje trvání „telického děje“ („od jeho začátku do jeho završení“), označeného dokonavým participiem. Naproti tomu v (24b) adverbium nemodifikuje trvání „telického děje“ („od jeho začátku do jeho završení“), označeného dokonavým participiem, nýbrž trvání stavu rezultátu. Co je pozoruhodné, je to, že pokud přidáme do struktury ESSE InfP a MOVERE InfP časová adverbia, obě se chovají stejně:

(25) Otec byl nakoupit *za dvě hodiny / dvě hodiny

(26) Otec šel nakoupit *za dvě hodiny / dvě hodiny

Srov. rozdíl proti strukturám bez ESSE / MOVERE, v nichž je taky sloveso dokonavého vidu:

(27) Otec nakoupí za dvě hodiny / *dvě hodiny

Z toho ovšem vyplývá, že časové adverbium v obou případech modifikuje porci struktury [MOVERE / ESSE InfP), jejíž hlavou je MOVE/ESSE, tedy že v obou konstrukcích je ESSE resp. MOVERE je vkládáno do struktury ve funkční doméně IP. [13]

---

[12] Jestliže jsou ve větě dvě lokální PP/AdvP (*Petr jel do Olympie nakoupit v Tescu / Petr byl v Olympii nakoupit v Tescu*), pak z toho musí vyplývat interpretace, že obě lokální určení musejí být ve vztahu inkluze, tj. druhé místo je obsaženo v prvním.

[13] Z toho ale současně vyplývá, že nejen ESSE, ale i MOVERE má zde status funkční kategorie (což u sloves pohybu nepřekvapuje, viz data: v (i) je MOVERE lexikální sloveso, v (ii) a (iii) je to funkční sloveso, a tomu odpovídají i rozdíly v syntaxi (a) – (aa”) a rozdíly v sémantické interpretaci:

(i) Petr půjde nakoupit
(a) Petře, pojď nakoupit
(a’) *Petra půjde zítra do kina
(a”) Petr půjde do školy

Umíme teď už derivovat téměř všechny významy ESSE konstrukce (3), a technickým problémem by bylo vytvořit analytickou reprezentaci (syntaktický strom), z níž by se tyto významy daly odvodit. To není úkol snadný, a ani to nemůže být úkolem této studie. Nadto se mi nepodařilo vysvětlit, co je zodpovědné za to, že rezultativní strukturu ESSE InfP typu (3) interpretujeme tak, že gramatický subjekt místo, kde byl za účelem vykonání toho, co popisuje InfP, opustil.

Druhý úkol této studie byl iniciován studií de Groota (2000), který při analýze struktur ESSE InfP s významem absentivu, tedy s významem (přibližně) „subjekt se vzdálil z místa X a je nepřítomen“, který není příliš vzdálený od významu, který má česká konstrukce typu (3) *Petr byl lovit ryby*, zjistil, že tato konstrukce existuje v osmi evropských jazycích (viz (28), cit. podle de Groot, 2000:695-6), češtinu ale neuvádí:

(28) Jan ist boxen — němčina
Jan ESSE INF
János boxolni van — maďarština
János INF ESSE
Jan is boksen — holandština
Jan ESSE INF
Gianni è a boxare — italština
Gianni ESSE PART INF
Jan as tu boksin — frízština
Jan ESSE PART INF
Jussi on nykkeile-mä-ssä — finština
Jussi ESSE INF-Casus:Inessiv
Jan er og boksar — norština
Jan ESSE AND INF
John är och boxar — švédština
John ESSE AND INF

Nechám stranou to, že absence češtiny mezi jazyky, které uvádí de Groot jako jazyky mající absentiv, přirozeně vybízí k prověření toho, zda jsou de Grootovy údaje správné. Jde teď o toto: Pozitivní evidence, že existuje (minimálně) devět (osm de Grootových + čeština) evropských jazyků, v nichž je doložena konstrukce ESSE InfP se společným významem absentivu (a v různých jazycích s možnými variantními významy dalšími), vede k otázce, zda jde o projev toho, co se obecně označuje jako jazyky

(ii) Petr půjde přesvědčit
(a) *Petře, pojď přesvědčit
(a') Petra půjde přesvědčit
(a") *Petr půjde do školy
(iii) *Petr si půjde tykat
(a) Petře, pojď si tykat
(a') *Petra si půjde tykat
(a") *Petr si půjde do školy

Dobrým příkladem, který demonstruje, že v MOVERE InfP je MOVERE funkční slovesem, je otázka: *Kdy půjdeš nakoupit?* Ta totiž zřejmě nemusí nutně startovat presupozici, že někdo MOVERE ve smyslu „jít“ coby lexikální kategorie „pohybovat se (krokem) po vlastních nohou“, ale presupozici, že někdo MOVERE, tj. třeba ve smyslu „jet autem“ apod. Ostatně, srov. přesvědčivou analýzu MOVERE v angl. (v americké angl.) nekoordinačních konstrukcích typu *I go and buy bread* u Carden & Pesetsky (1977), nebo v ital. (dialektech) konstrukcích typu *Vaju a pigghiu u pani* u Cardinaletti(ové) & Giusti(ové), 2001, jako funkčních kategorií.

v integraci. Na první pohled nabízející se podporu pro její kladné zodpovězení poskytuje fakt, že mezi 9 evropskými jazyky s ESSE InfP s významem blízkým absentivu jsou jazyky indoevropské i neindoevropské (finština, maďarština), což minimalizuje možnost vykládat shodu na bázi genetické příbuznosti.

Podíváme-li se však na oněch 8 konstrukcí označovaných funkcionalistou de Grootem jako absentiv, snadno bez jakékoli teorie zjistíme, že z hlediska formální syntaxe to vůbec nejsou stejné struktury. Lze je rozdělit (minimálně) – už na základě viditelné syntaxe – do čtyř skupin, a to podle toho, jakou formu má INF: (a) „holý" INF (čeština, němčina, maďarština, holandština), (b) partikule + INF (italština, frízština), (c) INF + pádový morfém inessivu (finština), (d) spojka AND + INF (norština, švédština).

Čeština se ovšem od všech konstrukcí, tedy i od zbývajících konstrukcí s povrchově identickou strukturou (a), liší tím, že ESSE je (asi) možné jen v minulém čase, jinde je (určitě) možné i v přítomném čase a v budoucím čase. Srov. rozdíly (aspoň s němčinou a italštinou):[14]

(29) Petr byl / ?je / *bude boxovat
Petr è / è stato / era / sarà a boxare
Petr ist / war boxen / Petr wird boxen sein

Na druhé straně asi není náhodné, že ESSE v češtině ani ital. a něm. není možné v imperativu:

(30) Petře, *buď boxovat
Pietro, *sii a boxare
Peter, *sei boxen!

A co je mimořádně důležité, je to, že v češtině, italštině i němčině není konstrukce ESSE InfP možná jako komplement kontrolových sloves (31), ale je možná jako komplement sloves modálních, ve všech třech případech jen s epistémickým čtením (32). To nemůže být náhoda:

(31) Petr slíbil *být boxovat
Pietro ha promesso *di essere a boxare
Petr versprach (der Mama) *boxen zu sein.

(32) Petr musel / *musí být boxovat
Pietro doveda / deve essere a boxare
Petr musste / muss boxen sein.

Závažným argumentem, který svědčí o tom, že česká konstrukce *Petr byl boxovat* má zcela jinou syntaktickou strukturu než de Grootovy absentivy, a to zase i ty absentivy, které povrchově vypadají stejně jako *Petr byl boxovat*, tj. struktury (a) ESSE „holý" INF, je rozdílná sémantická interpretace.

(a) Sémantickým rysem, který má česká konstrukce *Petr byl boxovat*, ale který postrádají de Grootovy konstrukce ESSE InfP, je rezultativnost.

(b) Sémantickým rysem, který je společný pro de Grootovy konstrukce ESSE InfP, je absentiv, ale tento rys česká konstrukce *Petr byl boxovat* nemá.

(c) Všechny de Grootovy konstrukce ESSE InfP jsou na základě formy interpretovatelné tak, že vyjadřují jedno nastání události popisované InfP. Rodilí mluvčí němčiny a

[14] Za kontrolu italských vět děkuji A. M. Perissutti(ové), za kontrolu vět německých baronu R. von Waldenfelsovi.

italštiny nicméně tvrdí, že z nich lze odvodit,[15] že subjekt InfP vykonává děj popisovaný VP pravidelně/vícekrát. Italská, resp. německá absentivní konstrukce *Gianni è a boxare*, *Jan ist boxen* tedy implikuje, že Jan pravidelně chodí/jezdí boxovat. V české rezultativní konstrukci *Jan byl boxovat* je však taková interpretace nepřístupná.

Závěr: Z konfrontace už jen dobře vnímatelných syntaktických a sémantických vlastností konstrukcí ESSE InfP, které se pokládají za absentiv, a české konstrukce typu *Petr byl boxovat*, lze dedukovat, že jde o konstrukce natolik různé, že není důvod pokládat je za projev toho, co se obecně označuje jako jazyky v integraci. Silným argumentem pro to, že tyto konstrukce lze derivovat na základě univerzálněgramatických principů je ale to, že modální slovesa v nich nutně mají epistémickou interpretaci. (Bohužel jsem neměl k dispozici studii T. Bergera, který se těmito konstrukcemi taky zabýval.)

Reference

1. BLASZCZAK, J. (2008): What HAS to BE used? Existential, Locative, and Possessive Sentences in Polish. In: Antonenko, A. & J. F. Bailyn ad. (eds.): *Formal Approaches to Slavic Languages*, 16. Ann Arbor: Michigan Slavic Publ., s. 31-47.
2. DOKULIL, M. (1949): *Byl jsem se koupat, naši byli vázat*. *NŘ*, 33, s. 81-92.
3. CAHA, P. & P. KARLÍK (2005): *Je vidět Sněžku*: Searching Modality. *SPFFBU*, A 53, s. 103-114.
4. CARDINALETTI, A. & G. GIUSTI (2001): "Semi-lexical"motion verbs in Romance and Germanic. In: Corver, N. & H. van Riemsdijk (eds.): *Semi-lexical categories. On the function of content words and the content of function words.* Berlin ad.: Mouton de Gruyter, 2001, s. 371-414.
5. CARDEN, G. & D. PESETSKY (1977): Double-Verb Constructions, Markedness, and a Fake Coordination. *Papers form the Thirteents Regional Meeting of the Chicago Linguistic Society*, 13, s. 82-92.
6. DE GROOT, C. (2000): The absentive. In: Dahl, Ö. (ed.): *Tense and Aspect in the Languages of Europe.* Berlin ad.: Mouton de Gruyter, s. 693-719.
7. HALLE, M. & O. MATUSHANSKY (2006): The Morphophonology of Russian Adjectival Inflection. *Linguistics Inquiry*, 37, s. 351-404.
8. JUNGHANNS, U. (1997): On *byt'* (and *byti*). In: Junghanns, U. & G. Zybatow (eds.): *Formale Slavistik* (= *Leipziger Schriften zur Kultur-,Literatur-, Sprach- und Übersetzungswissenschaft*; 7). Frankfurt am Main: Vervuert Verlag, s. 251-265.
9. KARLÍK, P. (2009): Sloveso *být* v češtině a jinde. In: Rusinová, E. (ed.): *Přednášky a besedy z XLII. běhu LŠSS.* Brno: MU, s. 83-92.
10. KARLÍK, P. & L. VESELOVSKÁ (v tisku): Infinitive Puzzle. In: Dočekal, M. & M. Ziková (eds.): *Czech in Formal Grammar*.
11. KOPEČNÝ, F. (1955): Problém českého „příčestí minulého činného" v historii českého mluvnictví. In: *Sborník v čest na akad. A. Teodorov-Balan*. Sofia, s. 293-300.
12. MIGDALSKI, K. (2006): *The Syntax of Compound Tenses in Slavic*. Utrecht: LOT.
13. PALA, K. (1972): Některé vztahy mezi významem a syntaktickými vlastnostmi českých sloves pohybu. *ČJ*, 23, s. 169-176.
14. WHALEY, M. (2000): *The Evolution of the Slavic ?Be(come)'-type Compound Future.* PhD. Diss., The Ohio State University.

[15] Princip, na základě něhož to lze ze struktury popisující jednu nastavší/nastávající událost, odvodit, nemohu teď analyzovat.

Statistical Collocability of Russian Verbs

M. Khokhlova[1,2] and V. Zakharov[1,2,3]

[1] Faculty of Philology and Arts, St. Petersburg State University, Russia
[2] Institute for Linguistic Studies of the Russian Academy of Sciences, St. Petersburg, Russia
[3] Erasmus Mundus International NLP & HLT Masters Programme, Universitat Autònoma de Barcelona, Spain

Abstract. The paper deals with the results of a corpus-based study of collocational behaviour of the most frequent Russian verbs. Association measures (MI, t-score, and log-likelihood) are suggested as instruments to deduce verbs' collocability and grammatical patterns. Given a set of collocates for each verb, it is possible to infer about its distribution. The paper discusses the obtained results and the measures used as applied to the problem. Directions of the further perfection are offered.

1 Introduction

The issue of collocability is highly important in modern linguistics. The investigation of collocability is closely connected to the study of syntagmatics as a deeper level of lexical relations. Though several criteria have been taken into consideration to describe these relations (lexical restrictions, repeatability etc.), the boundary between free and set phrases often has been placed quite subjectively.

The notion of *collocation* was introduced by the founder of London School of Structural Linguistics and the representative of British contextualism J.R. Firth [1]. The word meaning, in Firth's opinion, is closely connected with its ability to collocability. Collocation is a tendency of a word to a certain environment. Thus, he stated the hypothesis according to which it is possible for a word to be attributed to a group by its neighbourhood. The parts of collocation occupy certain positions and, thus, are characterized by mutual expectancy of appearance. Collocations can be viewed as forms of meaning [1].

It is also possible to apply the lexicographic approach to studying the phenomenon of collocation [2, 3, 4]. While in British contextualism collocation is defined on the basis of statistical assumptions about the probability of co-occurrence of two (or more) lexemes, and especially frequent combinations of lexical units are considered as collocations, the lexicographic approach considers collocation as a semantic-syntactic unit.

Although the term *collocation* appeared in Russian linguistics long ago [5], it is not generally recognized by Russian scholars and even is absent in the Russian Linguistic Dictionary [6]. There is no agreement among scholars how to call such lexical units; cf. "set verbal-noun expressions" [7], "analytic lexical collocations" [8], etc. The majority of modern authors understand under a collocation a statistically set phrase. Collocations can be put between free phrases and idioms on a scale of phrases. The monograph

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 125–133, 2009.

by E. Borisova [9] has proved to be the first work in Russian linguistics, completely devoted to the research of the concept of collocation on a material of Russian. One of the key properties of a collocation is "the impossibility of prediction of such combinations on the basis of meanings of their components" [9: 13]. In the "Meaning–Text" theory [10] collocations are considered as a subclass of more extensive class of set phrases, or phrasemes.

In our sense collocations should be defined simply as statistically set phrases.

2 Statistical background

Statistical methods are widely used in corpus linguistics. Nowadays there are several ways in linguistics to calculate the degree of collocates' coherence. There are different measures which calculate a degree of words nearness in a text, namely, MI (mutual information), t-score, log-likelihood (henceforth LL), z-score, chi-square, Dice coefficient, odds ratio etc. [11]. There are also other statistical methods supplementing measures of association. E.g., in [12] there is a description of a method to extract collocations based on bigrams rank distribution. The results of the method's application are compared with results obtained by measures of association.

It should be realized that statistical measures enable to find phrases of various nature, e.g., frequent free phrases as well as phrasemes (cf. lexical functions in Mel'čuk's theory "Meaning–Text"). For instance, the free word combination крепкий чай (strong tea) is usually not recorded in dictionaries. Nevertheless foreign speakers of Russian have to remember that the Russian usage here is крепкий (strong) rather than сильный (powerful). Such statistical data about the strength of syntagmatic relations between words may be useful in translation studies or language teaching and can be used as a raw material for a lexicographers. Probably, various measures extract word combinations of different types.

So far the association measures proposed in most works have not been evaluated in the perspective of extracting collocations in Russian. Unfortunately there are not so many Russian corpora having this kind of tools implemented for them. We can mention here only the corpora built at the University of Leeds by S. Sharoff[4] and the enthusiastic project by AOT team[5].

3 The goal and methods of study

The object of our research in the given work are the 10 most frequent verbs in Russian (see the Electronic Frequency Dictionary of Russian by S. Sharoff [13]). They are as follows: быть (be), сказать (say), мочь (can), говорить (speak), знать (know), стать (become), есть (eat), хотеть (want), видеть (see), идти (go). The study deals with finding bigrams for these verbs. The goal is to study collocational properties of these lexical items and to define opportunities of statistical methods as a whole and several measures in particular. The research has been led on a corpus created at the University of

[4] http://corpus1.leeds.ac.uk/ruscorpora.html
[5] http://aot.ru/demo/bigrams.html

Leeds on the base of texts from the Russian National Corpus (about 50 mln tokens) under the guidance of S. Sharoff.

In a search mode one can choose several statistical measures (namely, MI, t-score, LL), set a position of collocate and a span in words, it is also possible to set a part of speech of a collocate and to search either word-forms or lemmas.

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

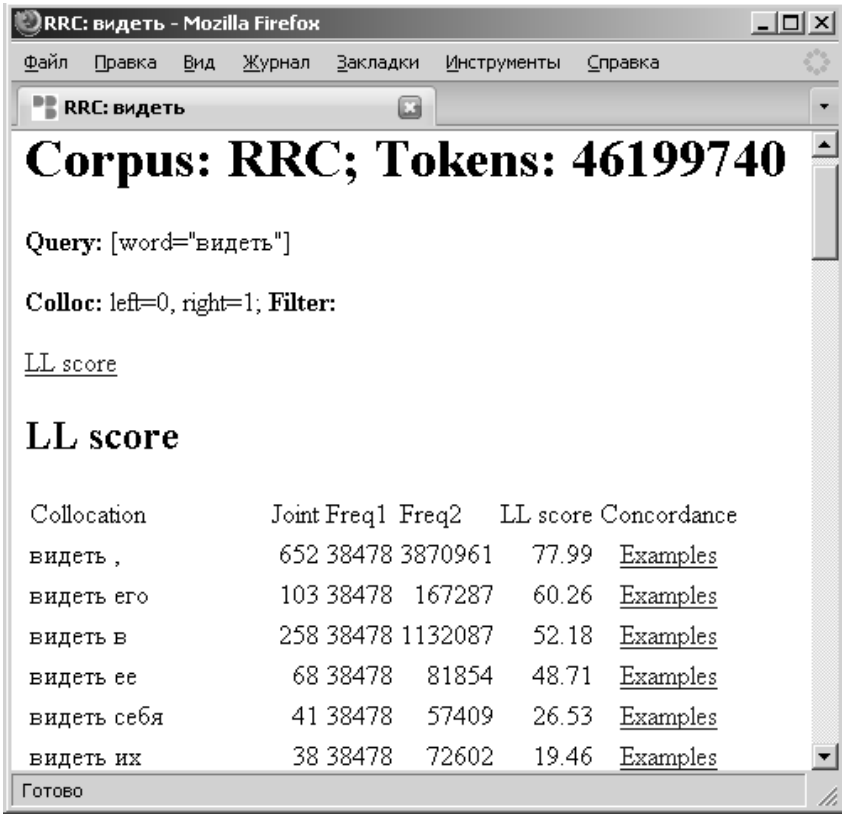


Fig. 1. Collocations for видеть in accordance with LL-measure.

It is necessary to mention that each element of the corpus including punctuation marks which stands before or after a blank is considered as a token. Hence there are meaningless combinations of verbs and punctuation marks, too.

Then the results obtained were brought to the one table representing three measures altogether with subtables for certain syntactic formulas (Adv+V, N+V, V+Adv, etc.) (see Table 1). Compilation of the above tables was made manually. At the same time the senseless collocations have been removed, e.g. the combinations of verbs

with conjunctions, interjections, particles, prepositions and punctuation marks. Every collocation has been assigned its rank.

Table 1. A fragment of combined results table for verb говорить (Adv+V) [6]

	Collocation	**Joint**	**Freq1**	**Rank MI**	**MI score (7,08–2,14)**	**Rank LL**	**LL score (1064,06–2,96)**	**Rank T-score**	**T-score (22,79–1,96)**
1.	честно говорить	527	2339	1.	7,08	1.	1064,06	101.	1,96
2.	постоянно говорить	62	4158	2.	7,04	14.	40,59	85.	2,26
3.	условно говорить	90	585	3.	6,53	8.	162,73	91.	2,11
4.	обиженно говорить	5	208	4.	6,46	77.	4,37	16.	6,52
5.	грубо говорить	130	988	5.	6,30	6.	224,23	93.	2,10
6.	умело говорить	23	2034	6.	6,20	33.	12,26	70.	2,40
7.	откровенно говорить	139	1203	7.	6,12	5.	230,24	94.	2,09
8.	собственно говорить	333	3114	8.	6,00	2.	538,32	99.	1,97
9.	жалобно говорить	6	481	9.	5,91	94.	3,45	25.	4,96
10.	.....								

The analysis of data of Table 1 (in total 101 collocations) shows that collocation ranks obtained by different measures do not coincide. The t-score measure behaves most differently, in contrast to MI and LL measures that often show similar results here (see Table 1, lines 1, 3, 5, 7, 8) as well as in tables for other verbs and other syntactic constructions.

Further the various analysis of data obtained was carried out; the results are partially given below.

4 Results and Discussion

Below there are fragments of collocations (a verb and a collocate) found for the verb сказать (sorted by measure LL):

4.1 Right context

V + Adv (Pred)

Collocation	Joint	Freq1	LL	MI	T-score
сказать трудно	54	6980	56.37	4.38	7.00
сказать нельзя	34	11557	20.33	2.98	5.09
сказать точно	30	9893	18.34	3.03	4.81
сказать вслух	15	1470	17.65	4.78	3.73
сказать особо	16	2012	16.90	4.42	3.81
сказать честно	14	2339	12.90	4.01	3.51

[6] Here and further 'Freq1' is frequency of a collocate

V + N

Collocation	Joint	Freq1	LL	MI	T-score
сказать правда	51	16453	31.68	3.06	6.28
сказать слово	53	41070	14.13	1.79	5.18
сказать гадость	5	390	6.44	5.11	2.17
сказать комплимент	4	387	4.73	4.80	1.93
сказать неправда	5	837	4.60	4.01	2.10
сказать тост	4	589	3.92	1.89	4.19

4.2 Left context

Adv (Pred) + V

Collocation	Joint	Freq1	LL	MI	T-score
можно сказать	1675	36011	3274.07	6.97	40.60
надо сказать	1046	37243	1768.42	31.91	6.24
нельзя сказать	315	11557	524.47	6.19	17.51
трудно сказать	279	6980	518.11	6.75	16.55
точно сказать	138	9893	183.63	5.23	11.43
нужно сказать	126	12993	145.33	4.70	10.79

V + V

Collocation	Joint	Freq1	LL	MI	T-score
хотеть сказать	1393	37897	2550.77	6.63	36.95
хотеться сказать	173	10127	247.53	5.52	12.87
успевать сказать	74	8437	81.72	4.56	8.24
следовать сказать	87	18608	70.24	3.65	8.59
забывать сказать	72	11464	68.04	4.08	7.98
смочь сказать	46	6167	47.21	4.43	6.44

For lack of space we present below the examples for one verb only, namely, говорить. As for the t-score a surprising steady, consistent pattern has been observed, its value being inversely proportional to the frequency of a collocation (see Table 2).

We have also compared statistical collocations obtained with dictionaries. For all collocations and for all verbs the same tendency has been observed: most of collocations (phrasemes) recorded in dictionaries fall on the top part of the list composed by one of the association measure application.

As for the verb говорить the Russian dictionaries specifically single out set combinations formed with this verb's adverbial participle говоря (saying). We have changed the conditions of the experiment accordingly. The verb говорить was presented with its exact form говоря on the search over the corpus. Thus, new, smaller lists of collocations have been obtained with new values of association measures (see table 3).

The analysis of the results obtained has shown that in this case concrete collocations overwhelmingly have significantly bigger value for all association measures. We can conclude that collocation extraction by statistical measures sometimes has to be done on the word form level rather then on the level of lemmas.

The collocations by the MI measure have been also extracted for the verb говорить (Adv+N) over the corpus of the well-known Moshkov library (680 mln. words) on the

Table 2. A fragment of combined results table for verb говорить (Adv+N), sorted by t-score.

Collocation	Joint	Freq1	Rank MI	MI score (7,08–2,14)	Rank LL	LL score (1064,06–2,96)	Rank T-score	T-score (22,79–1,96)
умоляюще говорить	4	85	15.	4,82	65.	4,80	1.	22,79
примирительно говорить	4	111	23.	4,43	80.	4,27	2.	17,96
скупо говорить	4	138	31.	4,12	86.	3,85	4.	15,46
……..								
восхищенно говорить	9	149	69.	2,91	34.	11,91	39.	3,24
неуверенно говорить	10	615	53.	3,25	49.	6,94	45.	3,01
убедительно говорить	10	502	17.	4,76	47.	7,87	44.	3,01
смело говорить	10	782	58.	3,09	56.	5,87	46.	2,99
……								
коротко говорить	267	6540	18.	4,61	4.	301,73	100.	1,97
собственно говорить	333	3114	8.	6,00	2.	538,32	99.	1,97
честно говорить	527	2339	1.	7,08	1.	1064,06	101.	1,96

Table 3. Association measures for verb говорить in the adverbial participle form (first an old value (for the lemma) / then, in italics a new value (for the word form)).

Collocation	Joint	Freq1	MI score	LL score	T score
искренне говоря	12/5	1562	2,94/4.92	4,49/6.11	2,74/2.16
точно говоря	66/41	9893	2,64/5.29	21,09/55.31	2,21/6.24
просто говоря	214/144	28089	2,19/5.60	79,38/209.98	2,02/11.75
откровенно говоря	139/104	1203	6,12/9.67	230,24/299.54	2,09/10.19
честно говоря	527/429	2339	7,08/10.98	1064,06/1690.55	1,96/22.33
объективно говоря	7/6	502	4,24/6.82	4,37/11.22	4,16/2.43
образный говоря	51/44	233	3,00/10.80	102,07/145.01	2,32/6.63
строгий говоря	166/146	4239	4,55/8.34	184,16/351.80	2,08/12.05
условно говоря	90/87	585	6,53/10.45	162,73/275.55	2,11/9.32
грубо говорить	130/127	988	6,30/10.24	224,23/392.55	2,10/11.26
мягко говоря	252/247	3916	5,27/9.22	341,77/672.86	2,01/15.69
коротко говоря	267/265	6540	4,61/8.58	301,73/662.08	1,97/16.24
собственно говоря	333/332	3114	6,00/9.97	538,32/996.77	1,97/18.20
упрощенно говоря	5/5	34	3,07/10.44	8,92/15.78	7,53/2.23

above mentioned site AOT. The comparative analysis of the two lists has shown their good coincidence in the upper part of the table (see Table 4).

At the same time for some collocations the MI indices prove to differ significantly. It is possibly connected with the fact that in Moshkov Library fiction is mostly presented while the Sharoff's corpus is more representative.

Table 4. Comparison of collocation ranks for verb говорить (Adv+N) got under MI measure calculated on base of two corpora of Russian.

Collocation	**Joint**	**Freq1**	**Rank MI**	**MI score**	**Joint AOT**	**Freq1 AOT**	**Rank AOT**	**AOT MI**
честно говорить	527	2339	**1.**	7,08	6294	21845	**2.**	7.316870
постоянно говорить	62	4158	**2.**	7,04				
условно говорить	90	585	**3.**	6,53				
обиженно говорить	5	208	**4.**	6,46				
грубо говорить	130	988	**5.**	6,30	11866	558	**11.**	4.701699
умело говорить	23	2034	**6.**	6,20				
откровенно говорить	139	1203	**7.**	6,12	2589	13037	**3.**	6.779979
собственно говорить	333	3114	**8.**	6,00	5220	28468	**4.**	6.664905
жалобно говорить	6	481	**9.**	5,91				
охотно говорить	13	1231	**10.**	5,44				
истинно говорить	33	2693	**11.**	5,32	482	5777	**5.**	5.528909
мягко говорить	252	3916	**12.**	5,27	22714	1165	**10.**	4.826944
……..								

5 Conclusion and Further Work

The experiment has shown the possibility to apply statistical tools in order to extract collocations from Russian texts. The results of this work (and the data about word collocability based on statistical measures), first of all, can be applied to dictionary compiling.

Yet it is difficult to decide on one statistical measure that could give allegedly perfect results. The work done by us has demonstrated significant discrepancies in different measures values for the same collocations. This points to the fact that the comparative functionality analysis of different measures should be continued. Maybe a combination of some measures could be used, e.g., sum of rank, and so on. We did try to sum MI and LL rank, and in some case obtained value shows good results. The results obtained leave open the question in which form the collocation elements have to be taken into account when calculating statistical association measures (see Section 4, Table 3).

Another task is to choose a threshold for each measure that can for sure indicate whether the phrase is a collocation or not.

One should take into account structural formulas which underlie collocations, too. The programmatic ways of eliminating from data combinations of so called "stop words" and punctuation marks (or just skipping them) have their importance, too.

Though our experiments and results are promising, nevertheless, they are only the first steps to demonstrate statistical methods and adopt them for the linguistic praxis. We are convinced that probabilistic-statistical methods should be more specified and additional programming tools are to be developed in the corpora study frameworks.

References

1. Firth, J.R.: Papers in Linguistics 1934–1951. London (1957)
2. Benson, M.: Collocations and idioms. In: Ilson, R. (ed.) Dictionaries, lexicography and language learning, pp. 61–68. Oxford (1985)
3. Hausmann F.J.: Kollokationen im deutschen Wörterbuch: ein Beitrag zur Theorie des lexicographischen Beispiels. In Bergenholtz, H. and Mugdon, J. (eds.) Lexicographie und Grammatik. Tübingen (1985)
4. Cowie, A.P.: General Introduction. In: Cowie, A.P. et al. (eds.) Oxford Dictionary of Current Idiomatic English, Volume 2, pp. 10–17. Oxford University Press, Oxford (1983)
5. Akhmanova, O.S.: Slovar' lingvisticheskikh terminov. Moscow (1966)
6. Yartseva, V.N, (ed.): Lingvisticheskiy enciklopedicheskij slovar'. Moscow (1990)
7. Deribas, V.M.: Ustoychivye glagol'no-imennye slovosochetaniya russkogo yazyka. Moscow (1983).
8. Teliya, V.N.: Russkaya frazeologiya: semanticheskiy, pragmaticheskiy i lingvokul'turologicheskiy aspekty. Moscow (1996).
9. Borisova, E.G.: Kollokacii. Chto eto takoye i kak ikh izuchat'. Moscow (1995)
10. Mel'chuk I.A.: Opyt lingvisticheskoy teorii "Smysl – Tekst". Moscow (1974).
11. Evert, S.: The Statistics of Word Cooccurences Word Pairs and Collocations. PhD thesis. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart (2004).
12. Cvrček, V.: Metoda zjišťování kolokační platnosti frekventovaných bigramů pomocí ranku. In Čermák F. (ed.) Kolokace.,p p. 36–55. Ústav Českého národního korpusu, Praha (2006)
13. Sharoff S.: Chastotnyy slovar' russkogo yazyka, `http://www.artint.ru/projects/frqlist.asp`

Supporting Visually Impaired People in Accessible Image Exploration and Creation of Personal Web Presentations

Ivan Kopeček and Fedor Tiršel

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
kopecek, xtirsel@fi.muni.cz

Abstract. This paper describes a framework enabling the visually impaired people to generate their personal web presentations with accessible images. A brief outline of the GATE (= Graphics Accessible to Everyone) system and BWG (= Browser WebGen) project is presented. The framework integrates these two systems in dialogue strategies, computer graphics, sonification of images and ontologies. We also present some examples and illustrations.

1 Introduction

Dialogue-based processing of graphics targets the applications that are convenient for inexperienced and handicapped, especially visually impaired, users. Up to now, the accessibility of graphics for the blind and visually impaired has been mostly connected with the use of tactile devices or use of sonification.

Recently, the SVG graphical format has been utilized to permit vector graphics to access object-oriented graphical information [1]. We utilize SVG format also for raster graphics enabling the user to obtain the required information in both verbal and non-verbal form.

The *BWG* system is designed to enable the creation of web pages and web presentations by means of dialogue. It enables creating a web site step-by-step, and allows the users to ask questions about the image and its structure within the *GATE* system. We present a brief description of the basic modules of the *GATE* system and an example illustrating how the both systems are integrated.

2 GATE System Overview

There are three main goals of the *GATE* project [2]. First, to develop utilities deployed for easy picture annotation. Second, to provide the blind users with support for exploring ("viewing") pictures. And finally, to develop a system for image generation by means of dialogue, enabling the blind to create computer graphics.

In this paper, we are concentrating on the problems concerned with exploring pictures, omitting the modules related to the image generation. One of the important tasks is

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 133–141, 2009.

to inform the user about the graphical content in a non-visual way. The *GATE* system provides two basic ways of informing the user verbally and by means of sound signal. Two basic tools support this type of communication.

The Recursive Navigation Grid (= RNG) is the orientation backbone of the system. The space is divided into nine identical rectangular sectors analogously to the layout of numerical keys 1-9 on the numerical keyboard. Each sector is subdivided in the same way recursively (see Figure 1). The *RNG* module enables the user to obtain the information about the investigated point or region with demanded precision. It also enables "zooming", i.e. assigning coordinates relatively to the recursive grid. [3].

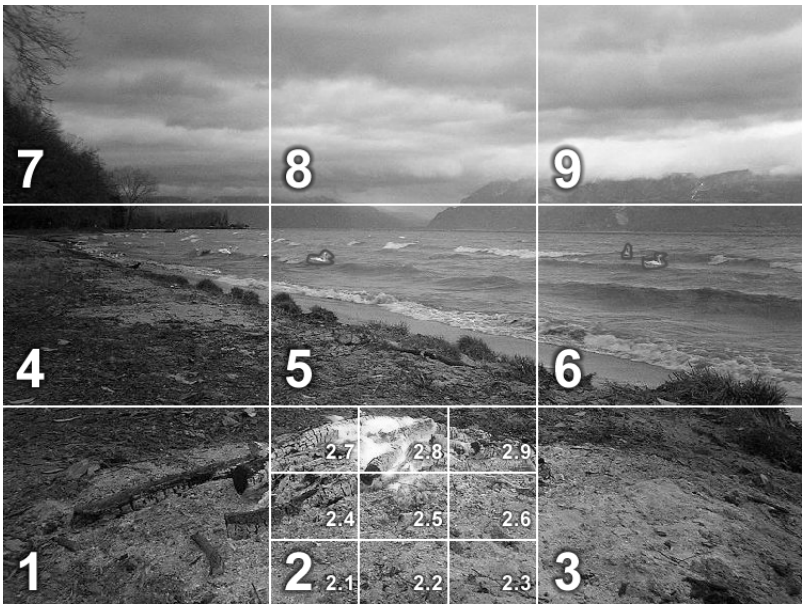


Fig. 1. Example of RNG

What-Where Language (= WWL) is the second basic tool. It is a simple fragment of English consisting of the sentences in the form *WHAT is WHERE* or *WHERE is WHAT*. This language has a simple syntax and to use it, practically no learning is needed. It enables putting simple questions about the objects in the scene and their position, for instance: "What is in the sector 2?", "Where is the swan?", "What is in the background?" etc. In what follows, an example of a conversation between a blind user *U* and the system *S* is presented (U = user, S = system).

U: (Enters the system by clicking a picture or button.)
S: Hi. This is a color photo of Genoa Lake from its shore. For more information, choose Description, Technical information, use keywords or What-Where

language.
U: What is in the foreground?
S: In the foreground, there is an open fireplace, shore and a small part of the lake.
U: What is in the lake?
S: There are two swans in the lake.
U: Where are the swans?
S: Do you wish the approximate localization or use Recursive Navigation Grid?
U: Approximate.
S: One is near the center of the picture; the second one is horizontally near the right-hand side, vertically near the middle.

Verbal Information Module (= VIM) is responsible for the verbal part of the dialogue including *WWL* communication control. Possible misunderstandings in the communication are solved by *VIM* by invoking dialogue repair strategies.

VIM supports two basic investigation strategies represented by two separate modules, called *GUIDE* and *EXPLORER*. Although these modules are independent, they closely cooperate and allow the user to switch between them. The basic function of *GUIDE* is to provide verbal information about the picture, exploiting both the pieces of information obtained by tagging the picture and also the pieces of information gained directly from the picture format, e.g. some technical parameters etc. The module provides relevant information for *EXPLORER* and utilizes the *VIM* and *RNG* modules.

Unlike *GUIDE*, the communication of *EXPLORER* is not primarily verbal, but analogue, by means of mouse, digitizer, or numerical keyboard. The output sound information is also primarily non-verbal. The pieces of information that are related to the pointed place, object or rectangle, are both verbal and non-verbal. Verbal pieces of information are provided from *GUIDE* by an information interface, whereby non-verbal pieces of information are provided directly by *EXPLORER*. Non-verbal communication serves for quick dynamic exploration of the non-annotated details of the picture.

3 BWG System Overview

The main goal of the *BWG* project is to simplify the creation of web presentations to blind users [4]. *BWG* system in consists of online application, which allows the user to create web presentations using dialogue, and a bridge to the *GATE's VIM* module. The process of the creation fulfils the conditions of accessibility standards and does not use a graphical interface. The system also does not require the installation of a special software or knowledge of web technologies, such as HTML or CSS. It utilizes just a web browser and a screen reader as basic equipments of the computers of the blind users.

The user chooses their language and identifies the types of information that appear in the presentation. Then they enter the data into the HTML questionnaire. The presentation is generated after the user selects the graphical layout of the page. The user should require the archive which contains the presentation and which can be uploaded on another server.

Photographs can be processed using the *GATE system*. The photo exploration is available by dialogue communication or image sonification [5]. An example of a generated web presentation is shown in Figure 2.

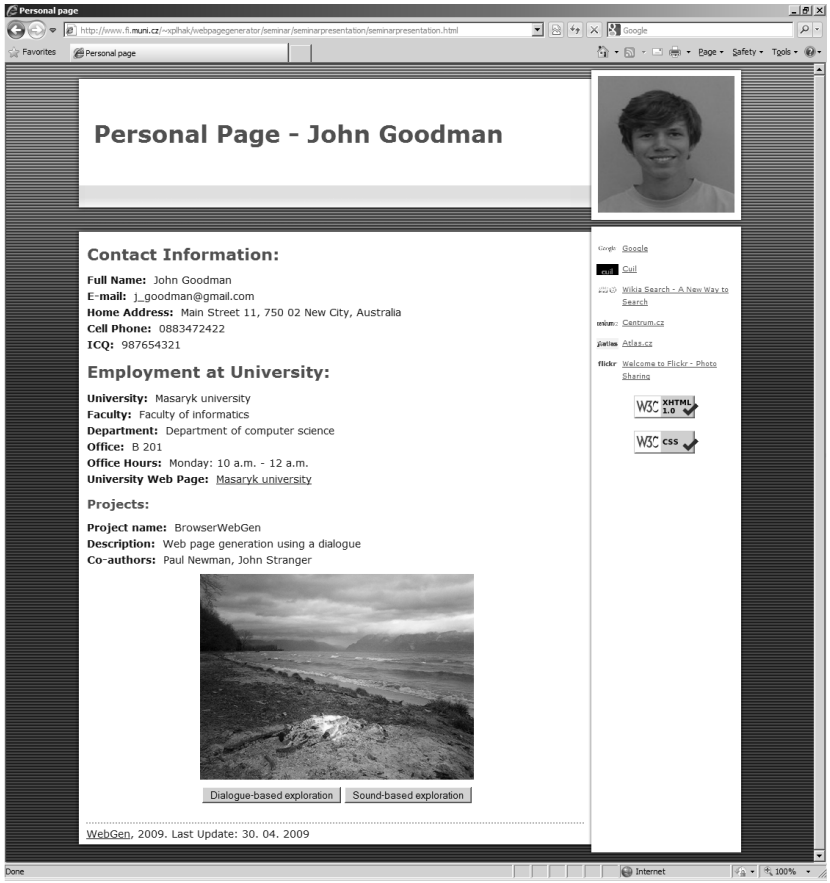


Fig. 2. Generated web presentation with "accessible" image

The **dialogue communication** within the *BWG* system is supported by the *DialogueStep* procedure [6]. This routine enables an easy programming and managing the dialogue strategies. It is based on some principles derived from the *VoiceXML* standard [7].

Entering an annotated picture, the user is made familiar with short basic information about the picture, including its graphical type (photography, painting, drawing, schema, diagram, clipart, mathematical picture, fractal, etc.), color type (full color, black and white, etc.) and a brief description. Then, they can choose the information describing the picture in more detail, technical information or comments about the picture.

The tree structure, into which the data describing the picture are organized, is forming an information backbone for the user's orientation and corresponds to the aggregate-component relationships of the ontology. This approach is complemented by the possibility to ask the system directly using the *WWL*. User question is send asynchronously to the server, where it is processed by the *GATE*. The result is sent back

to the browser, as Figure 3 shows. In this way, the web page with dialog conversation does not have to be reloaded. This approach is less time consuming, and current screen readers can handle *AJAX* (= Asynchronous JavaScript and XML) correctly.

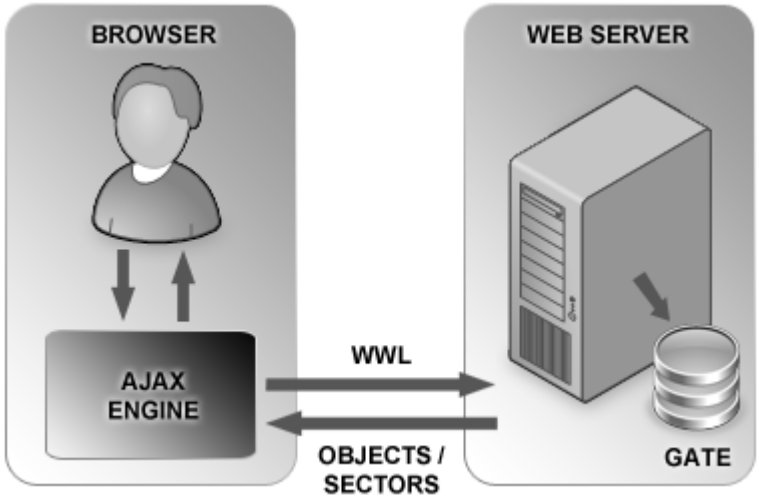


Fig. 3. Principle of querying the *GATE* within the *BWG* by *WWL*

The **image sonification** procedure provides a transformation of colors into the mixture of sounds. It is based on the sound representation of colors, assuming the sound information to be a combination of special sounds assigned to the primary colors of a suitable color model. To enhance the efficiency of this approach, a special graphical color model (see e.g. [8]) is used.

Manipulation with color components, e.g. red, green and blue in the RGB model, is uncomfortable for the investigation. For example, if the system informs the user about the color composed of the 45 % of red, 0 % of green and 55 % of blue, then the user has to realize that this is a shade of violet. Solution is to describe the color only by two primary colors and the luminosity.

We chose seven primary colors: red, green, blue, yellow, orange, violet and brown. Black and white is handled specially, by changing the intensity of the luminosity. Any color from the color spectra is expressed as the mixture of the two closest primary colors, e.g. 50 % of red and 50 % of green. Two sounds correspond to the two primary colors and the volume of each sound means the proportional representation of the color in the composition. Third sound is information about luminosity, where slower frequency of knocking means darker shade and faster knocking means lighter shade.

The sonification sounds has been selected in a way that any combination of the two sounds are distinguishable. We have chosen these sounds to evoke relevant primary color. Border crossing of the image or the grid region is signalized by a special sound.

4 Development Tools

An extension of Adobe Flash that has the ability of getting pixel color under mouse pointer has been chosen to support the image sonification technology. Adobe Flex [9] is a software development kit released by Adobe Systems for the development and deployment of cross-platform rich Internet applications based on the Adobe Flash platform. Flex applications can be written using Adobe Flex Builder or by using the freely available Flex compiler from Adobe. They can run in all modern browsers with installed Flash Player version 9 and above (today it is already over 90 % of all computers). This kit uses a combination of a markup-based language called *MXML* [10], and an object oriented language called *ActionScript* [11] that is based on the *ECMAScript* standard [12]. We exploited the easy work with events and a full fledged code debugger.

Unlike page-based HTML applications, the Flex applications provide a stateful client where significant changes to the view do not require loading a new page. Similarly, Flex and Flash Player provide many useful ways to send and load data to and from server-side components without requiring the client to reload the view. Though this functionality offered advantages over HTML and JavaScript development in the past, the increased support for *AJAX* in major browsers has made asynchronous data loading a common practice in HTML-based development as well.

5 Conclusions and Future Work

The presented framework enables us to generate web pages by mean of dialogue and work with both raster and vector graphical formats. The user has a wide variety of ways to investigate the images. A current version of *BWG* allows the creation of personal web sites.

The next version of the *BWG* system will include the editing, updating individual pages and also creating other types of sites, such as blog and photo gallery. An important task is also testing both systems.

The *GATE* and *BWG* systems are being developed in collaboration with Support Centre for Students with Special Needs of Masaryk University. The Center participates in testing and providing feedback.

Acknowledgment

The authors are grateful to the students and staff of the Support Centre for Students with Special Needs of Masaryk University for their collaboration. This work has been supported by the Czech Science Foundation under Contract No. GA 201/07/0881.

References

1. Mathis, R. M.: Constraint Scalable Vector Graphics, Accessibility and the Semantic Web. In *SoutheastCon Proceedings*, IEEE Computer Society, 2005, pp. 588-593.

2. Kopeček, I. and Ošlejšek, R.: GATE to Accessibility of Computer Graphics. In *Computers Helping People with Special Needs: 11th International Conference*. Berlin, Springer-Verlag, pp. 295–302, 2008.
3. Kopeček, I. and Ošlejšek, R.: Accessibility of Graphics and E-learning. In *The second International Conference on Information and Communication Technology & Accessibility*, Hammamet, Tunisia, 2009.
4. Bártek, L., Plhák, J.: Visually Impaired Users Create Web Pages. In *Computers Helping People with Special Needs: 11th International Conference ICCHP 2008*. Linz, Austria, Berlin: Springer-Verlag, pp. 466–473, 2008.
5. Daunys, G. and Lauruska, V.: Maps Sonification System Using Digitiser for Visually Impaired Children. In *ICCHP 2006*, Berlin, Springer-Verlag, pp. 12-15.
6. Kopeček, I., Ošlejšek, R., Plhák, J. and Tiršel, F.: Detection and Annotation of Graphical Objects in Raster Images within the GATE Project. *Brno: FI MU Report Series*, 2008. FIMU-RS-2008-11.
7. Voice Extensible Markup Language (VoiceXML) Version 2.0, available at `http://www.w3.org/TR/voicexml20`.
8. Kopeček, I., Ošlejšek, R.: Hybrid Approach to Sonification of Color Images. In *KO-The 2008 International Conference on Convergence and Hybrid Information Technologies*. Los Alamitos: IEEE Computer Society, pp. 722–727, 2008.
9. Open source framework Adobe Flex 3, available at `http://www.adobe.com/products/flex/`.
10. MXML 2009 Functional and Design Specification, available at `http://opensource.adobe.com/wiki/display/flexsdk/MXML+2009`.
11. ActionScript 3.0 overview, available at `http://www.adobe.com/devnet/actionscript/articles/actionscript3_overview.html`.
12. ECMAScript Language Specification, 3rd edition (December 1999), available at `http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-262.pdf`.

Czech Vulgarisms in Text Corpora

Vojtěch Kovář, Miloš Jakubíček, and Jan Bušta

NLP Centre, Faculty of Informatics, Masaryk University,
Botanická 68a, 602 00 Brno, Czech Republic

Abstract. This paper focuses on the occurence of vulgarisms in common Czech texts. We present frequencies of vulgarisms found in the Czech National corpus and a web corpus of vulgarisms that has been newly created. Based on these data, usage of vulgarisms in then further discussed.

1 Introduction

Vulgarisms represent an integral part of every language. Since the world doesn't consist of only nice, pretty, sweet and mild entities, we need to label, so to speak, all the disgusting, nasty and ugly aspects of it. Thus we need the vulgarisms. They play an important role in the language, including social and moral aspects. They help people express strong negative feelings to somebody or something.

But, how do we use them, at all? How often are they used in written texts? Which are the most frequent ones?

Text corpora are very valuable source of linguistic information. They can be used for studying almost all language phenomena, including vulgarisms.
In this paper, we present a small research of Czech vulgarisms, based on their frequencies in text corpora. We discuss the frequencies of the vulgarisms in available Czech corpora and also describe creation of a special, „domain" corpus of Czech vulgarisms. Based on these data, we formulate basic properties of Czech vulgarisms in written texts.

2 Available Data

In this section, we describe available data that are used in the next parts of the paper.

2.1 The List of Vulgarisms

It is not clear which words can be treated as vulgarisms – some words are vulgarisms in certain rare contexts only, other ones can be considered as vulgarisms in majority of contexts.

Since currently it is impossible to decide automatically which occurence of the certain word is vulgar and which is not, we used the public list[1] available as a part of the Wiktionary project. It contains 74 words that are usually vulgarisms in majority of contexts and therefore it is not necessary to handle particular occurences of these words. Although the list is not complete, it contains the most important vulgarisms and it is large enough to give us a basic idea of the behaviour of Czech vulgarisms.

[1] http://cs.wiktionary.org/wiki/Kategorie:cs-Vulgarismy

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 141–147, 2009.

2.2 Text Corpora

The most representative source of the Czech language data is undoubtedly the Czech National Corpus [1]. It contains about 100 million tokens and consists of a large number of texts from a variety of sources.

Another remarkable source is the DESAM [2] corpus created at the Masaryk University in Brno. This corpus is much smaller (about 1 million tokens) and contains manually disambiguated morphological tags. Due to the nature of included texts, this corpus contains minimum vulgarisms (less than 10) and we will not include it in further measurements.

The last corpus we used is a newly created corpus of vulgarisms. This corpus contains 2.2 million tokens and consists of texts downloaded from the web. The procedure of creation such corpus is described in the next section.

3 The Corpus of Vulgarisms

In this section, we briefly describe the procedure of collecting the domain-specific texts (such as vulgar texts). As a basic tool for collecting domain-specific texts, the WebBootCat program [3] was used. The basic texts obtained from WebBootCat were then further filtered and prepared for use in a corpus query system Bonito/Manatee [4].

3.1 The WebBootCat Tool

The WebBootCat is a user-friendly tool for building text corpora from the web. It combines a mechanism of searching for web pages based on the Yahoo! search engine [5] with extracting keywords from existing corpora. It starts with a set of so called *seed words*, collects the web pages retrieved from the search engine using these seed words and after application of a series of filters (encoding detection, boilerplate and html tagging removal etc.), builds a corpus from them. New keywords can be extracted and manually selected from this new corpus and can be used as seed words in the next iteration of the described process.

3.2 Creating the Corpus

As seed words, we used vulgarisms from the Wiktionary list introduced above. Using these seeds words we were able to create a corpus with 2.24 millions of tokens. After keywords extraction, no new Czech vulgarisms were found.

Among other extracted keywords, the word *reagovat* (*reply*) is very interesting which indicates that many vulgar texts in our corpus have their origin in internet discussions.

The corpus was compiled for use in a corpus query system Bonito/Manatee [4] and is available for searching.

Word	Frequency
dobytek	534
svině	424
hovno	317
krám	297
honit	278
prdel	268
kurva	212
děvka	137
hajzl	114
buk	93

Table 1. Top 10 vulgarisms in the Czech National Corpus

Word	Frequency
mrdat	1002
kurva	837
hovno	758
prdel	698
kokot	518
debil	343
píča	199
jebat	172
piča	165
kunda	156

Table 2. Top 10 vulgarisms in the new corpus of vulgarisms

4 Frequencies of Czech Vulgarisms

In this section, we show the results of frequency measurements of vulgarisms contained in the Wictionary list. For frequency computation, a newly implemented word list feature of the Bonito system was used. The frequency counts are based only on word forms, lemmata were not considered in the computation.

4.1 Vulgarism Density

Total number of vulgarisms found is 3,362 for the Czech National Corpus (CNC) and 6,826 for the corpus of vulgarisms. With regard to the sizes of the two corpora, this means that the vulgarisms density in the new corpus (3,047 vulgarisms per million words) is about 90 times higher than in the CNC (34 vulgarisms per million words). However, some of the items found in the CNC might not be vulgar in all contexts, e.g. *dobytek*, *honit* or *buk*, since these words have more senses including non-vulgar ones. Thus, the difference in density might be slightly higher.

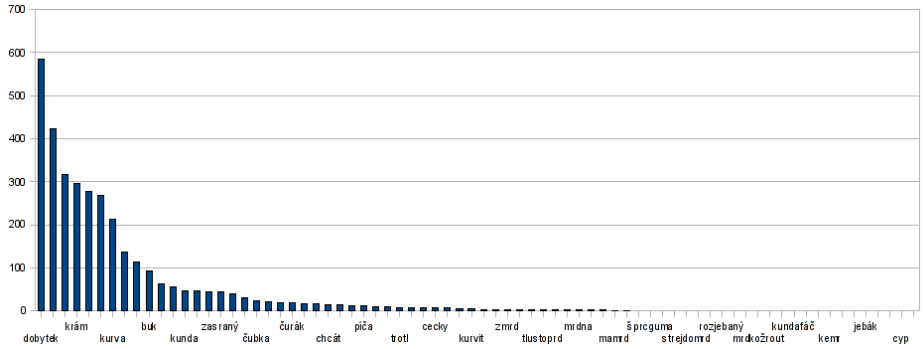


Fig. 1. Frequency distribution of vulgarisms in the Czech National Corpus

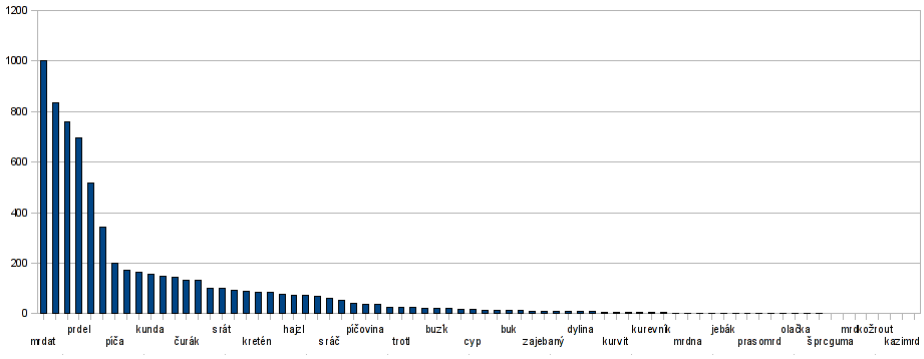


Fig. 2. Frequency distribution of vulgarisms in the corpus of vulgarisms

4.2 Most Frequent Vulgarisms

In Tables 1 and 2, we can see the most frequent vulgar words for both corpora. We can see that the lists differ in most items which indicates that the structure of using vulgarisms in internet texts is different from the general language, as it is captured in the CNC.

Words *kurva*, *hovno* and *prdel* are common for both lists as they probably belong to the most frequently used vulgarisms in both spoken and written language. However, about the spoken language we cannot say more than an expectation since there are no relevant statistics of their usage.

4.3 Frequency Distribution of Vulgarisms

In Figures 1 and 2, we can see the frequency distribution of the vulgarisms under examination. In both cases it can be well approximated by the curve of exponential distribution – the few most frequent vulgarisms form the majority of all occurences and the less frequent ones are not so important from the quantitative point of view.

This behaviour clearly corresponds to Zipf's law which is characteristic for many phenomena in corpus linguistics and statistical language processing [6].

5 Conclusions

In this paper, we described some of the basic properties of Czech vulgarisms in text corpora. We presented building new corpus of vulgarisms and then showed differences between this corpus and a reference corpus of Czech, the Czech National Corpus. We also discussed frequency distributions of the Czech vulgarisms in both corpora and formulated some basic observations about their behaviour. We hope this work could become a good starting point for future studies on vulgarisms in the Czech corpora.

References

1. Ústav Českého národního korpusu FF UK: Český národní korpus – SYN2000 (2000) `http://www.korpus.cz`.
2. Pala, K., Rychlý, P., Smrž, P.: DESAM — annotated corpus for Czech. In: Proceedings of SOFSEM'97, Springer-Verlag (1997) 523–530 Lecture Notes in Computer Science 1338.
3. Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P.: WebBootCat: a web tool for instant corpora. In: Proceeding of the EuraLex Conference 2006, Italy, Edizioni dell'Orso s.r.l. (2006) 123–132
4. Rychlý, P., Smrž, P.: Manatee, Bonito and Word Sketches for Czech. In: Proceedings of the Second International Conference on Corpus Linguisitcs, Saint-Petersburg, Saint-Petersburg State University Press (2004) 124–132
5. Filo, D., Yang, J.: Yahoo! Inc. (2007)
6. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)

K formování Českého akademického korpusu

Jan Králík

Ústav pro jazyk český AV ČR

Před 40 lety – aniž kdo tušil všechna budoucí rozcestí – začal na půdě Ústavu pro jazyk český ČSAV vznikat pro účely lingvistických studií první pražský projekt rozsáhlého souboru textových dat v elektronické podobě. Postupně formované obrysy projektu pak procházely od roku 1971 pravidelnými semináři, na kterých nechyběl ani nynější jubilant, tehdy ale i dnes stále duchem mladý a dynamický Karel Pala. Z okruhu brněnských lingvistů přinášel řadu otázek, svěží nápady řešení a nadšené odhodlání vstupovat do jakýchkoli problémů počítačového zpracování českých textů s odvahou a vynalézavostí. Spolu s ostatními v diskusích nemálo přispěl k nalézání ideální podoby projektu, jehož výsledkem byl – v pozdější terminologii – Český akademický korpus, byť se na jeho vzniku přímo nepodílel. Otázkami a vlastními kroky při jiných projektech ale pomáhal tříbit cesty, jejichž výsledek je stále nepominutelný. Z živých diskusí jistě čerpal i sám (Pala 1996 aj.). Protože historické utváření zůstalo zasuto, jubileum Karla Paly nabízí ojedinělou příležitost k ohlédnutí alespoň za některými stránkami formování Českého akademického korpusu.

Rozvoj počítačové techniky přinesl v závěru 60. let 20. století rychlejší a produktivnější možnosti také analýze textů přirozeného jazyka, zejména sběru a zpracování kvantitativních charakteristik, které náležely k tradicím pražské lingvistické školy a byly tedy připraveny teoreticky i v prvních krocích prakticky zachyceny. Aplikace počítačové techniky a tím vynucená týmová práce ovšem vyžadovaly sjednocení východisek a ostré definice jevů i prvků. Přirozený vliv tradičně rozmanitých lingvistických přístupů přitom ještě neměl alternativu a vylučoval ryze technické pohledy. Mnohahodinové diskuse na seminářích proto neformovaly pouze projekt sám, zejména výběr textů a míru jejich popisu pro získání kvantitativních charakteristik, ale formovaly i jednotlivosti, dnes vyřešené, a proto zdánlivě marginální. Hodiny se například opakovaně diskutovalo o tom, jak v projektu chápat „slovo“ – ve škále od ryze technického řetězce písmen mezi dvěma mezerami po termín zastřešující celé lemmatické hnízdo. Podobně živě diskutované byly i postupy lemmatizace (od sjednocování do typů všechen/všecken, brambor/brambora, zvedat/zdvihat po různé pohledy na reflexiva a tehdy módní sdružená pojmenování a nevlastní předložky). Probírala se celá řada dalších (menších i klíčových) problémů.

Nejasné zůstávaly samozřejmě také mnohé technické otázky – nejen ty publikované (Králík 1987) – například, zda připravená data zůstanou déle uchována do budoucna. Počítačové systémy se rychle a nekompatibilně měnily. V počátku projektu byla k dispozici tzv. III. generace počítačů (aktuální zhruba v letech 1964-1981), schopná pracovat se 100 000 až 5 miliony operací za sekundu při operační paměti 0,5 až 2,0 MB. Jakkoli to dnes vzbuzuje úsměv, šlo o velký pokrok umožněný dobovými novinkami: integrovanými obvody, polovodičovými paměťmi a magnetickými disky, tedy skokovým vývojem proti reléovým či elektronkovým začátkům v nulté a první generaci i proti tranzistorům, ferritovým operačním pamětím a zavedení magnetických pásek v generaci druhé. Stále šlo ještě o počítače se zvláštním režimem přípravy dat na děrných páskách

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 147–155, 2009.

nebo štítcích, o dlouhé nestandardizovatelné programování a ladění každého úkonu od vstupu a ukládání dat přes opravy a třídění po jakékoli kombinační výpisy výsledků. Například abecední třídění celého korpusu trvalo šest hodin a provádělo se na několik pokusů v noci. Váha výpisu na papíře byla v řádu kilogramů. Výstupy na displeji byly ještě neznámé.

Projekt „všestranné kvantitativní analýzy současné psané a mluvené spisovné češtiny" prosadila M. Těšitelová s kolektivem oddělení matematické lingvistiky Ústavu pro jazyk český ČSAV. Původním záměrem bylo navázat novým textovým materiálem a jeho počítačovým zpracováním na materiál ručně zpracovaného Frekvenčního slovníku slov, slovních druhů a tvarů v českém jazyce (Jelínek – Bečka – Těšitelová 1961). Projekt měl proto obsáhnout i krásnou literaturu a svými výsledky ilustrovat paralelně připravovanou akademickou Mluvnici češtiny 1-3 (1986-1987). Diskuse ke koncepci mluvnice se však začaly rozrůstat do míry, jakou už nebylo možno vtěsnat do technických limitů korpusového projektu. Tradičnímu lingvistickému myšlení se zdála spolupráce s jakoukoli technikou včetně počítačové omezením na příliš mechanické pohledy. Ačkoli pražská škola uměla pracovat s distinktivními rysy a s funkčními hledisky, sjednocování pohledů bylo teprve ve stadiu procesu, jehož konvergence se dala jen matně tušit. Korpusový projekt se proto osamostatnil, přidržel se systematické morfologie a syntaxe Vl. Šmilauera (Šmilauer 1972) a pro své účely generoval vlastní pohled na zachycování syntaxe. Morfologické kategorie byly zachyceny podle rozvrhu M. Těšitelové (Těšitelová 1984, 1985b) a M. Ludvíkové (Ludvíková 1983-1990), zásady a rozvrh pro syntax vypracovaly L. Uhlířová (Uhlířová 1983-1990) a I. Nebeská (Nebeská 1983-1990).

Morfologické kategorie se vázaly na slovní druh. U jmen se zachycovaly rod, číslo a pád, případně bližší určení slovního druhu (u adjektiv a číslovek), přítomnost předložkové valence, u adjektiv stupeň, u zájmen kratší či delší tvar, u číslovek vedle druhu i pád počítaného předmětu atd. U sloves byly plně popisovány osoba (číslo), slovesný čas, způsob, slovesný i jmenný rod, příp. imperativ, neurčité tvary, složenost (víceslovnost) atd. Podrobnější popis zůstal i u nesklonných druhů slov: adverbia byla rozlišována podle původu, případně podle stupně, u předložek vlastních i nevlastních byl naznačen pád přítomný v užité vazbě, u spojek se rozlišovala souřadicí nebo podřadicí úloha.

U morfologických kategorií, jejichž určení bylo obligatorní, se ale už na sondách ukázalo, že je třeba také zavést technickou variantu „nelze určit", a to nejen pro texty mluvené. Vzhledem k záměru obsáhnout v korpusu současné spisovné češtiny i mluvené texty se ukázalo jako nutné zavést také značení pro nespisovnost (v 70. letech pojímanou přísněji) a výplňkovost, případně naopak pro „zastaralost" (= knižnost).

Syntax byla popsána ve dvou úrovních jednak pro postavení jednotlivých slov ve větě, jednak pro větu a větný celek. Označovala se syntaktická platnost slova ve větě (subjekt, predikát, atribut, apozice, doplněk, adverbiále, základ věty jednočlenné, přechodný typ se všeobecným subjektem, samostatný větný člen, parenteze) a její případné bližší určení (typ predikátu: slovesný, spona, nominální část sponového predikátu, spona u jednočlenné věty atd.), adverbiále místa, času, způsobu, příčiny, původu, původce a výsledku, základ věty jednočlenné substantivní, adjektivní, citoslovečné, částicové, vokativní, příslovečné, infinitivní, slovesné, slovesně jmenné a zájmenné atd.

U slov závislých (nikoli řídících) se udávala poloha (pozice) a vzdálenost od řídícího slova a zachycovala se i informace o případné koordinaci členů koordinační řady, o případném sdruženém pojmenování atd. Koordinační řady a sdružená pojmenování

byly kódovány tak, aby je bylo možno plně rekonstruovat i v případech koordinace uvnitř sdruženého pojmenování, u spojkových a příslovečných dvojic apod.

Popis syntaxe byl zároveň z hlediska kódů koncipován tak, aby umožňoval jednoznačný lineární zápis, ze kterého by bylo možno kdykoli zkonstruovat graf. Stejnou podmínku splňovalo i zakódování charakteristik vět, soustředěných vždy u prvního slova věty nebo větného celku. Rozlišovaly se věty jednoduché, hlavní či vedlejší (subjektové, predikátové, atributivní, objektové, místní, časové, způsobové, příčinné a doplňkové). U vedlejších vět atributivních se zachovávala informace o poloze (vzdálenosti) řídícího jména. Doplňovala se také informace o vztazích mezi větami uvnitř souvětí (koordinace, parenteze, přímá řeč, parenteze v přímé řeči, uvozovací věta, parenteze v uvozovací větě).

Také v popisu syntaxe bylo třeba pamatovat na specifiku mluvených textů, tedy např. na nepřítomnost řídících výrazů, na nepravé věty vztažné nebo obecně na chyby ve stavbě souvětí.

Všechny tyto údaje a pohledy stále znovu sjednocované v poradách pracovního týmu byly ručně převedeny do kódů, revidovány, zapisovány na děrné štítky, snímány, archivovány a tištěny a znovu revidovány (Králík 1987). Intelektuálně i technicky šlo o mimořádně náročnou přípravu, při níž vyvstávaly samozřejmé otázky po možném automatizování – automatické lemmatizaci a značkování. Ze strany lingvistů se stále ještě mísila nedůvěra k automatizaci s lákavostí ulehčení práce. Důvěrná znalost proměnlivosti a vnitřní variability jazykového systému neumožňovala věřit ryze technickým postupům při přípravě a zpracování – dnes bychom řekli značkování – textů. Váhání dodnes trvá, ale tolerance, priority a vývoj se časem posunuly díky postupům navrženým na jiných pracovištích zejména J. Hajičem (Hajič – Hladká 1997ab, Hajič et al. 2001, Hajič 2004 aj.) a B. Hladkou (Hladká – Ribarov 1998, Hladká 2000, Vidová-Hladká 1994, 2000 aj.). Ale také stopa jubilujícího K. Paly náleží v tomto směru mezi nejvýraznější (Pala 1996, Pala – Osolsobě – Rychlý 1998, Pala – Sedláček – Veber 2004, Klímová – Oliva – Pala 2005 aj.). Hesitační fázi nebylo možno obejít. Časem ji však bylo možno dovést na kvalitativně vyšší úroveň.

První automatizování v pražském projektu mělo zvláštní negační povahu: díky úplnosti zápisu syntaxe bylo rozhodnuto rezignovat na zápis interpunkce. Z kódů byl plně rekonstruovatelný, pokud ovšem původní text neobsahoval chyby. Předpoklad vycházel z přísného kritéria spisovnosti vybraných textů. Automaticky se zkoušely doplňovat slovníkové podoby (lemmata) u nesklonných slovních druhů (u nich nebyla žádná diskuse), ale bylo možno i nabízet tvary lemmat tam, kde morfologický kód naznačoval identickou podobu s tvarem v textu (nejen u nominativů singuláru a infinitivů). Další návrhy byly tehdy – v počátcích projektu – vyslechnuty a s krajní nedůvěrou striktně odmítnuty. Zkoušely se tedy alespoň automatické kontroly shod v pádu přívlastku a rozvíjeného substantiva (následujícího i předcházejícího). Po určité zkušenosti přišel pokus automaticky kontrolovat přípustnost kombinací v kódu. Našla se tak řada lidských chyb, ale objevilo se i několik případů, na jejichž realitu do té doby nikdo nepřišel. Výčet by mohl pokračovat.

Výběr textů pro celý projekt vznikal za pochodu tak, aby zachytil jazyk právě aktuální, tedy z počátku 70. let 20. století, a zároveň aby naplnil rozhodnutí o procentuální struktuře různých stylů. V rámci tzv. věcného stylu (540 000 slov) se předpokládala jedna třetina textů publicistických, odborné texty v rozsahu 300 000 slov a zbylých 60 000 slov z textů administrativních. Hlavní důvod, proč se tehdy začínalo u věcného stylu, je zapomenut:

nauka a administrativa byly politicky neutrální, u publicistiky bylo možno s jistotou doby začít výběrem z Rudého práva. Původně zamýšlené doplnění beletrií na celek o rozsahu jednoho milionu slov se neuskutečnilo. U beletrie tehdy nebylo vůbec jisté, zda vybraný autor či dílo zůstanou mezi přijatelnými. Klasikové byli již zpracováni a mizení soudobých autorů v politickém propadlišti by s sebou mohlo strhnout i celý projekt (Uhlířová – Králík 2007).

Jako „texty" byly do korpusu zařazovány souvislé výběry o rozsahu 3000 slov z delších celků. Důvod byl statistický, ale opět i politický. Sondy ukázaly, že k relevantnímu zachycení základních kvantitativních charakteristik většiny gramatických kategorií postačí souvislý text v délce 2000 slov. Právě toto číslo bylo ale od roku 1969 politické tabu. Číslo 3000 nepřipomínalo nic a statisticky poskytovalo víc než potřebný výběrový komfort (Hladká – Králík 2006). Členění na výběry umožnilo posílit rozmanitost oborového záběru. Celek pak mohl být právem označen za korpus věcného stylu. Výběrem textů se zabýval Jiří Kraus.

V každé ze zvolených oblastí bylo jako záměrné novum plných 25 % mluvených textů, což přineslo v té době nemálo starostí, zejména u sběru „mluvené" administrativy. Zaměření na spisovný jazyk bylo menším omezením než požadavek pestrosti zdrojů. Nakonec se i zde podařilo překvapivé vyvážení, jak později doložilo speciální studium kvantitativních výsledků např. na slovnědruhové struktuře mluvené publicistiky (Králík 1991).

Práce na formování korpusu tím nabývala na objevnosti, dobrodružnosti i vnitřním napětí. Ačkoli příprava se zdála jako ryze mechanická, přinášela zajímavé detaily v poznání a zvyšovala touhu vidět první výsledky. Ty není třeba opakovat ani rozvádět, protože byly podrobně publikovány v celé ediční řadě (Těšitelová ed. 1980-1992).

Postup prací a systematické zveřejňování výsledků vzbudily nemalý zájem nejen v okruhu lingvistů, z nichž Karel Pala náležel k těm nejbližším. Jako jedna z prvních institucí mimo akademickou sféru projevila zájem o textový korpus věcného stylu agentura ČTK. Středem zájmu byla pochopitelně publicistická část (180 000 slov), vhodná pro programování zkušebních automatických vyhledávačů. V akademických kruzích pak vzbudily publikace a přednášky o prvních výsledcích živý ohlas také v Německu, Rakousku, Finsku a Španělsku. Na domácí půdě došel nakonec Český akademický korpus – díky zdařilé záchraně všech dat – logického uplatnění inkorporací do několika projektů, z nichž nejaktuálnější je konverze pro rámec Pražského závislostního korpusu (Hladká – Ribarov 1998, Hana – Zeman 2005, Hladká – Králík 2006 aj.).

Karel Pala nabídl brzy k některým postupům na svém pracovišti v Brně vlastní řešení v nových, širších paralelách a podílel se na rozpracování vlastních cest, již poučených na zmapovaných slepých uličkách, ale i na vyzkoušených jistotách. Zkušenosti z různých polí se pak ještě jednou sešly v roce 1991 při zakládání Skupiny pro počítačový fond češtiny (Čermák – Králík – Pala 1992). Ale to už by byla zase jiná vzpomínka.

Reference

1. Čermák, F. – Králík, J. – Pala, K. (1992): Počítačová lexikografie a čeština (počítačový fond češtiny). *Slovo a slovesnost 53*, s. 41-48.
2. Hajič, J. (2004): *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Prague: Karolinum, Charles University Press, Czech Republic.

3. Hajič, J. – Hajičová, E. – Pajas, P. – Panevová, J. – Sgall, P. – Vidová-Hladká, B. (2001): *Prague Dependency Treebank 1.0* CDROM. Praha: Linguistic Data Consortium.
4. Hajič, J. – Vidová-Hladká, B. (1997a): Probabilistic and Rule-Based Tagger of an Inflective Language - a Comparison. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington DC, USA, s. 111-118.
5. Hajič, J. – Vidová-Hladká, B. (1997b): Morfologické značkování korpusu českých textů stochastickou metodou. In: *Slovo a slovesnost*, 58, (4), s. 288-304.
6. Hana, J. – Zeman, D. (2005): *Manual for Morphological Annotation, Revision for the Prague Dependency Treebank 2.0*. Praha: Ústav formální a aplikované lingvistiky MFF UK.
7. Hladká, B. (2000): *Czech Language Tagging.* PhD thesis. Prague: Charles University.
8. Hladká, B. – Králík, J. (2006): Proměna Českého akademického korpusu. *Slovo a slovesnost* 67, č. 3, s. 179-194.
9. Hladká, B. – Ribarov, K. (1998): PoS Tags for Automatic Tagging and Syntactic Structures. In: E. Hajičová (ed.): *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, Karolinum, Charles University Press, Prague, Czech Republic, s. 226-240.
10. Jelínek, J. – Bečka, J. V. – Těšitelová, M. (1961): *Frekvence slov, slovních druhů a tvarů v českém jazyce* FSSDTČJ, SPN Praha.
11. Klímová, J. – Oliva, K. - Pala, K. (2005): Czech lexical database - First stage. In: F. Kiefer, G. Kiss, J. Pajzs (Eds.): *Papers in Computational Lexicography, Complex'05*. Budapest: Hungarian Academy of Sciences, s. 142-151.
12. Králík, J. (1983a): Some Notes on the Frequency – Rank Relation, *Prague Studies in Mathematical Linguistics 8.* Praha: Academia, s. 67-80.
13. Králík, J. (1983b): Statistika českých grafémů s využitím moderní výpočetní techniky. *Slovo a slovesnost 46*, s. 295-304.
14. Králík, J. (1987): *Kapitoly o výpočetní technice.* K problémům komunikace lingvista – programátor – počítač. ÚJČ ČSAV Praha.
15. Králík, J. (1991): Probabilistic scaling of texts. In: R. Köhler – B. R. Rieger (eds.): *Contributions to Quantitative Linguistics*, Kluwer Academic Publisher Dordrecht – Boston – London, s. 227-240.
16. Králík, J. (2007): Quantitative Linguistics within Czech Contexts. In: R. Köhler, P. Grzybek (eds.): *Exact Methods in the Study of Language and Text.* Mouton de Gruyter Berlin – New York, QL 62, s. 343-351.
17. Ludvíková, M. (1983): Quantitative Aspects of Verb Categories (Based on Present-Day Czech Non-Fiction Texts). *Prague Studies in Mathematical Linguistics 8*, Praha: Academia, s. 19-30.
18. Ludvíková, M. (1986): On the Semantics of Pronominal Adverbs from the Quantitatice Aspect. *Prague Studies in Mathematical Linguistics 9*, s. 43-52.
19. Ludvíková, M. (1990): Some Specific Features of the Semantics of Adverbs. *Prague Studies in Mathematical Linguistics 10*, Praha: Academia, s. 49-64.
20. *Mluvnice češtiny 1 – 3* (1986-1987), Academia Praha, 566 s, 536 s, 738 s.
21. Nebeská, I. (1983): Compound / Complex Sentences in Non Fiction Texts. *Prague Studies in Mathematical Linguistics 8*, Praha: Academia, s. 53-66.
22. Nebeská, I. (1986): A Contribution to the Semantics of Modal Verbs from the Quantitative Point of View. *Prague Studies in Mathematical Linguistics 9*, Praha: Academia, s. 31-42.
23. Nebeská, I. (1990): On Expressing Possibility and Necessity in Czech. *Prague Studies in Mathematical Linguistics 10*, Praha: Academia, s. 75-84.
24. Pala, K. (1996): Informační technologie a korpusová lingvistika (1) a (2). *Zpravodaj ÚVT MU VI*, č. 3 a 4, *s. 11-13.*
25. Pala, K. - Osolsobě, K. - Rychlý, P. (1998): Frekvence vzorů českých substantiv. *Sborník prací Filosofické fakulty brněnské university*. A, Řada jazykovědná, Brno: Filosofická fakulta brněnské university, s. 77-92.

26. Pala, K. – Sedláček, R. – Veber, M. (2004): Vztah mezi tvarotvornými a slovotvornými vzory v češtině. *Čeština – univerzália a specifika 5*, 2004, s. 151-162.
27. Šmilauer, Vl. (1972): *Nauka o českém jazyku*, Praha 1972.
28. Těšitelová, M. (1979): On Quantitative Linguistics in Czechoslovakia, ITL, *Tijdschrift van het Institut Toegepaste Lingüistiek 43*, Leuven, s. 53-73.
29. Těšitelová, M. ed. (1980a): *Frekvenční slovník současné administrativy*, ÚJČ ČSAV, Praha.
30. Těšitelová, M. ed. (1980b): *Frekvenční slovník současné české publicistiky*, ÚJČ ČSAV, Praha.
31. Těšitelová, M. (1981): On the Language of the Present-Day Publicist Prose. *Prague Studies in Mathematical Linguistics 7*, Praha: Academia, s. 9-26.
32. Těšitelová, M. ed. (1982a): Kvantitativní charakteristiky současné české publicistiky, *Linguistica II*, ÚJČ ČSAV, Praha.
33. Těšitelová, M. ed. (1982b): Kvantitativní charakteristiky současné české publicistiky (tabulky a grafy), *Linguistica III*, ÚJČ ČSAV, Praha.
34. Těšitelová, M. ed. (1982c): *Frekvenční slovník současné odborné češtiny*, ÚJČ ČSAV, Praha.
35. Těšitelová, M. (1983a): On the State of Quantitative Linguistics in Studies of Czech. *The Prague Bulletin of Mathematical linguistics 40*, s. 15-30.
36. Těšitelová, M. (1983b): Some Quantitative Characteristics of Non-Fiction Texts in Present-Day Czech. *Prague Studies in Mathematical Linguistics 8*, Praha: Academia, s. 9-18.
37. Těšitelová, M. ed. (1983a): *Frekvenční slovník češtiny věcného stylu*, ÚJČ ČSAV, Praha.
38. Těšitelová, M. ed. (1983b): *Kvantitativní charakteristiky gramatických jevů v současné administrativě (tabulky)*, ÚJČ ČSAV, Praha.
39. Těšitelová, M. ed. (1983c): Psaná a mluvená odborná čeština z kvantitativního hlediska (v rámci věcného stylu), *Linguistica IV*, ÚJČ ČSAV, Praha.
40. Těšitelová, M. ed. (1983d): Kvantitativní charakteristiky současné odborné češtiny (v rámci věcného stylu), *Linguistica VII*, ÚJČ ČSAV, Praha.
41. Těšitelová, M. (1984): Kvantitativní analýza češtiny s pomocí moderní výpočetní techniky. *Naše řeč 67*, s. 47-50.
42. Těšitelová, M. ed. (1984): *Kvantitativní charakteristiky gramatických jevů v češtině věcného stylu (tabulky a přehledy)*, ÚJČ ČSAV, Praha.
43. Těšitelová, M. (1985): K využití statistických metod v kombinaci s retrográdním uspořádáním jazykových jednotek. *Slovo a slovesnost 46*, s. s. 109-118.
44. Těšitelová, M. ed. (1985a): Současná česká administrativa z hlediska kvantitativního, *Linguistica XV*, ÚJČ ČSAV, Praha.
45. Těšitelová, M. ed. (1985b): *Kvantitativní charakteristiky současné češtiny*. Praha: Academia.
46. Těšitelová, M. (1986): On Semanic Quantitative Analysis. *Prague Studies in Mathematical Linguistics 9*, Praha: Academia, s. 9-18.
47. Těšitelová, M. (1987): Kvantitativní lingvistika a počítače. In: *Kvantitativní lingvistika*, Praha: SPN, s. 140-143.
48. Těšitelová, M. (1990): On Semantics of Nouns from the Quantitative Point of View. *Prague Studies in Mathematical Linguistics 10*, Praha: Academia, s. 9-24.
49. Těšitelová, M. (1992): Quantitative Linguistics. In: *Linguistic and Literary Studies in Eastern Europe 37*. John Benjamins Publishing Company.
50. Těšitelová, M. – Petr, J. – Králík, J. (1985): *Retrográdní slovník tvarů adjektiv v současné češtině*. Praha: ÚJČ ČSAV.
51. Těšitelová, M. – Petr, J. – Králík, J. (1986a): *Retrográdní slovník současné češtiny*, Praha: Academia.
52. Těšitelová, M. – Petr, J. – Králík, J. (1986b): On Some Issues of the Reverse Dictionary of Words and Forms. *Prague Studies in Mathematical Linguistics 9*, Praha: Academia, s. 65-74.
53. Těšitelová, M. – Uhlířová, L. – Králík, J. (1984): K automatickému zpracování textu při kvantitativní analýze přirozeného (českého) jazyka. *Slovo a slovesnost 45*, s. 145-150.

54. Uhlířová, L. (1983): Simple Sentence Structure from the Auantitative Point of view (Based on Present-Day Czech Non-Fiction Texts). *Prague Studies in Mathematical Linguistics 8*, Praha: Academia, s. 43-52.
55. Uhlířová, L. (1986): On Verbal Semantics from the Quantitative Point of view. *Prague Studies in Mathematical Linguistics 9*, Praha: Academia, s. 19-30.
56. Uhlířová, L. (1990): Beginning and the End of Sentence (A Quantitative Study on the Present-Day Czech). *Prague Studies in Mathematical Linguistics 10*, Praha: Academia, s. 65-74.
57. Uhlířová, L. – Králík, J. (2007): The Czech Academic Corpus (CAC), its history and presence. *Journal of Quantitative Linguistics 14*, Taylor & Francis, s. 265-285.
58. Uhlířová, L. – Nebeská, I. – Králík, J. (1982): Computational Data Analysis for Syntax. In: J. Horecký (ed.): *COLING 82*. Amsterdam - New York - Oxford, s. 391-396.
59. Vidová-Hladká, B. (1994): *Software Tools for Large Czech Corpora Annotation*. MSc thesis, Prague: MFF UK.
60. Vidová-Hladká, B. (2000): *Czech Language Tagging*. PhD thesis. Praha: Univerzita Karlova.

Tajemné spojení jazyka se světem

Pavel Materna

Fakulta informatiky, Masarykova univerzita, Brno

Abstrakt Co může a co nemůže jazyková konvence:
Jde o pojmenování jednotlivin? Proč by v takovém případě byla oprávněná výtka, že jde o 'mýtus muzea'? Jazyková konvence určuje výhradně podmínky identifikace předmětů. Výrazy proto nemohou označovat předměty samé, protože konvence není vševědoucí a nemůže vědět, které předměty splňují dané podmínky za daného stavu světa.
Gramatiku nelze chápat pouze jako pravidla spojování výrazů: jde o zakódování objektivních kvazialgoritmických procedur, které z jednotlivých abstraktních objektů vytvářejí nové objekty.
Toto pojetí dostalo přesnou podobu v Tichého Transparentní intenzionální logice, kde zmíněné procedury byly definovány jako tzv. konstrukce. Vzájemná spolupráce TIL a komputační lingvistiky je proto výhodná pro obě strany.

1 Logická analýza přirozeného jazyka a jazyková konvence

Jak si vysvětlíme pozoruhodný fakt, že sekvence symbolů *Na Marsu jsou živé organismy* říká něco zajímavého mluvčímu češtiny a je pouhou změtí znaků pro běžného Číňana? Jaký vztah má uvedená sekvence k abstraktnímu objektu, který nazýváme 'propozice' a který může být pravdivý nebo nepravdivý?

Je zřejmé, že otázkami, které vznikají v souvislosti s tímto fundamentálním tázáním, se zabývají z různých hledisek různé obory. Jde především o lingvistické obory a zejména o lingvistickou sémantiku, dále o filozofii jazyka, a konečně o logickou analýzu přirozeného jazyka (LANL). Zde se budeme zabývat právě LANL, a proto si předem vyjasníme specifický charakter LANL.

Lingvistické obory zkoumají různé stránky jazyka jakožto „přírodního jevu" a mají z tohoto hlediska podobný charakter jako přírodní vědy, tj. jsou *empirické*. LANL vychází z modelové situace, kdy jako mluvčí daného jazyka tomu jazyku rozumíme, takže jazyková konvence pro nás není předmětem empirického zkoumání, nýbrž je nám dána. Z toho hlediska se LANL liší od jakékoli jiné disciplíny zkoumající jazyk. Jazyk je pro LANL již dán, konvence je přijata. Co zbývá, je *odhalení těch abstraktních logických struktur, které jsou výrazy a gramatikou daného jazyka zakódovány*.[1]

Takto chápaná LANL se tedy nezabývá empirickým výzkumem: její metoda je skutečně logická, tj. *a priori*.

[1] LANL se tedy odlišuje na jedné straně od empirických oborů lingvistiky, na druhé straně od logické sémantiky, která se týká formalizovaných systémů a která *stanovuje interpretaci,* kdežto LANL *hledá a objevuje* významy výrazů. Pokud jde o filozofii jazyka, je LANL neslučitelný s jakoukoli formou pragmatického pojetí významu, jak je najdeme např. u Quinea a pozdního Wittgensteina.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 155–163, 2009.

Zde můžeme narazit na tuto námitku: Jsou známy četné případy, kdy výraz připouští různá 'čtení'. Jak chce LANL rozhodnout, jaké čtení zvolit, když k tomu potřebuje empirické fakty, tj. fakty, které rozhodují o tom, v jaké situaci je adekvátní volit jaké čtení?

Na tuto námitku odpovídá LANL následujícím způsobem: Mějme výraz A, který připouští dvě čtení: $A1$ a $A2$. Vycházíme z toho, jak již bylo řečeno, že rozumíme oběma čtením. Zkoumáme *daný* jazyk, tj. nepotřebujeme empirický výzkum, co výraz A, resp. $A1$, $A2$ znamená. Jde jen o to, dík jakým logickým strukturám vysvětlíme, že vznikla různá čtení.

Poznámka: Názor, že lze provádět desambiguaci na základě 'čisté syntaxe', je iluze.[2] Variantní stromové struktury, k nimž lingvistika dochází na základě různých metod parsing, jsou nutně spojeny s významy jednotlivých výrazů. LANL vidí tyto významy jako abstraktní mimojazykové objekty (viz dále) a umožňuje co možná přesnou aplikaci principu kompozicionality.

Ostatně skutečnost, že rozumíme i čtením $A1$, $A2$, má za následek, že dovedeme výrazu A přiřadit výrazy *A1', A2'* takové, že odpovídají v tomto pořadí čtením $A1$, $A2$.

Jako příklad uveďme výraz

A Karel se chce oženit s princeznou.

Čtením $A1$, $A2$ zřejmě odpovídají výrazy

A1' Existuje princezna, s kterou se chce Karel oženit.

A2' Karel chce, aby existovala princezna, s kterou by se oženil.

LANL pracuje tím způsobem, že přiřazuje jednotlivým podvýrazům daného výrazu abstraktní mimojazykové objekty ('významy') a přesně definovanou syntézou dochází k významu celého výrazu.[3] Protože předpokládá porozumění výrazům, ví, že např. *oženit se* je nějaký vztah mezi dvěma individui, což je jistě mimojazykový objekt, atd. atd.

Hlavní úkol řešený LANL lze shrnout takto: nalézat *významy* výrazů přirozeného jazyka a způsoby, jakými se z významů vytvářejí nové významy. V jistém kontextu jde o rehabilitaci pojmu *význam*, který Quine odsoudil jako „obscure entity" (viz Quine 1953 aj.), mj. protože nepochopil, že nelze definovat význam na základě analytičnosti a synonymie, nýbrž naopak že analytičnost a synonymie jsou definovatelné na základě definice významu.

LANL tedy vychází z plausibilního předpokladu, že jazyková konvence od samého začátku obdařovala výrazy přirozeného jazyka významy. Nečinila tak ovšem způsobem, jaký známe z konstruování umělých (např. formálních) jazyků, nýbrž živelně, takže úkol řešený LANL není triviální. Zamysleme se nyní nad tím, co by významy mohly být.

Začněme nejnaivnější představou, kterou můžeme charakterizovat jako „chybnou interpretaci knihy Genesis", kde Adam je vyzván, aby pojmenoval všechny živočichy. Příslušná pasáž je krásným obrazem vzniku jazyka. Představme si tedy, že Adam tvoří jazyk tak, že *pojmenovává jednotlivé živočichy*, tj. počíná si jako oni lapuťanští pseudoučenci z Gulliverových cest, kteří verbální komunikaci nahrazovali jednotlivými

[2] Viz např. Gamut, L.T.F. (1991): *Logic, Language and Meaning II., Intensional Logic and Logical Grammar.* Chicago University Press.

[3] V detailech, někdy podstatných, se ovšem různé systémy LANL liší, jmenovitě Montaguova a Tichého analýza.

předměty. Tak tedy Adam potkává jednotlivé živočichy a každého 'onálepkuje', opatří 'jménem'. Jak by asi vypadal takový jazyk?

Racionální výklad onoho 'pojmenování živočichů' je samozřejmě jiný. Když Adam potká slona, dá mu 'jméno' odpovídající českému *slon*. Když potká želvu, opatří ji 'jménem' *želva* atd. Důležité je, že když Adam potká jiného slona / želvu, nedá jim jiné jméno. Ty výrazy *slon, želva* atd. nejsou ve skutečnosti jména jednotlivých exemplářů, nýbrž označují *vlastnosti*. Adam – jakož i jakýkoli mluvčí jazyka – nemá a nemůže mít prostředek, jak přidělovat jména jednotlivinám tak, aby šlo skutečně o různá jména pro různé jednotliviny. Adam místo toho vyčleňuje *podmínky*, jaké něco musí splňovat, aby bylo možné o tom mluvit. V případě 'jmen' pro živočichy jde o *vlastnosti*, které jakákoli individua mají či nemají. V jiných případech jde o jiný druh podmínek, např. jde-li o výraz *papež*, jde o podmínku, kterou musí jedinec splňovat, aby bylo možno o něm mluvit jako o jedinci: takové podmínky nazval Church 'individual concepts' a Tichý užíval termín 'individual office' (dnes mluvíme o *individuových rolích.* V ještě abstraktnějším případě mluvíme o *pravdivostních podmínkách'* čili *propozicích*, které musí nabývat hodnoty Pravda / Nepravda podle toho, zda příslušná věta je pravdivá či nepravdivá.

V uvedených případech je splnění podmínek závislé na stavu světa v daném čase. Proto empirické výrazy nemohou označovat objekty, které danou podmínku splňují: jazyková konvence není vševědoucí, takže neví dopředu, které jednotliviny splňují podmínku v daném stavu světa. Nepatří proto do významu slova *slon*, která konkrétní individua jsou slony 21.4.1930, ani nemůže jazyková konvence zahrnovat informaci, kdo je papežem r. 2009. *Empirický výraz nemůže nikdy označovat něco, co je závislé na stavu světa v daném okamžiku.* Naproti tomu vlastnosti, individuové role, propozice apod. nejsou závislé na momentálním stavu světa. Že papežem v r. 2008 byl Jan Pavel II., to je fakt závislý na stavu světa v tomto roce, a proto irelevantní z hlediska významu, ale to, že papež je hlava katolické církve, je nezávislé na tom, kdo je momentálně nositelem toho úřadu, a proto jde jistě o součást významu slova *papež*. Podobně empirická věta *Na Marsu jsou živé organismy* je pravdivá či nepravdivá v závislosti na stavu světa, a proto nemůžeme tvrdit, že označuje pravdivostní hodnotu: nechtějme na jazykové konvenci, aby do významu té věty vkládala empirická, a tedy logicky nahodilá fakta. Naproti tomu příslušnou propozici můžeme chápat jako funkci, která nabývá pravdivostních hodnot v závislosti na stavu světa a je tedy jakožto tato funkce na stavu světa nezávislá.

Zůstaňme u empirických výrazů. Absurdnost pojmenovávání jednotlivin jako stanovení významu výrazů by skutečně mohla být zvýrazněna Quineovskou metaforou „mýtus muzea", podle něhož sémantika přirozeného jazyka, která vychází z přiřazování významů jednotlivým výrazům, degraduje jazyk na nálepkování jednotlivých muzejních exponátů. Quineův holismus ovšem vylévá s vaničkou i dítě: Podle něho nelze sémanticky ohodnocovat jednotlivé výrazy, nýbrž pouze celky typu jazyka nebo vědecké teorie. Avšak spojení výrazů s významem má hodně daleko k primitivnímu (a NB nemožnému) nálepkování jednotlivin. Viděli jsme především, že nemůže jít o nálepkování jednotlivin, a ony podmínky, o kterých jsme mluvili a které mají charakter *intenzí*, tj. funkcí z možných světů a časů, jsou jistě něco, co nepřipouští srovnání s muzejními exponáty. Jsou tedy *intenze* vhodnými kandidáty na roli *významů*?

2 Intenze jako významy?

Pokud chápeme intenze jako *funkce* z možných světů, které nabývají v daném možném světě v daném čase určitou hodnotu, pak nesplňují jeden ze znaků, které musí mít význam. Jsou totiž jednoduché v tom smyslu, že nemají části, tím spíše pak části, které by odpovídaly částem daného výrazu. Jako ilustraci uvažujme větu *Měsíc je menší než Země*. Kdyby význam této věty byla intenze, pak by to byla propozice, tj. funkce, která by možným světům a časům přiřazovala pravdivostní hodnoty. Našemu, 'aktuálnímu'světu by nyní přiřazovala Pravdu, ale protože nejde o matematické či logické tvrzení, není toto přiřazení nutné: jde o nahodilý fakt, takže v jiných možných světech by tato propozice nabývala hodnoty Nepravda a např. v těch světech, kde Měsíc neexistuje, by byla bez jakékoli pravdivostní hodnoty. Jde tedy o pouhé přiřazení hodnot argumentům (tj. světům a časům) a nic v takovém funkčním přiřazení neodpovídá jednotlivým podvýrazům naší věty. Také je zřejmé, že propozice přiřazená větě *Země je větší než Měsíc* by se absolutně ničím nelišila od propozice přiřazené první větě.

Obecně pak přiřazení intenze jakožto významu danému výrazu je přiřazení nestrukturované množiny (neboť funkce jsou množinové objekty) obecně strukturovanému výrazu, takže podstatná informace, kterou se máme dík významu dovědět, je ztracena.

Na nedostatečnost analýzy založené výhradně na intenzích upozornil 1947 Rudolf Carnap ve slavném spise *Meaning and Necessity,* ale obecně známou se stala kritika 'nestrukturovaného významu' až v sedmdesátých letech dík Cresswellovi a jeho (1975) a (1985)[4]: Cresswell se pokusil zachytit význam jako strukturovaný objekt na základě uspořádaných n-tic. Jeho pokus měl strukturovanost výrazu spojit se strukturovaností významu: jednotlivým podvýrazům pak odpovídaly jednotlivé složky příslušné n-tice. Jak ovšem ukázal Pavel Tichý (1994),[5] n-tice jsou sice jakýmsi seznamem významů jednotlivých podvýrazů, ale nemohou být něčím, co bychom chápali jako význam výrazu E: ten jistě není jen seznamem významů jednotlivých podvýrazů E.

3 Řešení

Pokud jde o intenze, je jistě intuitivní mluvit o tom, že empirické výrazy *označují* (netriviální) intenze. Výraz *největší planeta* nebude tedy označovat Jupiter, nýbrž 'individuovou roli', tj. funkci, která každému možnému světu přiřadí v jednotlivých okamžicích nejvýše jeden objekt (individuum), výraz *Měsíc je menší než Země* označuje nikoli pravdivostní hodnotu, nýbrž propozici, výraz *hnědý pes* neoznačuje třídu individuí, nýbrž vlastnost individuí atd. atd. To, co výraz označuje, tj. jeho denotát, není ovšem jeho význam. Význam je to, dík čemu výrazu rozumíme, a jistě tedy musí být strukturovaný. Viděli jsme, že Cresswellovy n-tice nesplňují naše očekávání, ale zatím jsme nenašli náhradu. Bez adekvátní explikace významu nepochopíme ovšem vazbu mezi výrazem a jeho denotátem. Význam zřejmě *vede k* denotátu, je to pojítko mezi výrazem a denotátem,

[4] (1975): 'Hyperintensional logic', *Studia Logica*, vol. 34, pp. 25-38, (1985): *Structured meanings*, Cambridge: MIT Press.

[5] Viz také Jespersen (2003) 'Why the tuple theory of structured propositions isn't a theory of structured propositions', *Philosophia*, vol. 31, pp. 171-83.

pokud možno jednoznačné. Co v té abstraktní říši může uspokojit naši intuici v tomto směru?

Již r. 1968 nalezl řešení Pavel Tichý (1968, 1969)[6] Tichý zavádí pojem (*abstraktní*) *procedury*, jak ho známe z teorie algoritmů (efektivních procedur). Když hledáme Fregovský *smysl* (zde budeme mluvit o *významu*) určitého výrazu E, pak narazíme na fakt, že jde o určitou proceduru, jejíž jednotlivé kroky odpovídají skládání určitých procedur, jež odpovídají významům jednotlivých podvýrazů E a jejichž výsledek je procedura odpovídající významu celého výrazu E.

Tento procedurální pohled je velmi přirozený a uspokojuje naše intuice: Tak jako matematické objekty jsou abstraktní a přitom s nimi zacházíme dnes a denně, tak významy jakožto abstraktní procedury mohou být uchopeny dík tomu, že známe významy jednotlivých slov a dovedeme s těmito významy zacházet, protože známe dík gramatice daného jazyka proceduru, která kombinací dílčích procedur dospívá k celé proceduře-významu.

Tichého idea procedurální definovatelnosti významů je nezranitelná Quineovou proslulou kritikou, protože se nesnaží definovat význam na základě pojmů analytičnosti a synonymie, nýbrž definuje význam nezávisle na těchto pojmech (což umožňuje naopak definovat analytičnost a synonymii na základě pojmu významu).[7] Z důvodů převážně mimoteoretických nebyla tato idea ve světové literatuře zaznamenána a dochází pozornosti až v době, kdy Tichý vytváří Transparentní intenzionální logiku (TIL)[8], v níž definuje nejdůležitější procedury jako tzv. *konstrukce,* jejichž teorii po formální stránce inspiroval (typovaný) λ-kalkul.

Montague[9] a další stoupenci jeho školy LANL rovněž využili typovaný λ-kalkul, ale nezabývali se explicitně konstrukcemi: λ-termy byly interpretovány jako funkce, které jsou konstrukcí konstruovány, nikoli jako konstrukce samy. Montaguovci tedy nemohou provádět analýzy, jejichž výsledkem jsou hyperintenzionální objekty. Nemohou tedy např. analyzovat postoje týkající se matematických objektů.

Sama idea strukturovaného významu je ovšem běžná (viz předchozí komentář ke Cresswellovi). Viz také Moschovakis, Y.N. (1994): 'Sense and denotation as algorithm and value', in: J. Väänänen and J. Oikkonen (eds.), *Lecture Notes in Logic*, vol. 2, Berlin: Springer, pp. 210-49.

4 Intermezzo: matematické výrazy

Souvislost významu a procedury lze dobře ilustrovat na analýze matematických výrazů.

Především je jasné, že matematické výrazy mají význam, přičemž nikdy neoznačují intenze. O významu jakožto intenzi se proto nikdy nemůže ani uvažovat. Konstrukce

[6] (1968): 'Smysl a procedura', *Filosofický časopis*, vol. 16, pp. 222-32, (1969): 'Intensions in terms of Turing machines', *Studia Logica*, vol. 26, pp. 7-25.

[7] Viz Materna (2007).'Once more on Analytic vs. Synthetic', **L**ogic and **L**ogical **P**hilosophy Volume 16 (2007), 3–43.

[8] Viz zejména Tichý (1988). *The Foundations of Frege's Logic*. De Gruyter.

[9] Montague (1974): *Formal Philosophy: Selected papers of Richard Montague,* Thomason, R. (ed.), New Haven.

jako význam matematického výrazu je naproti tomu velmi pochopitelná. Vezměme zcela jednoduchý příklad: výraz

a) $3 + 2 = 6 - 1$

Jak rozumíme tomuto výrazu: říká snad, že 5 = 5? To by se žáci procvičující takové příklady moc nenaučili. Číslo 5 není v a) zmíněno, protože v a) se nemluví o číslu 5, nýbrž o rovnosti výsledku sčítání na 3 a 2 s výsledkem odčítání 1 od 6. Význam věty a) je tedy určitá procedura, spočívající zhruba v těchto krocích:

i) Aplikuj operaci sčítání na dvojici (3, 2).
ii) Aplikuj operaci odčítání na dvojici (6, 1).
iii) Výsledek i) porovnej s výsledkem ii).
iv) iii): Jde o stejné číslo, rovnost je pravdivá.

Na matematických výrazech můžeme také ilustrovat problém *jednoduchých významů / pojmů.*

Za jednoduchý význam (jednoduchý pojem) pokládáme jednokrokovou proceduru spočívající v identifikaci příslušného objektu bez pomoci jiných procedur. Jak si můžeme ukázat právě na matematických výrazech, jednoduchý význam (pojem) je často neproveditelná procedura, když totiž jde o nekonečný objekt. Jednoduchý pojem prvočísla (reprezentovaný v TIL tzv. trivializací, tedy ^{0}prvočíslo) je procedura, která vede k aktuálnímu nekonečnu, v našem případě k celé nekonečné množině prvočísel. Jak si tedy vysvětlíme, že výraz *prvočíslo* je srozumitelný, že jeho význam je dostupný? Jde o to, že ^{0}prvočíslo není významem tohoto výrazu, je pouze ekvivalentní tomu významu: výraz *prvočíslo* vznikl jako zkratka, a kdo tomuto výrazu rozumí, ovládá některou z efektivních procedur, které jsou významem některé definice množiny prvočísel, přičemž tyto procedury nevedou k aktuálnímu, nýbrž jen potenciálnímu nekonečnu. Jedna taková procedura bude vypadat zhruba následovně:

Mějme nějaké (přirozené) číslo m.

i) Vytvoř třídu K čísel j takových, že m je dělitelné j.
ii) Obsahuje K přesně dva prvky? Jestliže ano, m je prvočíslo. Jestliže ne, m není prvočíslo.

Oba kroky obsahují konečně mnoho podkroků. Takto můžeme o každém čísle rozhodnout konečným počtem kroků, zda je či není prvočíslo.

Procedury, které jsou významem empirických výrazů, vedou k intenzím, tj. k funkcím z možných světů (+ časů) a neobsahují krok, který by určil, který svět je aktuální. Na rozdíl od matematických výrazů, kde význam vede k matematickému objektu, což je konečný cíl, jsou denotátem empirických výrazů funkce, jejichž hodnota v aktuálním světě (+ čase) nemůže být efektivně určena a kterou musíme teprve *empiricky* zjišťovat. Tuto hodnotu, kterou LANL nemůže 'vypočítat', nazýváme *referencí.* Například význam výrazu *největší planeta* je konstrukce takové funkce, která každému možnému světu přiřadí v daném čase nejvýše jeden objekt. Tedy: významem je ta konstrukce (k níž dojdeme analýzou toho výrazu), denotátem je ta funkce a referencí, tj. hodnotou té funkce v aktuálním světě nyní, je Jupiter. Cesta k denotátu je záležitost LANL, objev reference mají v popisu práce astronomové.

5 Úloha gramatiky

Představme si procedury:

A. Každé třídě objektů (individuí) přiřadí podle okamžitého stavu světa ten její prvek (pokud takový je), který je *největším* prvkem té třídy.
B. Každému okamžitému stavu světa přiřadí třídu objektů nazývaných *planeta.*
C. O každé dvojici individuí rozhodne v daném stavu světa, zda první člen *je menší než* ten druhý.
D. Danému stavu světa přiřadí individuum zvané *Slunce.*

Nyní uvažujme větu

Největší planeta je menší než Slunce.

Chybné chápání sémantiky by postupovalo tak, že by prostě přiřadilo procedury A. – D. po řadě výrazům *největší, planeta, menší než, Slunce.* Jak říká Tichý, tyto procedury či výsledky těchto procedur by z uvedené věty visely jako ozdoby na větvi vánočního stromku. Z tohoto pohledu by tyto jednotlivé významy nebo denotáty 'držely pohromadě' jen díky jazykovému výrazu, tj. rezignovali bychom na význam celé věty.

LANL v podobě analýzy na základě TIL nabídne konstrukci (w, t jsou proměnné po řadě možných světů a časů)

$$\lambda w \lambda t\ [^0\text{menší\_než}_{wt}\ [^0\text{největší}_{wt}\ ^0\text{planeta}_{wt}]\ ^0\text{Slunce}],$$

která reprezentuje význam celé věty, tj. je mimojazykovou procedurou, která není pouhým výčtem jednotlivých 'atomických' procedur, nýbrž je sama instrukcí, procedurou pracující s těmito atomy.

Z tohoto hlediska je gramatika daného jazyka souborem pravidel, která umožňují zakódovat způsob, jakým se jednotlivé konstrukce-významy spojují v nové konstrukce-významy. Stromy, které jsou výsledkem syntaktického parseru, by měly strukturálně odpovídat příslušné konstrukci, tj. vyhovovat principu kompozicionality, podle něhož ($m-$význam, F – syntaktická funkce, G – sémantická funkce, $\text{e}_1,\ldots,\text{e}_k$ podvýrazy výrazu E):

$$m(\text{F}(\text{e}_1,\ldots,\text{e}_k)) = \text{G}(m(\text{e}_1),\ldots,m(\text{e}_k)).$$

(viz Szabó, Z.G. (2005): "Compositionality". *The Stanford Encyclopedia of Philosophy* `http://plato.stanford.edu/archives/spr2005/entries/compositionality/`.)

Sémantika přirozeného jazyka a reálného světa – počítačové zpracování

Václav Matoušek, Roman Mouček, Pavel Mautner

Západočeská univerzita v Plzni
Fakulta aplikovaných věd, katedra informatiky a výpočetní techniky
{matousek, moucek, mautner}@kiv.zcu.cz

Abstrakt Článek se zabývá možnostmi počítačového zpracování sémantiky přirozeného jazyka a reálného světa a pokládá otázku, do jaké míry je toto zpracování možné a smysluplné. Odpověď pak hledá v kombinaci poznatků a zkušeností tří různých oborů – neurověd, lingvistiky a informatiky. Stručně je prezentován pohled neurověd na fungování lidského mozku a popsány paměťové složky mající vliv na zpracování sémantiky přirozeného jazyka a sémantiky vnějšího reálného světa. Krátce je představen i vnější, lingvistický pohled na přirozený jazyk a jeho sémantické roviny. Z informatických oborů jsou pak představeny přístupy umělé inteligence a softwarového inženýrství. Zmíněna je i vize sémantického webu.

1 Úvod

Zpracování sémantiky přirozeného jazyka[1] patří mezi problémy, se kterými si současné počítačové systémy a aplikace dokážou poradit jen částečně. Teorií, přístupů i experimentů, které se pokoušejí přirozený jazyk popsat a zpracovat i na sémantické úrovni, existuje samozřejmě velké množství, málokdy se však ptáme, proč se o vyřešení tohoto problému vůbec snažíme. Zvykli jsme si zadávat klíčová slova do internetových vyhledavačů, při telefonátech na linky mobilních operátorů či dopravních společností hledáme často nejkratší cestu k živému operátorovi, abychom se nemuseli „bavit“ s počítačovým dialogovým systémem a zda jsme nalezli ten správný dokument, který potřebujeme, poznáme nejlépe sami po alespoň letmém přečtení. Navíc si často nerozumíme ani mezi sebou (nedokážeme správně interpretovat promluvu jiného člověka). Existují kromě touhy po počítačovém „pokoření“ přirozeného jazyka i racionální důvody, proč se snažíme počítačově zpracovat i jeho sémantiku? Co nám toto zpracování může přinést a do jaké míry jej máme realizovat?

Sémantiku přirozeného jazyka můžeme chápat jako jazykovou interpretaci složitého vnějšího reálného světa a zároveň složitého vnitřního světa jednotlivce. Kromě smyslového zpracování vnější reality se na výsledné jazykové interpretaci podílí i řada interních procesů (schopnost kognitivního a emociálního zpracování, zkušenost, stupeň zvládnutí

[1] Počítačové zpracování sémantiky přirozeného jazyka tak, jak je chápáno v tomto textu, zahrnuje schopnost počítačového systému smysluplně interpretovat text či promluvu v přirozeném jazyce a posléze na tento text či promluvu adekvátně reagovat; zkoumána je schopnost počítačového „porozumění“ přirozenému jazyku.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 163–171, 2009.

daného jazyka apod.). Výsledné lidské zpracování sémantiky přirozeného jazyka je stejně jako další aspekty lidského chování a jednání ovlivněno jak geneticky, tak učením a životní zkušeností. Historicky lze počítačové zpracování sémantiky přirozeného jazyka spojit s vývojem řečových aplikací a automatických překladačů[2]. V aplikacích umělé inteligence je sémantická úroven jazyka řešena např. při vývoji dialogových informačních systémů založených na rozpoznávání přirozeného mluveného jazyka. Tyto systémy a aplikace se však v případě mluveného jazyka zaměřují většinou na zpracování signálu, určení slovních hypotéz a částečné zpracování syntaxe promluvy. Sémantika, pokud je vůbec zpracovávána, se obvykle omezuje na vyhledávání a porovnání významu slova vzhledem k systémové či doménové databázi. Systémy aplikující složité systémy pravidel, speciální formalismy, historii dialogu, pravděpodobnostní modely promluv apod. se objevily pouze v experimentální podobě (podrobněji např. [5]). Úspěšnější z existujících systémů pak „paradoxně“ využívají ke zpracování sémantiky hrubší sílu (rozsáhlé korpusy, statistika, jednoduchá pravidla), nežli sofistikované algoritmy a systémy složitých pravidel.

Dnes je pozornost věnována především počítačovému zpracování psaných dokumentů – vyhledávání informací v dokumentech, metodám organizace a klasifikace dokumentů (např. projekt WEBSOM [2]), vytváření významových sítí (např. projekt EuroWordNet [2]), anotaci korpusů, sumarizaci dokumentů, automatickému překladu apod. O zpracování sémantiky ve smyslu porozumění však často nejde. Řada metod přistupuje k dokumentu jako k souboru dat, který je nutné porovnat s jiným souborem dat (např. s dotazem uživatele nebo s jiným dokumentem), obsah souboru a jeho význam však dále nijak nezkoumá.

Velkou výzvou a potencionálním řešením problémů při pokusech o zpracování sémantiky přirozeného jazyka se stala snaha o realizaci tzv. sémantického webu. Zde však již nemůžeme hovořit o přirozeném jazyku a jeho zpracování (centrem zpracování není dokument), ale o datovém modelování reálného světa (zpracovávána jsou označkovaná data) a práce s konceptualizovanými daty ve webovém prostředí. I tato myšlenka a především pokus o realizaci (zaštiťovaný konsorciem W3C) však zatím nepřinesly očekávané výsledky. „Popovídat si“ s počítačem v přirozeném jazyce (ať textově nebo hlasově) je dnes stále možné jen ve velmi omezené doméně. Přes snahu odborníků z různých vědních disciplín navrhnout formalismy či metajazyky popisující libovolnou část sémantické roviny jazyka nedošlo v tomto směru zatím k významnějšímu pokroku. Proto je i v současné době vývoj aplikací, které zpracovávají sémantickou informaci přirozeného jazyka, velmi problematický. Proč je přirozený jazyk takto nezkrotný? Lze se v této situaci poučit u člověka, čili zkoumat, jakým způsobem zpracovává sémantiku lidský mozek? Je možné, že plné zpracování sémantiky přirozeného jazyka (jakéhokoli dokumentu či promluvy) současnými prostředky výpočetní techniky je nezvládnutelný úkol?

[2] V tomto případě je zpracovávána jak sémantika reálného světa, který se snaží přirozený jazyk postihnout (dialogový systém – modelování domény), tak sémantika samotného jazyka (dialogový systém – porozumění promluvě, automatický překlad).

2 Sémantika a lidský mozek

Z poznatků neurověd vyplývá, že funkční systémy lidského mozku (např. systém jazyk a řeč nebo paměť) vykazují znaky rozsáhlého distribuovaného systému, který je funkčně částečně modulární (podrobněji např. [4]). Tyto funkční systémy mají tzv. zúžené profily informačního chodu[3]. Součástí dlouhodobé vědomé paměti jsou i tzv. epizodická a sémantická paměť. Epizodická paměť je paměťová složka sloužící k zapamatování událostí vázaných na kontext, prostor a čas (události autobiografického charakteru), sémantická paměť slouží k zapamatování faktů, pojmů a významů „kontextově nezávislých". Obě paměti jsou částečně vázány na odlišné části mozku a částečně se překrývají. Vykazují vzájemnou spolupráci, jsou však schopny fungovat i samostatně, např. při významném poškození jedné z nich. Kromě epizodické a sémantické paměti lze identifikovat i další částečně oddělené paměťové složky, které se podílejí na zpracování přirozeného jazyka, např. lexikální a syntaktické informace[4]. Na výslednou interpretaci sémantiky přirozeného jazyka má pak vliv i emoční paměť.

Jestliže budeme předpokládat, že epizodická i sémantická složka paměti mají vliv na výsledné promluvy jednotlivce i na dokumenty, které tento jednotlivec napíše, pak stejně tak mají tyto paměťové složky vliv na interpretaci promluv jiných osob či dokumentů napsaných jinými lidmi. Jestliže vliv sémantické paměťové složky skýtá možnost smysluplného zpracování výsledných promluv nebo dokumentů, pak příspěvek epizodické složky paměti je pro počítačové zpracování sémantiky patrně neřešitelným problémem. Sepětí této paměťové složky s osobní zkušeností jednotlivce, která se navíc neustále vyvíjí a mění v čase, by znamenalo nutnost přizpůsobení počítačového systému této zkušenosti. Umíme dostatečně přesně modelovat všechny aspekty zkušenosti? Jsme schopni v reálném čase tyto údaje počítači předávat?

Větší možnosti nám poskytuje sémantická složka paměti. Nezávislost na kontextu odstinuje do značné míry osobní zkušenost, přesto by bylo iluzorní obsah sémantické složky paměti považovat za shodný u všech lidských jedinců. Můžeme však předpokládat, že vzájemná shoda bude narůstat u lidí žijících v stejném časovém období, ve stejné kulturní oblasti, v podobných sociálních podmínkách, s podobnou úrovní vzdělání atd. Vzájemnou shodu sémantické složky paměti podporuje v některých životních sférách i postupující globalizace.

Položky sémantické složky paměti vykazují vzájemný asociativní vztah na obecnější rovině. Tento vztah založený na „chaotickém" propojení obrovského množství synapsí je v reálných aplikacích modelován mnoha prostředky a formalismy (pravidla, gramatiky, umělé neuronové sítě, rámce, sémantické sítě, objektově orientovaný návrh apod.). Avšak každý z těchto prostředků se osvědčil pouze ve velmi specifických úlohách. Při experimentech na rozsáhlejších doménách dochází k nekontrolovatelnému nárůstu reálně nesmyslných vazeb.

Jak je možné, že v sémantické složce paměti vznikají „nesmyslné" vazby v menší míře než při použití modelovacích prostředků? Můžeme předpokládat, že vyšší reálnost sémantické složky paměti je výsledkem mohutné výpočetní kapacity neuronových

[3] Zúžený profil informačního chodu označuje místo, jehož poškození je kritické vzhledem k funkci příslušného funkčního systému.

[4] Některé experimenty připouštějí existenci hierarchie objektů ve spánkových lalocích.

obvodů? Znamená to, že modelovací prostředky jsou nepřesné jen proto, že obsahují málo výpočetních prvků?

Znamená to, že i základem lidského usuzování není nic jiného než dostatečný počet neuronů a synaptických spojení vznikajících nejprve na genetickém základě a poté na základě mapování reálných událostí nejprve do epizodické a posléze sémantické složky paměti? Znamená to, že pokud nemáme k dispozici dostatečnou výpočetní kapacitu, pak nemá smysl se pokoušet vymýšlet sofistikované formalismy pro zpracování přirozeného jazyka, neboť ty pak nezvládnou rozsah uchovávaných znalostí, nebo naopak produkují množství nesmyslných znalostí? Statistické metody jsou např. ve srovnání s formalismy založenými na systému pravidel relativně úspěšné, jejich spolehlivost roste s množstvím dostupných trénovacích dat.

3 Sémantika a lingvistika

Lingvistické teorie obecně zkoumají přirozený jazyk na základě jeho vnějších projevů. Sémantika jako lingvistická disciplína zaujímá ke zpracování sémantické informace přirozeného jazyka rozdílné postoje. Připouští, že vágnost a nejednoznačnost přirozeného jazyka je jeho základní vlastnost (teorie odpovídá konceptu paměťových složek). Tento přístup pak kontrastuje se snahou lingvistiky vybudovat obecnější úroven reprezentace sémantiky a obecnou sémantickou hierarchii. Někteří lingvisté se pak pokoušejí najít kontextově nezávislou úroven přirozeného jazyka (odpovídá „globalizované“ sémantické složce paměti), zatímco jiní přiznávají mnohoznačnost tzv. jednoduchého jazykového znaku[5], závislost významu na doméně (odpovídá sémantické složce paměti v rámci určité domény), situaci a individuální interpretaci (odpovídá epizodické složce paměti). Pak jen tzv. konceptuální a kolokační významová složka (vysvětleno např. v [5]) jednoduchého jazykového znaku může sloužit jako základ pro vytváření sémantických hierarchií a ontologií.

Sémantické teorie složitého jazykového znaku (promluvy) jsou velmi různorodé a neucelené. Připouštějí (podobně jako v případě jednoduchého jazykové znaku) jak existenci a neoddělitelnost jednotlivých významových složek, tak možnost pracovat s několika částečně nezávislými stupni porozumění. Výsledné popisy přirozeného jazyka a jeho sémantických rysů bývají bud vágní (a tedy obtížně aplikovatelné) pro počítačové zpracování přirozeného jazyka, nebo obsahují značné množství pravidel a výjimek a počítačové zpracování pak generuje značné množství nereálných jazykových konstrukcí či jejich interpretací (problém vzájemného mapování abstraktních vrstev popisu sémantiky, aplikace gramatik atd.). Mezi teorie zabývající se sémantikou přirozeného jazyka patří např. kontextový přístup, princip kompozicionality, syntakticko-sémantické větné vzorce a mimojazykové mikrosituace (podrobně viz [3]), model lidského chování nebo teorie aktuálního členění věty v češtině (podrobně v [8]).

Moderní lingvistika se zabývá i proměnlivostí sémantiky přirozeného jazyka a sémantického pole v čase, tedy posunu interpretace významu promluv a textů u skupiny lidí, která jazyk používá v různých časových obdobích.

[5] Pojmy jednoduchý a složitý jazykový znak jsou podrobně vysvětleny např. v [5].

4 Sémantika a informatika

Informatika se od svého vzniku věnuje především jazykům formálním. Přesto prudký rozvoj tohoto oboru vedl až k situaci, kdy jazykem formálním chceme zpracovávat jazyk přirozený. Zpracováním sémantiky přirozeného jazyka se v rámci informatiky nejprve zabývala umělá inteligence. Sémantické reprezentační a interpretační systémy vytvořené v rámci vývoje inteligentních dialogových systémů (přehled např. [5]) se vyznačují mnohými společnými vlastnostmi. Významné je především centrální postavení slovesa a volba abstraktní úrovně popisu významu ostatních složek věty či promluvy (nazývané koncepty, tematické role,...). Jelikož užití pouze jedné abstraktní úrovně popisu významu se ukazovalo jako nedostatečné, tvůrci dialogových informačních systémů přistupovali k definici několika abstraktních úrovní jak doménově závislých, tak doménově nezávislých. Vyvíjené metody umělé inteligence pro (nejen) zpracování sémantiky jsou dobře použitelné v úzkých a specifických situacích, v případě zpracování rozsáhlejší domény se i v této oblasti stále více prosazují metody založené na statistice, rozsáhlé datové základně a výkonném hardwaru.

Významný krok vpřed při modelování sémantiky reálného světa (a některých významových složek přirozeného jazyka) udělalo i softwarové inženýrství. Nelze zde samozřejmě hovořit o zpracování sémantiky přirozeného jazyka, lze však pozorovat zvyšující se úroven vyjádřitelné abstrakce, kterou dovoluje využít zvolený formální jazyk. Prostředky objektově orientované analýzy a objektově orientovaného návrhu tak poskytují aparát pro částečný popis sémantické paměťové složky (typicky asociativní vazba). Tento popis je samozřejmě zjednodušený, přesto však dostupnější a srozumitelnější širší odborné komunitě nežli prostředky a metody umělé inteligence.

Objektově orientovaná analýza definuje pohled na reálný svět bez ohledu na implementační prostředky. Analytický model vystihuje podstatu omezeného světa ve formě tříd objektů a jejich vzájemných asociativních vazeb. Je možné jej přirovnat k abstrakci sémantické složky paměti definující základní koncepty (pokud přijmeme i lexikální složku paměti, pak je modelována i tato) a jejich vzájemné asociace. Objektově orientovaný návrh pak analytický model dále rozšiřuje o definici vlastností a chování typizovaných objektů a o vzájemnou interakci těchto objektů. Modelování vlastností se zde liší od prostředků umělé inteligence. Zatímco v klasických formalismech jsou vlastnosti často definovány na stejné úrovni abstrakce jako související objekty, zde se stávají přímo součástí objektového paradigmatu a principu zapouzdření. Dalším modelovacím prostředkem je pak princip dědičnosti typů objektů (zavedení hierarchie).

Nalezneme podobné struktury i v oblastech sémantické složky paměti? Lze aktivaci neuronových oblastí v daném čase úspěšně modelovat definicí vlastností a způsobů chování typizovaných objektů? Můžeme předpokládat, že dva jevy prezentované na neuronové úrovni aktivací částečně společných souborů neuronů budou na strukturální úrovni prezentované alespon částečně stejnými vlastnostmi a chováním[6]? Z hlediska modelování reálného světa výše popsaným strukturálním způsobem je zřejmá snaha související jevy udržet pohromadě, a to asociativními vazbami nebo přímo na úrovni definice vlastností a chování objektů.

Další pohled na softwarovou realizaci jednotlivých složek paměti představuje model uchování persistentních objektů. Sémantika persistentních dat je nejčastěji definována

[6] Experimenty potvrzují, že související jevy aktivují částečně společné skupiny neuronů.

relačním modelem (v případě použití databáze) či tzv. sémantickými značkami na úrovni XML dokumentu. Z hlediska dalšího zpracování dat se však volnost při definici sémantických značek ukazuje jako nepraktická, neboť je nutná jejich další interpretace. Obecně pak hrozí nebezpečí vzniku několika vrstev popisů – metadat na úrovni různých vrstev abstrakce. Mapování mezi jednotlivými vrstvami abstrakce je pak obecně velmi problematické.

Za významný mezistupen (a možný bod setkání akademičtější umělé inteligence a praktičtějšího softwarového inženýrství) mezi komplexnějším zpracováním sémantiky přirozeného jazyka a typickým modelováním sémantiky reálného světa počítačovými systémy lze považovat vizi tzv. sémantického webu. Sémantický web [7] lze charakterizovat jako rozšíření současného webu takovým způsobem, že bude možná kombinace a integrace dat z různých zdrojů, a tak se významně zlepší spolupráce jak mezi lidmi, tak mezi počítačovými systémy. Tento přístup zahrnuje myšlenkový posun od zpracování a výměny dokumentů ke zpracování a výměně dat. Tato vize předpokládá, že data prezentovaná na internetu budou mít přesně definovaný význam a tím bude umožněno jejich strojové zpracování. Celá myšlenka sémantického webu tak počítá s konceptualizací dat (existencí doménových ontologií), existencí aktivních inteligentních komponent zabezpečujících požadavky uživatelů a standardizovaného popisu webových zdrojů[7]. Idea sémantického webu byla představena již v roce 2001 [1] a je podporována konsorciem W3C.

5 Závěr

Současné přístupy k počítačovému zpracování sémantiky reálného světa a přirozeného jazyka se střetávají s mnoha problémy a otázka možností zpracování sémantiky na úrovni strojového porozumění je stále otevřená. Přesto je z kombinace poznatků a zkušeností různých oborů (neurověd, lingvistika, informatika) zřejmé, kde se nachází v současné době realizovatelná hranice. Modelování světa a přirozeného jazyka jsou doménou oborů umělé inteligence a softwarového inženýrství. Zatímco klasická, abstraktněji orientovaná umělá inteligence staví na formalismech a zobecnění na úrovni gramatik či různých systémů pravidel, softwarové inženýrství využívá modelovacích prostředků prakticky využitelných při realizaci konkrétních aplikací ve velmi omezených doménách. Oba obory se pak potkávají při využití statistických metod či hardwarového výkonu. Jejich vzájemné provázání je pak pravděpodobné i při realizaci myšlenek sémantického webu.

Vezmeme-li v úvahu např. uspořádání neuronů a synapsí v lidském mozku, existenci asociativních vazeb, spolupráci sémantické a epizodické paměti, zásadní roli epizodické paměti v běžných životních situacích, emoční paměť jako významný interpretační mechanismus nebo zkušenosti s modelováním reálného světa v počítačových systémech, je jen velmi obtížné si představit komplexní zpracování přirozeného jazyka a vytvoření obecné abstraktní úrovně sémantické reprezentace přirozeného jazyka. I v případě modelování sémantiky jednoduchých domén jsou vytvořené modely velmi složité a neodrážejí komplexitu celé situace, či zavádějí nereálné vazby, vlastnosti a způsoby chování. Je tedy možné, že jediným modelovacím prostředkem pro popis a porozumění

[7] Práce s webem by byla obdobná práci s relační databází; význačným přínosem je značná relevance odpovědi na položený dotaz.

přirozenému jazyku je prostředek výpočetní kapacitou, uspořádáním a fungováním obdobný lidskému mozku.

Na začátku jsme se ptali, proč se snažíme počítačově zpracovávat přirozený jazyk. Možná to nejen neumíme, ale ani nepotřebujeme, a např. vize sémantického webu může určovat hranici, ke které z pohledu zpracování sémantiky jazyka a reálného světa stačí v informatice dojít.

Reference

1. Berners-Lee, T., Hendler, J., Lassila, O.: *The Semantic Web,* Scientific American, May 2001, dostupné online na http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2.
2. EuroWordNet, dostupné online na http://www.illc.uva.nl/EuroWordNet/.
3. Grepl, M., Karlík, P.: *Skladba češtiny*, Olomouc, 1998.
4. Koukolník, F.: *Paměť a její poruchy*, *Lidský mozek. Funkční systémy. Norma a poruchy*, Portál, Praha, 2002.
5. Machová S., Švehlová M.: *Sémantika & Pragmatická lingvistika*, Univerzita Karlova, Praha, 2001.
6. Mouček R.: *Sémantika v dialogových systémech,* disertační práce, Západočeská univerzita v Plzni, Plzeň, 2004.
7. Semantic Web, dostupné online na http://www.w3.org/2001/sw.
8. Sgall P., Hajičová E., Panevová J.: *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Academia, Praha, 1986.
9. WEBSOM, dostupné online http://websom.hut.fi/websom/.

Kajícný a nevěřícný – adjektiva na *-cí/-cný*: slovníky, gramatiky, korpusy

Klára Osolsobě

osolsobe@phil.muni.cz

Karlu Palovi ve vzpomínce na některá nelingvistická témata našich rozhovorů

Abstract. *Kajícný* and *nevěřícný* – adjectives on *-cí/-cný*: dictionary, grammar, corpora
The aim of this paper is to describe one type of the Czech deverbative adjectives on *-cí/-cný*. Relevant data extracted automatically (`http://deb.fi.muni.cz/deriv`) from the representative machine readable dictionary of Czech and from the corpora of literary Czech (ČNK: SYN2000, SYN2005, SYN2006PUB) will be presented. The description of the pairs on *-cí/-cný* in Czech dictionaries (PSJČ, SSJČ, SSČ) and in grammar will be examined and the characterisation of them compared with data mined from corpora. Afterwards the classification of this type will be suggested and the lemmatisation-rules for the automatic morphological analysis formulated. Finally the corpus based analysis of differences *kající/kajícný* and *nevěřící/nevěřícný* for lexicographical purpose will be tested.

1 Adjektiva na *-cí/-cný*: problém automatické morfologické analýzy

Ke zkoumání adjektiv na *-cí/-cný* dala podnět práce na derivačním slovníku češtiny navazující na algoritmický popis odvozování některých téměř paradigmaticky tvořených derivačních typů (Osolsobě 1996), mezi jinými derivace adverbií od adjektiv a tvoření syntetických tvarů II. a III. stupně adjektiv i adverbií. Při hledání lingvistických podkladů pro formulaci pravidel generování syntetických tvarů komparativu/superlativu adjektiv na *-cí* a možností automatické derivace adverbií a substantivních názvů vlastností na *-ost* od těchto adjektiv, jsme narazili na některé nepřesnosti ve stávajících popisech adjektiv na *-cí/-cný* (gramatiky, slovníky), které se odrážejí v praxi automatických morfologických analyzátorů používaných v českém prostředí. Na základě analýzy jazykových korpusů se pokusíme navrhnout řešení sporných míst přijatelná pro praxi automatické morfologické analýzy a uspokojivá z hlediska lingvistické teorie.

2 Automatický nástroj Deriv

Deriv (`http://deb.fi.muni.cz/deriv`) je nástroj vyvinutý na FI MU (srv. Hlaváčková, Osolsobě, Pala, Šmerk, 2009), který umožňuje jednoduchým způsobem vyhledávat ve strojovém slovníku automatického morfologického analyzátoru *ajka* (Sedláček, 2004)

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 171–186, 2009.

lemmata podle formálně zadaných pravidel (úvodní řetězec, koncový řetězec, morfologická značka) a vytvářet touto cestou seznamy slov, která s velkou mírou pravděpodobnosti patří k jednomu derivačnímu typu, respektive jsou tvořena jedním derivačním prostředkem. Se seznamy lze následně pracovat, je možné je prohlížet ve dvou modech, a to jako prosté seznamy nebo jako seznamy s uvedením frekvence vyhledaných jednotek v korpusu SYN2000 (viz níže). Nástroj tedy umožňuje rychlé prohledání a extrakci dat z rozsáhlého strojového slovníku českých kmenů (Osolsobě 1996, Sedláček 2004), je propojen s korpusem SYN2000 (100 milionů slovních tvarů), a umožňuje tak v korpusu ověřovat frekvenci jednotek zahrnutých ve strojovém slovníku.

Poznámka: Strojový slovník češtiny (Osolsobě 1996) byl vybudován na základě hesláře SSJČ a doplněn o řadu dalších slov na základě testování aplikací automatických morfologických analyzátorů *lemma* (Ševeček 1996) a *ajka* (Sedláček 2004) na korpusech češtiny (korpusy ČNK, korpusy budované na FI MU). Strojový slovník analyzátoru *ajka* zahrnuje zhruba 400 000 jednotek (kmenů), k nimž na základě formálních pravidel (deklinačních vzorů) generuje 6 milionů slovních tvarů.

2.1 Analýza materiálu

Pro vyhledání informací o možných slovotvorných vztazích adjektiv na *-cí/-cný*, adverbií na -c*ně* a substantiv na *-cnost* jsme použili automatický nástroj *Deriv* (webové rozhraní). Po přihlášení (uživatelské jméno/heslo) v *Deriv* zvolíme z nabídky funkci **Hledání slov podle pravidla**. Nejdříve zadáme do nabídky příkaz pro vyhledání slov končících na *-cí* a majících značku k2.* (adjektivum) a uložíme je do souboru (1). Pak zadáme příkaz pro vyhledání slov končících na *-cný* a majících značku k2.* (adjektivum) a uložíme je do souboru (2).

Vytvořené soubory (1) a (2) sloučíme a vytvoříme z nich soubor jeden (3). Na takto vzniklý soubor použijeme funkci **Hledej základová slova** a zadáme, že chceme v příslušném souboru najít dvojice slov takových, že jeden člen páru končí na *-cí* a druhý je slovo vzniklé odtržením *-cí* a jeho nahrazením *-cný*. Dále aplikujeme funkci **Rozdělení souboru** a získáme dva soubory, jeden, který obsahuje nalezené dvojice (4) a druhý, který obsahuje zbylá slova, ke kterým se nepodařilo najít podle příslušného pravidla slovo do „páru“ (5). Soubor (4) pak obsahuje kandidáty na hledaný derivační typ. Postup opakujeme pro vytřídění dvojic na *-cí/-cně* a *-cí/-cnost*. V tabulce 1 uvádíme výsledky.

Poznámka: Nástroj *Deriv* prochází rekonstrukcí směřující k optimalizaci. Použitá verze je nyní překonána. Pro účel naší studie byla vyhovující.

Poznámka: Dále sledujeme pouze deverbativní adjektiva na *-cí/-cný*. Do automaticky generovaného seznamu jsme doplnily nepravidelně tvořené *bojácný* a *budoucný*. Stranou našeho zájmu zůstalo nedeverbativní adjektivum *domácný* (uvádí je PSJČ i SSJČ, v korpusech se nevyskytlo ani jednou). Na internetu je doložen i v česky psaných textech tvar nejdomácejší (patrně vliv slovenštiny srv. Stich 1969: 64). K adjektivům (?) horoucí, žádoucí, ... se vrátíme níže.

3 Adjektiva (deverbativa) na *-cí/-cný* ve slovnících

V tabulce 2 je uveden přehled popisu adjektiv na na *-oucný /–ícný* v českých slovnících.

Tabulka 1. Tabulka výsledků při použití nástroje *Deriv*.

.*cí/k2.*	.*cný/k2.*	.*cně/k6.*	.*cnost/k1gF.*	komentáře
budoucí	-	*budoucně*	*budoucnost*	
domácí	*domácný*	*domácně*	*domácnost*	NEDEVERB
horoucí	*horoucný*	*horoucně*	*horoucnost*	?
jsoucí	*jsoucný*	*jsoucně*	*jsoucnost*	
kající	*kajícný*	*kajícně*	*kajícnost*	
mohoucí	*mohoucný*	*mohoucně*	*mohoucnost*	
obojlící	-	*obojlícně*	-	NEDEVERB
oboulící	-	*oboulícně*	-	NEDEVERB
nejsoucí	-	-	*nejsoucnost*	
nicí	-	-	*nicnost*	přegenerovávání
plecí	-	*plecně*	-	přegenerovávání
pomíjející	*pomíjejícný*	*pomíjejícně*	*pomíjejícnost*	
prací	*pracný*	*pracně*	*pracnost*	přegenerovávání
přehoroucí	*přehoroucný*	*přehoroucně*	-	?
přející	*přejícný*	*přejícně*	*přejícnost*	
přežádoucí	*přežádoucný*	*přežádoucně*	-	?
srdcervoucí	*srdcervoucný*	*srdcervoucně*	-	
strhující	*strhujícný*	*strhujícně*	-	
šikmolící	-	-	*šikmolícnost*	NEDEVERB
veležádoucí	*veležádoucný*	*veležádoucně*	-	?
vroucí	*vroucný*	*vroucně*	*vroucnost*	
vynikající	-	*vynikajícně*	-	
všemohoucí	*všemohoucný*	*všemohoucně*	*všemohoucnost*	
vševědoucí	*vševědoucný*	*vševědoucně*	*vševědoucnost*	
vědoucí	*vědoucný*	*vědoucně*	*vědoucnost*	
věřící	*věřícný*	*věřícně*	*věřícnost*	
zlolající	*zlolajícný*	*zlolajícně*	*zlolajícnost*	
žádoucí	*žádoucný*	*žádoucně*	*žádoucnost*	?
žhoucí	*žhoucný*	*žhoucně*	-	?
živoucí	*živoucný*	*živoucně*	*živoucnost*	?

Poznámka: „+“ značí, že slovo figuruje jako samostatné heslové slovo v příslušném slovníku, „0“ značí, že slovo je uvedeno jako existující české slovo uvnitř hesla adjektiva na *-cí* nebo slovesa, „-“ značí, že ve slovníku slovo zaznamenáno není (implicitně se předpokládá, že není sporu o tom, jak je odvozeno a jaký je jeho význam).

4 Adjektiva (deverbativa) na *-cí/-cný* v synchronních korpusech

Podívejme se ještě, jak jsou tvary na *-cí, -cný, -cnější* doloženy v synchronních korpusech. Tabulka 3 uvádí přehled výskytu adjektiv na *-cí*/*-cný*+*-cnějš-* ve třech synchronních korpusech. „X“ označuje výskyt adverbia *-cně*.

Tabulka 2. Adjektiva (deverbativa) na *-cí/-cný* ve slovnících

	PSJČ	SSJČ	SSČ
**bojící/bojácný*	-/+	-/+	-/+
budoucí/budoucný	+/-	+/-	+/-
horoucí/horoucný	+/+	+/+	+/0
přehoroucí/přehoroucný	+/+	+/-	-/-
jsoucí/jsoucný	+/+	-/+	-/-
mohoucí/mohoucný	+/+*	+/-	-/-
nemohoucí/nemohoucný	-/+	-/0	+/-
všemohoucí/všemohoucný	+/+**	+/-	+/-
srdvervoucí/srdcervoucný	+/-	+/-	+/-
vědoucí/vědoucný	+/+*	+/-	-/-
vševědoucí/vševědoucný	+/+**	-/-	+(pod *vše-*)/-
vroucí/vroucný	-/+	-/+	+/+
žádoucí/žádoucný	+/+	+/0	+/-
přežádoucí/přežádoucný	+/+**	+/-	-/-
veležádoucí/veležádoucný	+/+**	+/-	-/-
žhoucí/žhoucný	-/+**	-/-	+/-
živoucí/živoucný	+(*ne-*)/+**	+/+***	+/-
kající/kajícný	-/+	-/+	-/-
zlolající/zlolajícný	+/+*	+/0*	-/-
pomíjející/pomíjejícný	-/+	-/0***	+(*pominout*)/-
(ne)přející/(ne)přejícný	-/+	-/+	+(*přát*)/0
(ne)věřící/(ne)věřícný	-/-	-/-	+(*věřit*)/-

*Heslová adjektiva jsou označena křížkem (+).
** Heslová adjektiva jsou označena hvězdičkou (*).
*** U adjektiv je uveden příznak řidč., kniž.

Poznámka: Značkování sledovaných korpusů rozlišuje na úrovni tagsetu případy adjektivizovaných přechodníků (*AG.**), které mají na desáté pozici (atribut stupeň) hodnotu „-“ (nestupňuje se), a „obyčejných“ adjektiv (*AA.**), která mají na příslušné pozici uvedenu hodnotu stupně (1., 2. nebo 3.). Na úrovni slovníku však toto rozlišení dodržuje pouze v případech, kdy adjektivizovaný přechodník a adjektivum na *-cí* nejsou homonymní (tedy např. *žádoucí/žádající, ...*). Pokud jsou homonymní, rozlišena nejsou. (Např. se na úrovni morfologického značkování nerozlišuje adjektivizovaný přechodník *... aby strany <**přející**/AG.*> režimu převzaly ...* od dezaktualizovaného adjektiva *... uličnická, <**přející**/AG.*> a odvážná ...*). Druhým nedostatkem automatického značkování je nerozlišování substantivizovaných adjektiv na *-cí* (*všemohoucí, vševědoucí, věřící, nevěřící, nemohoucí, ...*). Uvedené počty adjektiv na *-cí* je tudíž třeba brát s touto rezervou.

4.1 Lemmatizace tvarů na *-cnější* v korpusech ČNK

V tabulce 4 je přehled tvarů se sufixem *-cn-ější* tvořených od adjektiv na ?*-cí/-cný* v českých korpusech. Lemmatizace naznačuje jisté rozpaky v popisu těchto tvarů (lemma je v některých případech tvar na *-cí,* v jiných tvar na *-cný,* jinde tvar sám).

Tabulka 3. Adjektiva (deverbativa) na *-cí/-cný* v synchronních korpusech

lemma	SYN2000	SYN2005	SYN2006PUB	internet*
bojící/bojácný	16/207+11	14/275+5	37/586+12	+/+
budoucí/budoucný	10778/1	8907/0	33466/2	+/+
horoucí/horoucný	177/1+7	288/4+12	169/1+2	+/+
přehoroucí/přehoroucný	1/0	0/0	0/0	+/X
jsoucí/jsoucný	215/0+1	214/0+1	100/0	+/+
mohoucí/mohoucný	14/1	4/0	3/0+1	+/+
nemohoucí/nemohoucný	137/2	87/0+2	65/0+2	+/+
všemohoucí/všemohoucný	404/0	458/0	174/0	+/+
srdcervoucí/srdcervoucný	48/0+1	95/0	90/0+3	+/+
vědoucí/vědoucný	144/0+3	187/0+1	226/0+7	+/+
vidoucí/vidoucný	89/0	168/0	130/0	+/+
vroucí/vroucný	49822/0+25	71230/0+30	525/60+41	+/+
žádoucí/žádoucný	3809/0+23	3753/0+44	6836/0+27	+/+
žhoucí/žhoucný	45/0	97/0	23/0	+/+
živoucí/živoucný	415/0+2	594/0+1	650/0	+/+
kající/kajícný	59/61	56/111+1	58/103+3	+/+
zlolající/zlolajícný	0/0	0/0	0/0	+/+
pomíjející/pomíjejícný	50/1+1	54/0	50/0	+/+
přející/přejícný	49/6+2	50/5	90/16+1	+/+
nepřející/nepřejícný	68/20	56/15+1	196/42	+/+
věřící/věřícný	2281/1	2988/0	737/0	+/+
nevěřící/nevěřícný	427/32	516/68	6032/89	+/+

*K dokladům z internetu neuvádíme pochopitelně frekvence, pouze výskyt. Tvar *srdcervoucný* se vyskytuje jenom v citátu z Haškova Švejka. Adjektivum *žhoucný* pouze v přehledu slovní zásoby mající vztah k počasí.

Poznámka: Ve druhém sloupci se uvádí počet lemmat adjektiv, která mají tvar II./III. stupně tvořen od derivačního kmene na *-cn-*. Ve třetím sloupci se uvádí dotaz na příslušný tvar (pomocí regulárních výrazů, kde „.“ znamená libovolný znak a „*“ opakování libovolného předchozího znaku), za lomítkem následuje lemma a opět za lomítkem počet výskytů v příslušném korpusu.

5 Slovníky a gramatika

Praxe popisu adjektiv (deverbativ) na *-cí/-cný* v českých slovnících je poněkud nepřehledná. Adjektiva na *-cí* (*-ou-cí/-í-cí-*) paradigmaticky tvořená od sloves (adjektivizované přechodníky/aktivní slovesná adjektiva) se neuvádějí. Výjimkou jsou případy (ne všechny), kdy se takto utvořené adjektivum osamostatnilo, nevyjadřuje již pouze aktuální vlastnost plynoucí z děje (dezaktualizuje se). Také vztahy mezi dvojicemi adjektiv na *-cí/-cný* jsou v každém ze sledovaných slovníků zachyceny jinak a uvnitř jednotlivých slovníků nejednotně (srv. výše).

Tabulka 4. Lemmatizace tvarů na *-cnější* v korpusech ČNK

	cnější	dotaz/lemma/frekvence
SYN2000	11	*.*vroucnější.*/ vroucný/ 25, .*žádoucnější.*/ žádoucný/ 23, .*horoucnější.*/ horoucný/ 7, .*bojácnějš.*í/bojácný/11, .*vědoucnější.*/ vědoucný/ 3, .*živoucnější.*/ živoucný/ 2, .*přejícnější.*/ přejícný/ 2, .*vidoucnější.*/ vidoucnější/ 1, .*srdcervoucnější.*/ srdcervoucnější/ 1, .*jsoucnější.*/ jsoucný/ 1, .*pomíjejícnější.*/ pomíjejícný/ 1*
SYN2005	14	*.*žádoucnější.*/ žádoucný/ 44, .*vroucnější.*/ vroucný/ 30, .*horoucnější.*/ horoucný/ 12, .*bojácnějš.*í/bojácný/5, .*nemohoucnější.*/ nemohoucný/ 2, .*matoucnější.*/ nejmatoucnější/ 1, .*živoucnější.*/ živoucný/ 1, (D|d)ivoucnější.*/ (D|d)ivoucnější/ 2, .*vědoucnější.*/ vědoucí/ 1, .*jsoucnější.*/ jsoucný/ 1, .*strhujícnější.*/ strhující/ 1, .*nepřejícnější/přejícný/ 1, .*zdrcujícnější.*/ zdrcujícnější/ 1, .*kajícnější.*/ kajícný/ 1*
SYN2006PUB	17	*.*vroucnější.*/ vroucný/ 41, .*žádoucnější.*/ žádoucný/ 27, .*bojácnějš.*í/bojácný/12, .*vědoucnější.*/ vědoucí/ 7, srdcervoucnější.*/ srdcervoucnější/ 3, .*horoucnější.*/ horoucný/ 3, .*kajícnější.*/ kajícný/ 3, .*nemohoucnější.*/ nemohoucný/ 2, sebevroucnější.*/ sebevroucnější/ 1, .* mohoucnější.*/ mohoucnější/ 1, .*zatěžujícnější.*/ nejzatěžujícnější/ 1, .*alarmujícnější.*/ nejalarmujícnější/ 1, .*nepřejícnější.*/ přejícný/ 1, .*odstrašujícnější/odstrašujícnější/ 1, .*strhujícnější.*/ strhující/ 1, .*znevažujícnější.*/ nejznevažujícnější/ 1, .*sebenepřejícnější.*/ sebenepřejícnější/ 1*

O tvarech II. a III. stupně na *-cnější* se ve slovnících neříká nic. Nejsou zmíněny, přestože slovníky uvádějí jak nepravidelné tvary II. (III.) stupně od supletivních kmenů (*lepší, horší, ...*), tak „nepravidelné“ stupňování (*sladší, bližší, ...*).

O „dezaktualizaci“ aktivních verbálních adjektiv se hovoří v Mluvnici češtiny I. (Dokulil 1986: 322, 330) jsou uvedena adjektiva tvořená sufixem *-ný* od aktivních slovesných adjektiv na *-c(í)*. Cituji: „Patří sem adj. z prézentních kmenů sloves jako *kajícný, přejícný, mohoucný, pomíjejícný, vroucný, vědoucný, jsoucný, horoucný, živoucný.* Pasívní význam má *žádoucný*, anomálně je tvořeno *bojácný*.“ S tímto citátem si dovolíme polemizovat. K adjektivům *horoucný, živoucný* nejsme schopni najít pravidlo podle něhož by se synchronně tvořila od „(!) prézentního kmene slovesa“. Podobně by synchronně pravidelně tvořeným adjektivem od prézentního slovesného kmene bylo **žádajícný* nikoli *žádoucný*. Pasivní význam má nejen adjektivum *žádoucný,* ale i adjektivum *žádoucí*. Anomálně tedy není dle našeho názoru tvořeno jenom adjektivum *bojácný*, ale i *horoucný, živoucný, žádoucný*. Formální anomálie adjektiva *bojácný* je ovšem zvýrazněna absencí příslušné varianty na *-cí* (viz níže).

Šmilauer v partii věnované popisu tvoření syntetických tvarů II. a III. stupně adjektiv na *-cí* píše, že tato adjektiva lze stupňovat přidáním -n-: *vroucí – vroucnější* (srv. Šmilauer 1971: 127). V Encyklopedickém slovníku češtiny (Karlík, Nekula, Pleskalová 2002: 447) stojí, že adjektiva na *-cí* a subjektově posesivní adjektiva na *-ův/-in* stupňovat nelze. Obojí tvrzení je třeba upřesnit. Synteticky lze stupňovat pouze dezaktivizovaná adjektiva

na *-cí* homonymní s aktivními slovesnými adjektivy/adjektivizovanými přechodníky, a to tak, jak uvádí Šmilauer (např.: SYN2006PUB ... *Za <**nejalarmujícnější**> považuje Čarnogurský ...*). Příklad, který Šmilauer zvolil, není dle našeho názoru zvolen šťastně, protože je ponechána stranou možnost, že stupňované tvary *vroucnější* se tvoří od tvaru *vroucný* nikoliv *vroucí*. K tvrzení v ESČ lze kromě upřesněného Šmilauerova pravidla dodat, že aktivní slovesná adjektiva/adjektivizované přechodníky lze stupňovat analyticky. Např.: *... přísný musulman, <**bojící se více**> jelita než střelné rány...* Analyticky se mohou stupňovat i další adjektiva na *-cí* (*... do hrnce s <**více vroucí**> vodou ...*) a nejen ona. Uvedený příklad není paradoxním stupňováním absolutní vlastnosti (= *bod varu*), ostatně ani to by nebylo, vzpomeňme na Orwella, v přirozeném jazyce ničím zvláštním, ale příkladem toho, jak se původní aktivní slovesné adjektivum (*... kapalina, <**vroucí**> při 129,5° ...*) dezaktualizuje a stává se synonymem k *vřelý, horký* (*... pokrmy ještě <**vroucí**> přemístěte ...*).

Otázkám stupňování dezaktualizovaných adjektiv a tvarům typu *nejvzrušující* se věnuje Stich (1969). V korpusu SYN2000 jsme nalezli doklady *nejvzrušujících/1, nejodstrašující/1, nejpomíjející/1, nejvzrušujícímu/1, nejpovznášejícím/1* (lemmatem je tvar sám a ve značce se uvádí na první pozici *X* – neznámý slovní druh). Pod vlivem slovenštiny lze na internetu nalézt tvary typu *nejdomácejší.*

Slovníky ani gramatiky neřeší problém lemmatizace, totiž zda jsou tvary na *-cnější* odvozeny od adjektiv na *-cný* nebo od adjektiv na *-cí*. Ačkoliv to není explicitně řečeno, implicitně se patrně předpokládá, že tvary na *-cí* a *-cný* jsou vůči tvarům na *-cnějš-* synonymní a je tudíž lhostejné, zda jde o derivaci *-cí*>*-cnější* nebo *-cný*>*-cnější.*

6 Návrh klasifikace adjektiv na *-cí/-cný*

V tabulce 5 uvádíme návrh systematické klasifikace vztahů aktivních slovesných adjektiv, dezaktualizovaných adjektiv (DA) na *-cí* a adjektiv na *-cný* tvořených od stejných základů.

1. Adjektiva *horoucí, žádoucí, ...* nejsou synchronně pravidelně tvořenými tvary aktivních slovesných adjektiv (srv. Gebauer 1909: 90). Tato adjektiva jsou synonymy adjektiv na *-cný* (*horoucí/horoucný, žádoucí/žádoucný, ...*). 2. K adjektivu *bojácný* neexistuje diachronně pravidelně tvořené **bojácí/bojácný* ani synchronně pravidelně tvořené **bojící/*bojícný*. Má tedy více anomálií než adjektiva 1. skupiny. 3. Adjektivum na *-cí* má dva (?více) významy(ů). První souvisí s primárním významem motivujícího slovesa, druhý je metaforický, specifikující atd. Adjektivum na *-cný* se vztahuje pouze k jednomu z více významů. 4. K synonymním dvojicím adjektiv na *-cí/-cný* neexistuje aktivní slovesné adjektivum. Patří sem jednak dvojice *budoucí/budoucný*, dále kompozita s druhým členem deverbativním adjektivem na *-cí*. 5. Dezaktualizovaná adjektiva na *-cí* homonymní s aktivními slovesnými adjektivy a synonymní s tvary na *-cný*. 6. Dezaktualizovaná adjektiva na *-cí* homonymní s aktivními slovesnými adjektivy, k nimž se tvoří tvary II. (III.) stupně přidáním *-c-n-ějš-*.

7 Lemmatizace tvarů na *-cnější* a formální analýza

Otázka, na niž je třeba při aplikaci pravidel pro formální popis odpovědět, je, jak lemmatizovat tvary na *-cn-ějš-í,* tj. jaké lemma se má přiřadit, zda tvar (nom. sg. mask.

Tabulka 5. Návrh klasifikace adjektiv na *-cí/-cný*

	základové sloveso	**A** : akt. sloves. adj.	**B** : DA na -c*í*	**C** : adj. na -*cný*
1.	žít	žijící (kde/kdy)	žijící (*klasik*)	0
		0	živoucí (*bytost*)	živoucný
	hořet	hořící (kde/kdy)	?hořící (*keř*)	0
		0	(pře)horoucí (*peklo*)	(pře)horoucný
	žádat	žádající (o koho/co)	?žádající (*ruka*)	0
		0	žádoucí (*otěhotnění*)	žádoucný
	žhnout	žhnoucí (kde)	žhnoucí (*slunce*)	0
	+žíci	0	žhoucí (*uhlíky*)	žhoucný
2.	bát (se)	bojící se (koho/čeho)	0	0
			0	bojácný
3.	vřít	vroucí (kde)	vroucí$_1$(*voda*)	0
		0	vroucí$_2$(*láska*)	vroucný
	(ne)věřit	(ne)věřící (komu/čemu)	(ne)věřící$_1$ (*žid*)	0
		0	(ne)věřící$_2$ (*pohled*)	(ne)věřícný
4.	být	0	budoucí (*matka*)	budoucný
4a.	moci	0	všemohoucí (*stvořitel*)	všemohoucný
	rvát	0	srdcervoucí (*výkřik*)	srdcervoucný
	vědět	0	vševědoucí (*vypravěč*)	vševědoucný
	lát	0	zlolající	zlolajícný
5.	být	jsoucí (kde/kdy)	jsoucí (*život*)	jsoucný
	(ne)moci	(ne)mohoucí (dělat co)	(ne)mohoucí (*pacient*)	(ne)mohoucný
	vědět	vědoucí (o kom/čem)	vědoucí (*úsměv*)	vědoucný
	vidět	vidoucí (koho/co)	vidoucí (*zraky*)	vidoucný
	kát (se)	kající se (z čeho)	kající (*hříšník*)	kajícný
	pomíjet	pomíjející (koho/co)	pomíjející (*čas*)	pomíjejícný
	(ne)přát	(ne)přející (komu/čemu)	(ne)přející (*povaha*)	(ne)přejícný
6.	alarmovat	alarmující (koho/co)	alarmující (*zjištění*)	*alarmujícný/-alarmujícnější
	odstrašovat	odstrašující (koho od čeho)	odstrašující (*příklad*)	*odstrašujícný/-ostrašujícnější

živ. pozitiv) na *-cn-ý,* nebo na *-c-í,* tedy je-li lemmatem tvaru *kajícnější ?kajícný* nebo *?kajícící.*

V případě existence synonymních lemmat je třeba stanovit přesná pravidla pro to, kterému lemmatu dává příslušný popis (příslušný automatický morfologický analyzátor) přednost.

Jednoduše by mohla fungovat následující pravidla: 1. V případě, že je doložen tvar na *-cný*, je lemmatem tvaru na *-cnější* tvar na *-cný* (skupina 1-5). 2. V případě, že není doložen tvar na *-cný*, je lemmatem tvaru na *-cnější* tvar na *-cí* (skupina 6).

Jednoduchost tohoto formálního pravidla by mohl narušit jedině FBL (fucking bloody linguist), a to v případě, že by z existence nesynonymních tvarů adjektiv na *-cí/-cný*, vyvodil existenci homonymních tvarů II. a III. stupně adjektiv na *-cnější*.

Podobným případem je i lemmatizace tvarů II. a III. stupně adjektiv lišících se v I. stupni pouze sufixem *-ní/-ný* např. *sluneční/slunečný*. Ty by pak bylo třeba při automatické morfologické analýze podrobit disambiguaci (skupina 3).

Další problém automatické lemmatizace představují tvary na *-(e/ě)jší* s prefixoidem *sebe.*

Poznámka: Lemmatem tvarů II. a III. stupně je zpravidla tvar pozitivu. Tvary *sebe.***ější* jsou formální komparativy. Proto jsou adjektiva (*sebevětší, ...*) lemmatizována lemmatem *sebe.***ší*. Na 10. pozici (stupeň) je uvedena hodnota 1 (pozitiv). Lemmatizace není provedena důsledně. Řada adjektiv (mezi nimi i *sebevroucnější, sebenepřejícnější)* mají značku $X.*$ (neznámý slovní druh) a jako lemma je uveden příslušný tvar sám.

7.1 *Kajícný* a *nevěřícný* („šetření FBL“)

Slovníky (především praxe SSČ) napovídají, že až na výjimky jsou adjektiva na *-cný* variantami dezaktualizovaných adjektiv na *-cí*.

Poznámka: Odpovědět na otázku, jakými směry (co od čeho) se ubíraly derivace adjektiv na *-cný,* adverbií na *-cně* a substantiv na *-cnost*, ponecháváme v této studii stranou (srv. více Stich, 1969).

Podíváme-li se podrobněji na statistické údaje získané analýzou korpusů, na první pohled zaujme adjektivum *kajícný*. Zatímco ve všech ostatních případech mají tvary adjektiv na *-cí* výrazně vyšší frekvenci než tvary adjektiv na *-cný,* u tohoto adjektiva převažují, a to v korpusech SYN2005 a SYN2006PUB dokonce výrazně převažují, tvary na *-cný*. Relativně vyšší frekvenci než ostatní adjektiva na *-cný* má i adjektivum *nevěřícný*.

Otázka, kterou si na základě empirického pozorování klademe zní: Signalizuje toto „vybočení z řady“ něco, nebo jde o náhodu? Jak jsou tato adjektiva zaznamenána v českých slovnících?

	PSJČ	SSJČ	SSČ
kající/kajícný	-/+	-/+	+/0
nevěřící/nevěřícný	-/-	-/-	+/-

Poznámka: „+“ značí, že slovo figuruje jako samostatné heslové slovo v příslušném slovníku, „0“ značí, že slovo je uvedeno jako existující české slovo uvnitř hesla adjektiva na *-cí* nebo slovesa, „-“ značí, že ve slovníku slovo zaznamenáno není (implicitně se předpokládá, že není sporu o tom, jak je odvozeno a jaký je jeho význam).

Přestože mají adjektiva *kajícný* a *nevěřícný* po adjektivu *bojácný* v korpusech nejvyšší frekvenci ve srovnání s ostatními adjektivy na *-cný*, nejsou zaznamenána v SSČ jako samostatná heslová slova. Adjektivum *kajícný* je uvedeno pod heslem *kající, nevěřícný* (ani *věřícný*) uvedeno není.

Poznámka: SSČ má samostatné heslové adjektivum *vroucný* i *vroucí* (2 významy) a *horoucný, nepřejícný* pod hesly *horoucí* a *nepřející*. Adjektiva *věřící/věřícný, přející/přejícný* jsou v korpusech častěji doložena jako negativa, tedy *nevěřící/nevěřícný, nepřející/nepřejícný*. Ačkoliv jsou adjektiva (nejen adjektiva) s prefixem *ne-* lemmatizována automatickými morfologickými analyzátory používanými v českém prostředí tvarem nom. sg. mask. bez prefixu *ne-*, zdá se, že kvantitativní analýza naznačuje jisté rozdíly v distribuci tvarů s +/- *ne-*. Tyto rozdíly pramení z toho, že v případě

některých adjektiv nejde o čistě záporová opozita, nýbrž o opozita, která se lexikalizují v různých významových odstínech (viz níže).

Pokusme se na základě kolokací (vyhledaných automaticky pomocí statistik MI-score) dvojic *kajícný/kající, nevěřícný/nevěřící* porovnat významy dvojic, které jsme v našem přehledu zařadili do různých skupin (3 a 5) – viz tabulka 6.

Tabulka 6. Kolokace dvojic *kajícný/kající, nevěřícný/nevěřící*

	SYN2000	SYN2005	SYN2006PUB
kajícný	doznání, odsouzenec, hříšník, mafián	doznání, půst, žalm, bohoslužba	doznání, odsouzenec, hříšník, mafián
kající	hříšnice, Magdalena, bohoslužba	Indikoplev, diákon	Magdaléna, mafián, hříšník
nevěřící	Tomáš, pes, člověk, občan	úžas, zraky, Tomáš	Tomášová, Tomáš, výraz, pes
nevěřícný	kroucení, úžas, výraz, hlava, pohled	údiv, úžas, výkřik, výraz, pohled	kroucení, úžas, údiv, výraz, úsměv

Poznámka: Práci s kolokacemi do jisté míry znesnadňuje stav anotací sledovaných korpusů. Na úrovni morfologické značky nejsou rozlišena adjektivní a substantivizovaná užití lemmat na *-ící* (*nevěřící*) (srv. Wagner, 2005). Tento nedostatek je patrný např. z výskytu kolokátu *věřící* (opozice substantiv *věřící nevěřící*). Z kolokací jsme tudíž vybírali substantiva (adjektivum rozvíjí příslušné substantivum).

Adjektiva *kající/kajícný* mají některé kolokáty společné (*mafián, hříšník, bohoslužba*), což by mohlo svědčit o tom, že jde o blízká (kontextově nahraditelná) synonyma (srv. výše jejich zařazení do skupiny 5).

Malá teologická poznámka: Jestliže se někdo *kaje* (aktuálně je *kající se*), není možné, aby týž člověk (obdařený svobodnou vůlí) zaručil, že se mu podaří odstranit v dalším životě to, co je předmětem jeho *pokání*. (A to ani tehdy, bude-li tento *kající se kajícník* mít to, čemu katolická dogmatika říká účinná lítost, tj. předsevzetí změnit to, z čeho *se kaje*). Měl by ale zůstat *kajícím* či *kajícným*, i tehdy, když by se podařilo realizovat jeho předsevzetí. Napětí mezi moralismem a milostí plyne z toho, že moralistovi stačí, že dodržuje zákon a Boha nepotřebuje. K dosažení odpuštění nestačí dle Evangelia dodržení zákona, ale milost, t.j. vědomí, že potřebujeme Boha a jeho odpuštění, nikoliv jen zákon a jeho dodržení. Tím se zákon neruší, pouze se staví až na druhé místo. (srv. Ratzinger 2007, s. 91 výklad k Lk. 18, 9-14).

Ačkoliv se adjektiva *nevěřící/nevěřícný* shodují v kolokátech *úžas, výraz*, ze seznamu konkordančních řádků je na první pohled patrné, že vzájemná záměna není možná ve všech kontextech. Ačkoliv jsou kolokace *nevěřící Tomáš* (narážkou na něj je i kolokace *nevěřící Tomášová*), *nevěřící pes* frazémy/idiomy, na internetu se objevuje i *nevěřícný Tomáš*. Význam adjektiva *nevěřící* je 1. jiné víry, bez víry (*Za oltářem dva <nevěřící> vojáci močili*), 2./(?3). jsoucí k nevíře, nedůvěřivý (*... nad nímž se srdce svíralo zhnuseným, <nevěřícím> úžasem ... ... na tváři se jí objevil <nevěřící> výraz ...*). V prvním významu

nelze adjektivum *nevěřící* nahradit adjektivem *nevěřícný*. Rozbor kolokací založený na korpusech potvrzuje vzájemnou nahraditelnost dvojic na *-cí/-cný* v případě adjektiv *kající/kajícný*, nikoli *nevěřící/nevěřícný*. Pokud bychom připustili, že lze použít v češtině tvar *nevěřícnější* ve významu *ateističtější*, pak by lemmatem tohoto tvaru muselo být *nevěřící* nikoli *nevěřícný* a zároveň bychom dokázali, že z výše formulovaného pravidla lemmatizace mohou existovat výjimky.

Použití výše uvedeného pravidla pro praxi automatických anotačních nástrojů přesto pokládáme za vyhovující. Z uvedených „šetření" je patrné, že „zjednodušující řešení" (které se navíc týká jednotek okrajových, jež lze předpokládat, nikoli doložit) nebrání v objevování rozdílů, které automaticky vkládané anotace ponechávají a mohou i nadále ponechávat stranou. Anotace korpusů nejsou a nemají být alfou a omegou lingvistických analýz (srv. Leech, 1993), nýbrž praktickým nástrojem usnadňujícím prohledávání rozsáhlých dat.

Poznámka: Potenciální *nevěřícnější (ateističtější), vroucnější (vřelejší/?vařícnější)* jsme ověřili elicitací.

Výsledky dotazníku shrnuje tabulka. Respondenti (informátoři) byli studenti FF (nebohemisté) a gymnazisté. Adjektiva byla uvedena v krátkých textech (viz níže).

Dotazník:

I.

*„"**Nevěřícnější** národ, než jsou Češi, abys pohledal, řekl smutně a podíval se po mně zkoumavě, co tomu říkám.*

„Jako že jsme ateisti?"

„To má být otázka?"

„Spíš odpověď," myslím si a mažu rukou klikyháky, které jsem si bezmyšlenkovitě vyryl do písku uhlazeného vlnkami narážejícími na břeh.

Usmál se. „Snad mi neřekneš, abych hodil první kamenem."

II.

Postavila před Karlíka talíř zelňačky. Ponořil lžíci do polívky, nabral a nesl k ústům. Tu se nesnesitelně rozeřval.

„Neječ!!!" zapištěla.

Dítě se rozeřvalo ještě hlasitěji. Nečekaně ji ruka vylétla. Řev zesílil.

„Já už ti nic vařit nebudu."

Hysterka, pomyslela si. Stává se jí to čím dál častěji, co je sama.

„Nechutná ti to?"

*Zvedl k ní uslzenou zarudlou tvářinku. „Je to ještě "**vroucnější** než včera a pusu mám spálenou už dvakrááát, protáhl fňukavě a popotáhl.*

Zarýmovanej, na zítra zase hlídání místo školky.

„Nebreč," osopila se na něj. „Dám ti do toho kostku ledu," dodala smířlivě. Sama bych si tak dala led do campari, pomyslela si a už se jí ani nechtělo stydět.

	nepřijatelné		zvláštní, ale hodí se		přijatelné	
nevěřícnější = ateističtější	11	25,6 %	32	74,4 %	0	0 %
vroucnější = vřelejší (polévka)	29	67,4 %	9	21 %	5	11,6 %

Přestože se zdá, že pro rodilé mluvčí je podstatně přijatelnější spojení *nevěřícnější národ* než *vroucnější polévka* napadá nás, jak by asi vypadala disambiguace v případě, že by automatický analyzátor nabízel všechny možnosti, které se honí hlavou FBL.

Lingvisticky ne zcela vyhovující by se mohlo jevit to, že všechna adjektiva, jejichž tvary lze automaticky generovat od tvarů sloves (přechodníků přítomného) mají v korpusech ČNK značku *AG.....-.*,* tedy na desáté pozici (atribut stupeň) je uvedena hodnota „-“ (neurčuje se). Adjektiva jsou automaticky pokládána za nestupňovatelná. Případná změna anotačního schématu (srv. výše 6. skupina) se nám zdá být schůdnější než pokus o automatické rozlišování (disambiguaci) homonymních aktivních slovesných adjektiv/adjektivizovaných přechodníků přítomných a dezaktivizovaných „obyčejných“ adjektiv na úrovni automatické morfologické analýzy. Toto řešení odpovídá i praxi automatického morfologického analyzátoru *ajka* (Sedláček 2004), kde se na úrovni morfologické značky nerozlišuje stupňovatelnost (kombinace hodnot 2. pozice a 10. pozice v systému značkování ČNK srv. Hajič 2004) ale pouze hodnota stupeň (u „nestupňovatelných“ adjektiv je uveden 1. st.).

8 Závěr

Cílem prezentované analýzy bylo zjistit vztahy mezi adjektivy na *-cí* a adjektivy na *-cný*. Pomocí automatického nástroje *Deriv* jsme vyhledali dvojice na *-cí/-cný, -cí/-cně* a *-cí/-cnost.* Snažili jsme se vysledovat souvislosti mezi blokací kombinovatelnosti sufixů s *-c—* se sufixy (*-ě/e)jší, -ě/e, -ost* a existencí adjektiv (deverbativ) na *-cí/-cný* (srv. Trávníček, 1951, s. 354).Zabývali jsme se popisem automaticky vybraných dvojic ve slovnících a gramatikách. Sledovali jsme výskyt adjektiv na *-cí/-cný* i výskyt tvarů komparativu/superlativu na *-cnější* v korpusech (na internetu). Na základě porovnání popisu zkoumaného typu adjektiv na *-cí/-cný* ve slovníku/gramatice s korpusovými daty jsme se pokusili systematičtěji popsat jeden úsek české slovotvorby. V těchto souvislostech jsme se zabývali stanovením pravidel lemmatizace tvarů na *–cnější* pro možné aplikace automatické morfologické analýzy.

Dvojice adjektiv *kající/kající* a *nevěřící/nevěřící*, která jsou ve všech třech sledovaných korpusech doložena lépe než ostatní adjektiva, jsme zkoumali z hlediska významových diferencí. (Využili jsme statistik MI-score.) Korpusová analýza potvrdila, že zatímco první dvojice je prakticky synonymní, u druhé lze rozlišit významy synonymní i nesynonymní. Snažili jsme se tak poukázat na možnost využít korpusů při lexikografickém zpracování jevů, které dosavadní slovníky zachycují nejednotně.

Poznámka: Zajímavé by mohlo být sledovat příslušná adjektiva diachronně. V korpusu diakorp je doloženo pouze adjektivum (*neprobyšúcný/1, vroucný/2, vroucnější/7*), adverbia *vroucně, budoucně, horoucně* a substantiva *budoucnost, vroucnost, všemohoucnost, horoucnost, nemohoucnost, (ne)pomíjejícnost, nekajícnost*. Řídké doklady neskýtají oporu pro korpusově podložený výzkum.

Reference

1. Dokulil, M., Komárek, M. a kol.: Mluvnice češtiny I, II., Praha : Academia 1986.
2. Filipec, J. a kol.: Slovník spisovné češtiny pro školu a veřejnost (SSČ). Praha : Academia 2005.

3. Gebauer, J. : Historická mluvnice jazyka českého III., Praha : Unie 1909.
4. Hajič J.: Desambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum : Praha 2004.
5. Havránek, B. a kol.: Slovník spisovného jazyka českého (SSJČ). Praha : Academia 1989.
6. Hlaváčková, D., Osolsobě, K., Pala, K., Šmerk, P.: Relations between formal and derivational morphology in Czech, Abstract for CFG conference 12.2. - 14.2. 2009.
7. Hladká, Z., Karlík, P.: Kam s ním? (problém stupňování adjektiv). In: Karlík, P., Pleskalová, J. (eds.), Život s morfémy. Brno : FF MU 2004, s. 73-93.
8. Hujer, O., Smetánka, E., Weingart, M., Havránek, B., Šmilauer, V., Získal, A. (red.): Příruční slovník jazyka českého (PSJČ). Praha : Státní nakladatelství 1935-1957.
9. Karlík, P., Nekula, M., Pleskalová, J.: Encyklopedický slovník češtiny. Praha : Nakladatelství Lidové noviny 2002.
10. Leech, G.: Corpus annotation schemes, Literary and linguistic Computing 8 (4), 1993, s. 275-281.
11. Osolsobě, K.: Algoritmický popis české morfologie a strojový slovník češtiny, Brno : FF MU, (disert. práce) 1996.
12. Osolsobě, K.: Čeho je moc, toho je příliš aneb jaké má čeština komparativy a superlativy? In: Přednášky a besedy z XLI. běhu LŠSS, Brno, 2008, s. 147-160.
13. Osolsobě, K.: K jednomu typu vyjadřování stupně v češtině. Bohemica Olomucensia (3 (I), řada Lingvistica Juvenilia. V tisku.
14. Ratzinger, J.: Jesus von Nazareth. Verlag Herder : Freiburg – Basel – Wien 2007.
15. Rejzek, J.: Český etymologický slovník. Voznice : LEDA. 2001.
16. Rychlý, P.: Bonito – grafické uživatelské rozhraní systému Manatee, Verze 1.49. 1998-2003. Dostupné z `http://ucnk.ff.cuni.cz/bonito/`
17. Sedláček, R.: Morphematic analyser for Czech. Brno : FI MU, (disert. práce) 2004.
18. (Morfologický analyzátor *ajka* dostupný z `http://nlp.fi.muni.cz/projekty/wwwajka`.)
19. Stich, A.: Stupňování přídavného jména „vzrušující“, Naše řeč 52, 1969, s. 62-64.
20. Šimandl, J.: Deverbativní adjektiva a jejich konkurenty. In Karlík, P.(ed.): Korpus jako zdroj dat o češtině. MU : Brno 2004, s. 135-143.
21. Šmilauer, V.: Novočeské tvoření slov. Praha : Státní pedagogické nakladatelství 1971.
22. Trávníček, F.: Mluvnice spisovné češtiny. Praha : Slovanské nakladatelství 1951.
23. Wagner, R.: Das Auffinden von reflexiven Verbalsubstantiven im tschechischen Nationalkorpus: Grenzen der morphologischen Annotation. In Štícha, F., Šimandl, J.: Gramatika a korpus 2005. Praha : Ústav pro jazyk český, 2005, s. 295-304.
24. Český národní korpus. Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: `http://ucnk.ff.cuni.cz`.
25. Deriv `http://deb.fi.muni.cz/deriv`
26. DEBDict – internetový prohlížeč slovníků `http://deb.fi.muni.cz/debdict`
27. Internetový prohlížeč Google: `http://www.google.com/`

Postavení příklonek v české klauzi v korpusech současné češtiny

Vladimír Petkevič

Ústav teoretické a komputační lingvistiky FFUK

0 Úvod

V tomto příspěvku se budu zabývat především hlavními stálými příklonkami (srov. Karlík et al. 1995, s. 647–651) v češtině, tj. slovními tvary *by, bys, ... , ho, -li, mi, mu, se, ses, si, sis, tě, ti,* a dále některými příklonkami nestálými: *je* (jakožto tvar zájmenný i slovesný), *ji, jí, jich, jim, jsem, jsi, jsme, jste, mě, mne, mně, sme, ste* a jejich vzájemným slovosledným postavením. O příklonkách v dnešní češtině se v české jazykovědné bohemistice psalo již vícekrát, psali o nich v poslední době mj. Avgustinova & Oliva (1995), Toman (2001), Uhlířová (2001), Oliva (2001) a Rosen (2001), v tomto příspěvku bude však zkoumání příklonek snad poprvé opřeno o velké korpusy češtiny SYN, SYN2005, SYN2000 a SYNEK (srov. Český národní korpus 2000, 2005, 2006, 2007). Příspěvek si klade tyto dva hlavní cíle:

a) stručně popsat slovosled stálých příklonek, a přispět tak ke zpřesnění teoretického popisu syntaxe češtiny
b) předvést, jak se zjištěných faktů dá využít k automatické morfologické disambiguaci a povrchové syntaxi českých textů.

Příspěvek se skládá z těchto částí:

1. Přehled nejběžnějších příklonek
2. Slovosledné postavení vybraných příklonek
3. Postavení příklonek a automatická slovnědruhová a morfologická disambiguace češtiny
4. Shrnutí

1 Přehled nejběžnějších příklonek

V příspěvku se budu zabývat zejména stálými příklonkami, a to jak z hlediska jejich vzájemného slovosledného postavení, tak z hlediska využití zjištěných faktů k automatické disambiguaci češtiny. Níže uvádím přehled nejběžnějších příklonek (tedy stálých a nestálých) a ty dělím podle dvou kritérií:

A) na:

– příklonky stálé (níže jsou podtržené), které nikdy nejsou nositeli přízvuku
– příklonky nestálé (níže nepodtržené), které mohou, ale nemusí být nositeli přízvuku

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 185–198, 2009.

(srov. Karlík et al. 1995, s. 647-651).

B) podle jemněji rozčleněných slovních druhů na:

B1) osobní zájmena: *ho, je, ji, jí, jich, jim, mě, mi, mne, mně, mu, nám, nás, ti, tě, vám, vás*

B2) reflexiva:*se*, *ses*, *si*, *sis*

B3) ukazovací zájmeno*: to*

B4) slovesné tvary*: je, jsem, jsi, jsme, jste* a nespisovné tvary *sem, si, sme, ste*; dále *byl*, *byla*, ...

B5) tvary kondicionálu:*bych*, *bys*, *by*, *bychom*, *byste* a nespisovné *bysme*

B6) adverbia*: sem, tu*

B7) spojka: *-li*

Z uvedeného přehledu dále vyberu úzkou skupinu příklonek, zejména stálých (tedy výše podtržených), a zaměřím se na jejich vzájemné slovosledné postavení.

2 Slovosledné postavení vybraných příklonek

Nejobsáhlejší popis příklonek ve všeobecných gramatikách současné češtiny jsem nalezl v Příruční mluvnici češtiny (Karlík et al. 1995) na s. 647–651 a v práci Rosen (2001). Na s. 649 uvádějí autoři Příruční mluvnice češtiny následující vzájemné slovosledné pořadí příklonek a příslušné příklonky doprovázejí typickými příklady:

Příklonky stálé:

- spojka *-li*
- tvary *být*, morfémy (morfy) *bych, bys* apod.
- zvratné *se, ses* a *si, sis*
- jednoduché tvary osobních zájmen (etický dativ, dativ, akuzativ, genitiv): *mu*, *ho*...
- kondicionálové *byl*

Příklonky nestálé:

- zájmeno *to*
- adverbiální výrazy *tu, tam, ted'*
- osobní zájmena s předložkou: *s ním*
- modifikační částice: *vlastně* apod.

Na s. 648 se praví, že „poměrně přesně je určeno vzájemné pořadí jednotlivých (zvláště stálých) **příklonek**, pokud se jich ve výpovědi objeví víc“. Dále je na s. 650 uvedeno důležité, ač všeobecně známé tvrzení, že „příklonka stojí za prvním přízvučným slovem, větným členem apod. [...] nebo (přízvučným) spojovacím výrazem“ a následují příklady. Já doplňuji obecně známý fakt, že nejen „příklonka stojí za prvním...“, ale platí to o celém příklonkovém shluku. Podívejme se tedy na postavení především příklonek stálých ve velkých korpusech češtiny. Jelikož v krátkém pojednání nemohu postihnout celou problematiku vzájemného slovosledu všech stálých příklonek, omezím se jen na některé frekvenčně nejvýznamnější příklonky, a sice na tyto:

stálé – ty nejsou v žádné ze svých funkcí a v žádném slovosledném postavení přízvučné: *bych, bys, by, bychom, byste* a obecněčeské *bysme*; *ho*, *-li*, *mi, mu*, *se* (jakožto reflexivum, nikoli vokalizovaná předložka!)*, ses, si* (jakožto reflexivum, nikoli slovesný tvar!)*, sis, tě, ti* (jako osobní zájmeno);

nestálé – ty mohou být v některé ze svých funkcí a v určitém slovosledném postavení přízvučné:
je (jakožto zájmenný i slovesný tvar), *ji, jí, jich, jim, jsem, jsi, jsme, jste, mě, mne, mně, sem* (jakožto obecněčeský slovesný tvar, nikoli adverbium!), *si* (jakožto obecněčeský slovesný tvar, nikoli reflexivum!), *sme, ste*.

Uvedenými nestálými příklonkami se budu zabývat zejména v takové funkci a takovém slovosledném postavení, kde nenesou přízvuk, a chovají se tedy jako příklonky stálé.

(a) Spojka *-li*

Spojka-*li* je v korpusech až na krajně nečetné výjimky (viz níže) připojena vždy těsně k předcházejícímu přízvučnému slovesnému tvaru nebo k některým dalším slovům. V této souvislosti je slovesným tvarem v korpusech výhradně:

- -lové příčestí
- prézentní/futurální slovesný tvar

Spojka *-li* se v korpusech dále připojuje především k těmto slovům:

- *ne* (*ne-li*)
- *málo* (*málo-li*)
- *mnoho* (*mnoho-li*)
- *moc* (*moc-li*)
- *možno* (*možno-li*)
- *nebo* (*nebo-li*)
- *než* (*než-li*)
- *pak* (*pak-li*)
- *prý* (*prý-li*)
- *zda* (*zda-li*)
- *zrovna* (*zrovna-li*)
- *neřku* (*neřku-li*)

To tedy znamená, že spojku *-li* opravdu – ve shodě s mluvnicemi – nepředchází v příklonkovém shluku žádná stálá příklonka. V dalším budu bez přílišné újmy na obecnosti předpokládat, že před *-li* předchází jen slovesný tvar.

(b) Tvary lexému *být*: *jsem, jsi, jsme, jste, sem, si, sme, ste;* tvary kondicionálu *bych, bys, by, bychom, bysme, byste*

Korpusy potvrzují, že pokud se v klauzi uvozené konfigurací

slovesný_tvar - *li*

kde *slovesný_tvar* je -lové příčestí nebo prézentní/futurální tvar, nachází některý z uvedených tvarů slovesa *být* nebo morf(ém) kondicionálu, nachází se těsně za *-li*:

(1) *Chtěl-li* **jsem**, aby si vzal potápěčský...
(2) Neboť *byli-li* **jsme** generace či aspoň...
(3) *Učinil-li* **by** to před měsícem
(4) bude samozřejmě nemožné, *měly-li* **by** se za příklad brát...

V tomto případě (pochopitelně) bez výjimky platí, že *slovesný_tvar* je nutně -lové příčestí, které s prézentním tvarem slovesa *být* tvoří minulý čas 1. nebo 2. osoby a s kondicionálovým morfémem *bych, bys, ...* tvoří přítomný nebo minulý kondicionál všech osob, jmenných rodů a čísel.

Obecně, pokud se v příklonkovém shluku vyskytne tvar kondicionálu i prézentní tvar slovesa *být*, následuje tento prézentní tvar bezprostředně za kondicionálovým morfémem, srov. nespisovné:

(5) Chtěl-li **by jsem** to udělat.
(6) Pak **by jsi** neměl co na práci.

(c) Reflexivní *se, si, ses, sis*

(c1) Reflexivní *se*

Podle Příruční mluvnice češtiny by po reflexivním *se* neměla následovat spojka *-li*, prézentní příklonné tvary slovesa *být* ani kondicionálové morfémy. Korpusy vyjevují, že: jestliže po reflexivním *se* následuje prézentní tvar slovesa *být*, tj. *jsem, jsi, jsme, jste, sem, si, sme, ste*, není reflexivum volným morfémem nějakého slovesa následujícího (ne nutně bezprostředně) za tvarem slovesa *být*:

(7) *Pak **se jsem** opravdu velmi snažil, abych to místo dostal.

Tato věta je negramatická, reflexivum *se* nijak nesouvisí se slovesným tvarem *snažil*. Reflexivum *se* se ovšem může objevit těsně před prézentním tvarem slovesa *být*, pak je však nutně volným morfémem bezprostředně předchozího slova, které reflexivizuje. Repertoár takovýchto slov je však omezený, může jím být pouze jedna z těchto možností:

- infinitiv
- přechodník
- deverbativní adjektivum utvořené od přítomného nebo minulého přechodníku
- deverbativní substantivum nebo deadjektivní substantivum utvořené od deverbativního adjektiva

Uvedu některé příklady:

Infinitiv

(8) *Mýlit **se*** | **jsem** opravdu **nechtěl**.
(9) Jejich neschopnost *domluvit **se*** | **jsi** jistě **považoval** za hrubou politickou chybu.
(10) *Snažit **se*** | **jste** v tomto případě **pokládali** za zbytečné.
(11) Zatímco *omlouvat **se*** | **jsme** nějak rychle **zapomněli**.

Přechodník

(12) *Potácejíce **se*** | ***jsme*** **vykročili** do mého pokoje.
(13) *Utíraje **se*** | **jsi** právě **vycházel** z koupele.

Deverbativní adjektivum/substantivum

(14) že dospělých *učících **se*** | **jsme** měli ve vyspělých průmyslových zemích dostatek

Deverbativní substantivum

(15) Na druhou stranu šance našich juniorů na *prosazení **se*** | **jsme** pokládali za velké
(16) *Přizpůsobení **se*** | **jsem** se snažil minimalizovat.
(17) *Zřeknutí **se*** | **jsem** pokládal za zradu.

Poznámka. Jestliže po reflexivním *se* následuje slovesný tvar *je*, nastávají tyto případy:

1. Reflexivum není volným morfémem nějakého slovesa následujícího (ne nutně bezprostředně) za tvarem slovesa *být*. Takové *se* je v tomto případě volným morfémem bezprostředně předcházejícího slovesa, které reflexivizuje. Tímto slovem je opět pouze jedna z těchto možností:

– infinitiv
– deverbativní adjektivum utvořené od přítomného nebo minulého přechodníku
– deverbativní substantivum nebo deadjektivní substantivum utvořené od deverbativního adjektiva

Uvedu nyní některé příklady:

Infinitiv

(18) *Mýlit **se*** | **je** lidské.
(19) *Ukázat **se*** | **je** to pravé slovo.
(20) Jejich neschopnost *domluvit **se*** | **je** v této situaci trestuhodná.

Deverbativní adjektivum/substantivum

(21) Tím, že *omlouvajícími **se*** | **je**(~~Pron~~, Verb) současné národní či jiné společenství, se vytvářejí...

Deverbativní substantivum

(22) *Sdružování **se*** | **je** totiž logickým důsledkem...
(23) Na *vzdání **se*** | **je** času vždycky dost

2. Na rozdíl od předchozího případu slovesný tvar *je* však také připouští, aby reflexivum *se* bylo volným morfémem slovesa následujícího (ne nutně bezprostředně) za tvarem *je* (reflexivum i sloveso, jemuž reflexivum přísluší, jsou uvedeny tučně):

(24) a věřte, že ***se*** **je** na co *těšit*
(25) a při počtu šesti lidí na pódiu ***se*** **je** na co *dívat*
(26) Hlavně ***se*** **je** třeba dobře *vyspat*
(27) že je to něco, čemu ***se*** **je** nutno *podřídit* v manželství
(28) v otázce restituce majetku židovské církve ***se*** **je** koalice schopna *dohodnout* mnohem lépe než v případě majetku církve katolické

Polohu *se* ve větách (26)–(28) pokládám za silně periferní.

Mimo tyto případy jsem našel v korpusech i následující zvláštnosti:

(29) Došlo mi, že **se je** výjimkou
(30) že **se je** výjimkou a současně **se je** nadále normálním člověkem
(31) Básníkem **se je**, a ne bývá
(32) Načež se zeptalo, kolik **se je** dlužným
(33) ... že **se je** poezií, anebo že **se je** básníkem?

Upozorňuji, že tyto doklady jsou vždy překlady z polštiny. Takovéto užití reflexiva *se* je buď vědomý překladatelský záměr, nebo neobratnost překladatele.

V ostatních případech není slovo *je* slovesný tvar, nýbrž tvar osobního zájmena a jako takové by se mělo disambiguovat:

(34) nebyl líný *naučit **se*** **je**(Pron,~~Verb~~) na stará kolena
(35) a jak ***se*** **je**(Pron, ~~Verb~~) odtud nyní „někdo“ *pokouší*
(36) že ***se*** **je**(Pron, ~~Verb~~) Clintonova administrativa *rozhodla* loni zrušit
(37) *Nepokoušejte **se*** **je**(Pron, ~~Verb~~) překonat tím, že
(38) *nepodaří*-li ***se*** **je**(Pron, ~~Verb~~) celosvětově zvládnout

(c2) Reflexivní *si*

Reflexivum *si* se chová obdobně jako *se* až na jisté odlišnosti ve vztahu k prézentním tvarům slovesa *být*. Reflexivum *si* kanonicky následuje za prézentními tvary slovesa *být*, které spoluvytvářejí minulý čas:

(39) Nedávno **jsem *si předsevzal***...
(40) ***Uvědomil*** **jsem *si*** moc dobře...

Opačný slovosled je mimo případy studované výše u reflexiva *se* možný jen v případě, že prézentní tvar slovesa *být* není pomocným tvarem pro tvoření minulého času, tj. není to příklonka. Mimo standardní:

(41) **Jsem *si*** vědom, že ...
(42) **Jsi *si*** jist, že

je podle rešerší v korpusech možné i:

(43) Opravdu ***si*** **jsem** vědom, že ...

Zcela běžné je to u tvaru *je*, neboť vedle standardního:

(44) **Je *si*** vědom, že

(45) **Je *si*** jist, že

Máme i:

(46) On ***si*** **je** vědom/jist, že

(c3) Reflexiva *ses, sis*

Na základě rešerší v korpusech lze konstatovat, že před tvary *ses* a *sis* v příklonkovém shluku předcházejí příklonky *-li* a *by*, nikdy za nimi nenásledují. Tvary *ses* a *sis* se nacházejí těsně za kondicionálovým morfémem *by*.

(47) Třeba **by *ses*** dokázal *zabít*

(48) *Nevěděla* **by *sis*** s nimi jinak rady

(d) Osobní zájmena a reflexiva se/si/ses/sis

Podle Příruční mluvnice češtiny (Karlík et al., s. 649) následují po reflexivním *se/ses/si/sis* v příklonkovém shluku osobní zájmena. Zkoumal jsem tato zájmena:

dativní: *mi, mně, mu, jí, jim, ti*

akuzativní: *ho, je, ji, mě, mne, tě*

genitivní: *ho, jich, mě, mne, tě*

Pokud osobní zájmeno předchází před reflexivem *se/ses/si/sis*, pak nastávají tyto tři možnosti:

(d1) Osobní zájmeno syntakticky nesouvisí s reflexivním nebo reflexivizujícím slovesem:

(49) Viděl **ho** | ***se*** *brodit* po pás ve vodě.

(50) Nařídil **mu** | ***se*** *převléci*.

(51) Přikázal **mu** | ***si*** *obout* boty.

(52) Donutil **je** | ***se*** pořádně *oholit*.

(53) Najít **ji** | ***se*** mu *nezdálo* vůbec těžké

(54) Najít **ji** | ***ses*** usilovně *snažil*

Je zřejmé, že osobní zájmena tu jsou předměty předcházejících sloves a reflexivum souvisí s jiným slovesem, které je (ne nutně bezprostředně) následuje.

(d2) Osobní zájmeno není příklonka (to ovšem není možné u stálých příklonek):

(55) Právě **ji *si*** ale stárnoucí herec a režisér ***zavolal***.

(56) Ale **mě *si*** nikdo ***nevšímal***.

Tučná osobní zájmena **ji**, **mě** tu nejsou příklonky, jsou přízvučná.

(d3) Osobní zájmeno je součástí předložkové skupiny (to ovšem není možné u stálých příklonek):

(57) Před svého milého postavila pivo a **na mě** ***se*** usmála.

Osobní zájmeno **mě** se nachází v předložkové skupině **na mě**.

Zvláštní případ

Výjimku v postavení dativních osobních zájmen ve vztahu k reflexivům tvoří **etický dativ**, který může předcházet v jednom příklonkovém shluku před reflexivem, srovnej:

(58a) Úředník ***se*** **ti** to *snažil* doručit.
(58b) *Úředník **ti** ***se*** to *snažil* doručit.
(59) Já **ti**(etický dat.) ***se*** *snažil* a ono to stejně nevyšlo.

Etický dativ však může předcházet dokonce i prézentní příklonkový tvar slovesa *být*, srovnej:

(60a) Já ***jsem*** **ti** *neumožnil* pracovat v týmu.
(60b) *Já **ti** ***jsem*** *neumožnil* pracovat v týmu.
(61) Já **ti**(etický dat.) ***jsem*** se začal *dívat* za sebe a vtom...

(e) Osobní zájmena a vztah pádů

Zkoumáním tří pádů osobních zájmen, tj. genitivu, dativu a akuzativu, jsem si v korpusech ověřil známou skutečnost, že *dativní zájmena v příklonkovém shluku předcházejí před osobními zájmeny v genitivu a akuzativu*. Pokud osobní zájmeno v dativu následuje osobní zájmeno v genitivu či akuzativu, pak platí obdobné varianty jako (d1)–(d3) uvedené výše:

(e1) Osobní zájmeno v dativu syntakticky nerozvíjí stejné sloveso jako osobní zájmeno v genitivu/akuzativu:

(62) *Vyrobit* ***ji*** | ***mu*** *trvalo* jediný měsíc.
(63) *Líbat* ***tě*** | ***mi*** *připadá* úplně skvělý.

Ve větě (62) je *mu* předmětem slovesa *trvalo*, *ji* však předmětem slovesa *vyrobit*, a tak zájmena *mu* a *ji* rozvíjejí různá slovesa, přičemž ovšem patří do téhož příklonkového shluku. Obdobně je tomu ve větě (63).

(e2) Jedno z osobních zájmen není příklonka (to ovšem není možné u stálých příklonek):

(64) A **mě**(acc) **mu**(dat) sultán daroval teprve včera
(65) A **mě**(acc) **mu**(dat) vždy kladli za vzor
(66) Ale **je**(acc) **mu**(dat) představili už včera.
(67) ... aby dýku odepjal a podal **ji**(acc) **mně**(dat) nebo komukoli z velmožů

Zájmenné tvary **mě**, **je** v akuzativu a **mně** v dativu nejsou v těchto větách příklonky.

Existují nicméně případy, které pravidlo o slovosledné přednosti dativních zájmenných příklonek porušují. Genitivní a akuzativní osobní zájmena jsou v následujících příkladech příklonky, nenesou přízvuk. Srov:

(68) A přitom dodavatelé dvora **jich**(gen) **jí**(dat) denně dodávali tucty.

(69) Dal **ji**(acc) **jí**(dat) někdo?

(70) Tu přidáme rodičce do vany nebo **ji**(acc) **jí**(dat) vmasírujeme na záda.

(71) Nemohl jsem pochopit, že mě odtrhává od mé dcery, že **mne**(acc) **jí**(dat) bere a ničí náš život.

(72) Chtěl jsem **tě**(acc) **jí**(dat) představit.

(73) Já **tě**(acc) **jim**(dat) nedám.

(74) Kdo **ho**(acc) **jim**(dat) naučil?

Tučná zájmena v genitivu, resp. v akuzativu, jsou v uvedených větách příklonky, a přece předcházejí v příklonkovém shluku před nestálými dativními příklonkami **jí**, resp. **jim**.

(e3) Osobní zájmeno je součástí předložkové skupiny (to ovšem není možné u stálých příklonek)

(75) Na **mě**(prep-acc) **mu**(dat) poslali zatykač.

(76) Ale ode **mě**(prep-gen) **mu**(dat) nebezpečí nehrozí.

Osobní zájmeno **mě** se nachází v předložkové skupině **na mě**. Podobně je tomu s předložkovou skupinou **ode mě** ve větě (76).

Ve vztahu genitiv vs. akuzativ se zdá, že genitiv předchází před akuzativem:

(77) Pak **jí**(gen) **ho**(acc) zbavili.

(78) ?Pak **ho**(acc) **jí**(gen) zbavili.

I v souvislosti se vzájemným postavením příklonných zájmen je zvláštním případem etický dativ, který může následovat za příklonným zájmenem v akuzativu:

(79) Přišel mně zabavit slepice a **mě**(acc) **ti**(etický dat.) popadl vztek.

Srovnej:

(80a) Pak **ti**(dat) **ho**(acc) prodal.

(80b) *Pak **ho**(acc) **ti**(dat) prodal.

Možné jsou i případy těsného sousedství dvou příklonných zájmen v témže pádě, kdy jsou patrně možné obě slovosledné varianty:

(81a) A přitom dodavatelé dvora **jich**(gen) **jí**(gen) denně *dodávali* tucty.

(81b) A přitom dodavatelé dvora **jí**(gen) **jich**(gen) denně *dodávali* tucty.

(82a) Otec **ji**(acc) **ho**(acc) naučil.

(82b) ?Otec **ho**(acc) **ji**(acc) naučil.

Zájmeno *jí* ve větách (81a) a (81b) může mít jistě i genitivní interpretaci.

3 Postavení příklonek a automatická slovnědruhová a morfologická disambiguace češtiny

Uvedených zákonitostí ve slovosledu příklonek lze s výhodou využít k formulaci několika jednoduchých pravidel pro automatickou morfologickou disambiguaci. Povšimněme si těchto víceznačných slovních tvarů:

je (sloveso | zájmeno)
se (reflexivum | předložka)
sem (adverbium | sloveso)
si (reflexivum | sloveso)
ti (osobní zájmeno | ukazovací zájmeno)

a těchto bigramů:

(a) *je se*
(b) *je si*
(c) *ho se*
(d) *ti se*
(e) *ti si*
(f) *ji/ho/mě/mne ti*
(g) *ji se*

(a) bigram *je se*

Konstatuji, že konfigurace ***je***(pron,~~verb~~) ***se***(refl,~~prep~~) je možná jen za těchto podmínek:

Nestojí-li slovní tvar **je** hned na začátku věty, je slovním tvarem předcházejícím tvar ***je*** buď sloveso, jehož je tvar ***je*** předmětem, nebo zdůrazňovací slovo, které znemožňuje příklonkovou interpretaci tvaru ***je***. Za reflexivem ***se*** musí následovat v téže klauzi alespoň jedno sloveso.

(83) Ani **je *se*** nepodařilo *obelstít.*

V opačném případě je tvar ***je*** sloveso nebo tvar ***se*** je předložka.

(b) bigram *je si*

Konfigurace ***je***(pron,~~verb~~) ***si***(refl,~~prep~~) je možná jen za těchto podmínek:

Nestojí-li slovní tvar **je** hned na začátku věty, je slovním tvarem předcházejícím tvar ***je*** buď sloveso, jehož je tvar ***je*** předmětem, nebo zdůrazňovací slovo, které znemožňuje příklonkovou interpretaci tvaru ***je***. Za reflexivem ***se*** musí následovat v téže klauzi alespoň jedno sloveso.

(84) Odmítnout **je *si*** nikdo *nedovolil.*

(c) bigram *ho se*

Konfigurace ***ho se***(refl,~~prep~~) je možná jen za těchto podmínek:

Slovním tvarem bezprostředně předcházejícím tvar ***ho*** je sloveso, jehož je tvar ***ho*** předmětem. Za reflexivem *se* musí následovat v téže klauzi alespoň jedno sloveso.

(85) Sepsat ***ho se***(refl,~~prep~~) *pokoušel* celý víkend.

(d) bigram *ti se*

Konfigurace ***ti***(pronpers,~~prondem~~) ***se***(refl,~~prep~~) je možná jen za těchto podmínek:

Nemá-li tvar ***ti*** funkci etického dativu (ten se ovšem stanoví velmi obtížně!), je slovním tvarem bezprostředně předcházejícím tvar ***ti*** sloveso, jehož je tvar ***ti*** předmětem.

Tvar ***ti*** je tedy osobní zájmeno v dat. sg., nikoli demonstrativum mask. anim. nom. pl. lemmatu *ten*. Za reflexivem ***se*** musí následovat v téže klauzi alespoň jedno sloveso.

(86) Vyhovět ***ti***(pronpers,~~prondem~~) ***se***(refl,~~prep~~) mi nikdy *nedaří*.

V opačném případě je tvar ***ti*** demonstrativním zájmenem:

(87) A ***ti***(~~pronpers~~,prondem) ***se***(refl,~~prep~~) *obávali* o vlastní bezpečnost

nebo tvar *se* je prepozice.

(e) bigram *ti si*

Konfigurace ***ti***(pronpers,~~prondem~~) ***si***(refl,~~verb~~) je možná jen za těchto podmínek:

Nemá-li tvar ***ti*** funkci etického dativu, je slovním tvarem bezprostředně předcházejícím tvar ***ti*** sloveso, jehož je tvar ***ti*** předmětem. Tvar ***ti*** je tedy osobní zájmeno v dat. sg., nikoli demonstrativum mask. anim. nom. pl. lemmatu *ten*. Za reflexivem ***si*** musí následovat v téže klauzi alespoň jedno sloveso.

(88) *Nařídil* ***ti***(~~pronpers~~,prondem) ***si***(refl,~~prep~~) *připravit věci*.

V opačném případě je tvar ***ti*** demonstrativním zájmenem:

(89) *A* ***ti***(~~pronpers~~,prondem) ***si***(refl,~~prep~~) s broukem efektivně *poradí*.

(f) bigram *ji/ho/mě/mne ti*

Konfigurace ***ji/ho/mě/mne***(pron) ***ti***(pronpers,~~prondem~~) je možná jen za těchto podmínek:

Nestojí-li slovní tvar ***ji/mě/mne*** hned na začátku věty, je slovním tvarem předcházejícím tento tvar sloveso, jehož je tvar ***ji/ho/mě/mne*** předmětem, nebo předložka (jen pro tvary ***mě/mne***). Za tvarem ***ti*** musí následovat v téže klauzi alespoň jedno sloveso.

(90) Pověřil ***ho ti***(pronpers,~~prondem~~) vynadat.

(g) bigram *ji se*

Konfigurace ***ji***(pron) ***se***(refl,~~prep~~) je možná jen za těchto podmínek:

Nestojí-li slovní tvar ***ji*** hned na začátku věty, je slovním tvarem bezprostředně předcházejícím tvar ***ji*** sloveso, jehož je tvar ***ji*** předmětem, nebo zdůrazňovací slovo, které znemožňuje příklonkovou interpretaci tvaru ***ji***. Za reflexivem ***se*** musí následovat v téže klauzi alespoň jedno sloveso.

(91) Získat ***ji se*** po všech nepříjemnostech nakonec *pokoušel* bezúspěšně.

V opačném případě je tvar *se* předložka.

4 Shrnutí

Zkoumání rozsáhlých korpusů češtiny ukázalo, že popis vzájemného postavení příklonek v těch současných gramatikách, jež se příklonkami zabývají, a dalších speciálních pojednáních v zásadě odpovídá jazykovému úzu odrážejícímu se v současných korpusech češtiny. V tomto příspěvku jsem se snažil zachytit jen některé hlavní zákonitosti vzájemného postavení nejfrekventovanějších příklonek ve velkých korpusech. Naznačil jsem také možné využití zjištěných zákonitostí pro praktický účel: slovnědruhovou a morfologickou disambiguaci češtiny. Dlužno dodat, že většina z naznačených úvah byla

již prakticky implementována v systému disambiguace češtiny založené na pravidlech (srov. např. Petkevič 2006, Čermák et al. 2005). Jak však ukázaly některé zvláštní příklady, jež se vymykaly z dosud platných a běžně přijímaných pravidel, tato problematika si zaslouží zevrubný popis.

Reference

1. Avgustinova, T. & K. Oliva (1995): The Position of Sentential Clitics in the Czech Clause. In: *Wiener Slavistischer Jahrbuch*, vol. 41. Verlag der Österreichischen Akademie der Wissenschaften, Wien, s. 21–42.
2. Čermák F. & V. Petkevič (2005): Linguistically Motivated Tagging as a Base for a Corpus-Based Grammar. In: Pernilla Danielsson and Martijn Wagenmakers (eds.), Proceedings of Corpus Linguistics 2005. The Corpus Linguistics Conference Series. At http://www.corpus.bham.ac.uk/PCLC/#grammar, Birmingham.
3. Český národní korpus (2000, 2005, 2006, 2007). Korpusy SYN, SYN2005, SYN2000, SYNEK. Ústav Českého národního korpusu FFUK. http://ucnk.ff.cuni.cz.
4. Karlík P., M. Nekula & Z. Rusínová (eds.) (1995): *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny, Praha.
5. Oliva K. (2001): Některé aspekty komplexity českého slovního nepořádku. In Hladká Z. & P. Karlík (eds.) (2001): *Čeština – univerzália a specifika 3*, Brno. Proceedings of the conference in Šlapanice near Brno, 22–24 November 1999, s. 163–172.
6. Petkevič V. (2006): Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In: M. Šimková (ed.), *Insight into the Slovak and Czech Corpus Linguistics* (Šimková M. ed.). Veda (Publishing House of the Slovak Academy of Sciences & Ludovít Štúr Institute of Linguistics of the Slovak Academy of Sciences), Bratislava, s. 26–44.
7. Rosen Alexandr (2001): *A constraint-based approach to dependency syntax applied to some issues of Czech word order*. Disertační práce. Univerzita Karlova v Praze, Filozofická fakulta.
8. Toman J. (2001): Prosodické spekulace o klitikách v nekanonických pozicích. In Hladká Z. & P. Karlík (eds.) (2001): *Čeština – univerzália a specifika 3*, Brno.Proceedings of the conference in Šlapanice near Brno, 22–24 November 1999, s. 73–79.
9. Uhlířová L. (2001): Klitika jako kategorie diskusní. In Hladká Z. & P. Karlík (eds.) (2001): *Čeština – univerzália a specifika 3*, Brno.Proceedings of the conference in Šlapanice near Brno, 22–24 November 1999, s. 29–35.

Between Chaos and Structure: Interpreting Lexical Data through a Theoretical Lens

James Pustejovsky and Anna Rumshisky

Dept. of Computer Science, Brandeis University
Waltham, MA USA {jamesp,arum}@cs.brandeis.edu

Abstract. In this paper, we explore the inherent tension between corpus data and linguistic theory that aims to model it, with particular reference to the dynamic and variable nature of the lexicon. We explore the process through which modeling of the data is accomplished, presenting itself as a sequence of conflicting stages of discovery. First-stage data analysis informs the model, whereas the seeming chaos of organic data inevitably violates our theoretical assumptions. But in the end, it is restrictions apparent in the data that call for postulating structure within a revised theoretical model. We show the complete cycle using two case studies and discuss the implications.

1 Introduction

This paper is an attempt to demonstrate both the theoretical significance of data and the empirical significance of theory. In short, it is an essay on the relationship between data and theory in linguistics. We begin by examining the role theory plays in the analysis of language.

Within analytic linguistics, our initial assumptions on what structures should exist in the data provide us with predictive force and guide us through an often muddled and contradictory set of facts requiring analysis. This initial stage of investigation, what we could call first-level data analysis, uses phenomenological data that are constructed by matching words to expressions predicted by analytic grammar. We refer to these as *synthetic data*. When real data do not fit, we add new structure or, just as often, we idealize such data away. In fact, it could be argued that no data analysis is ever performed without some theoretical bias. This approach has been the operative standard in most language analysis since Aristotle. Except of course, within corpus linguistics, to which we turn next.

Contrary to analytic linguistics, corpus linguists and lexicographers have long stressed the role that extensive analysis of text in use plays in any language description [1,2,3]. Such work emphasizes the importance of looking at the data without theoretical pruning (cf. [4,5,6,7], among others) The basis of this approach is the examination of *organic data* (as opposed to synthetic) in order to form hypotheses regarding language and linguistic behavior. Lexicographic studies of concordancing and collocations have long been used as a means for examining the data [8,9].

In this paper, we use a corpus-driven approach to test a theoretically motivated first-level data analysis of two linguistic phenomena. We will see that theory predicts behavior that is not attested, and behavior exists that is not predicted by theory. These data will be

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 197–214, 2009.

used to inform and update the theory, and in some cases modify or drop theoretical assumptions. The resulting process is an interplay of data analysis and theoretical description.

2 Theoretical Preliminaries

For this exercise in how theory and data interact, we adopt the model of Generative Lexicon, a theory of linguistic semantics which focuses on the distributed nature of compositionality in natural language [10]. Unlike purely verb-based approaches to compositionality, Generative Lexicon (henceforth, GL) attempts to spread the semantic load across all constituents of the utterance. Overall, GL is concerned with explaining the creative use of language; we consider the lexicon to be the key repository holding much of the information underlying this phenomenon. More specifically, however, it is the notion of a constantly evolving lexicon that GL attempts to emulate; this is in contrast to currently prevalent views of static lexicon design, where the set of contexts licensing the use of words is determined in advance, and there are no formal mechanisms offered for expanding this set.

Traditionally, the organization of lexicons in both theoretical linguistics and natural language processing systems assumes that word meaning can be exhaustively defined by an enumerable set of senses per word. Lexicons, to date, generally tend to follow this organization. As a result, whenever natural language interpretation tasks face the problem of lexical ambiguity, a particular approach to disambiguation is warranted. The system attempts to select the most appropriate 'definition' available under the lexical entry for any given word; the selection process is driven by matching sense characterizations against contextual factors. One disadvantage of such a design follows from the need to specify, ahead of time, all the contexts in which a word might appear; failure to do so results in incomplete coverage. Furthermore, dictionaries and lexicons currently are of a distinctly static nature: the division into separate word senses not only precludes permeability; it also fails to account for the creative use of words in novel contexts.

GL attempts to overcome these problems, both in terms of the expressiveness of notation and the kinds of interpretive operations the theory is capable of supporting. Rather than taking a 'snapshot' of language at any moment of time and freezing it into lists of word sense specifications, the model of the lexicon proposed here does not preclude extensibility: it is open-ended in nature and accounts for the novel, creative, uses of words in a variety of contexts by positing procedures for generating semantic expressions for words on the basis of particular contexts. To accomplish this, however, entails making some changes in the formal rules of representation and composition. Perhaps the most controversial aspect of GL has been the manner in which lexically encoded knowledge is exploited in the construction of interpretations for linguistic utterances. Both lexical items and phrases encode the following four types of information structures:

(1) a. LEXICAL TYPING STRUCTURE: giving an explicit type for a word positioned within a type system for the language;
b. ARGUMENT STRUCTURE: specifying the number and nature of the arguments to a predicate;
c. EVENT STRUCTURE: defining the event type of the expression and any internal

event structure it may have, with subevents;
d. QUALIA STRUCTURE: a structural differentiation of the predicative force for a lexical item.

The qualia structure, inspired by [11] interpretation of the *aitia* of Aristotle, are defined as the modes of explanation associated with a word or phrase in the language, and are defined as follows [12]:

(2) a. FORMAL: the basic category which distinguishes the meaning of a word within a larger domain;
b. CONSTITUTIVE: the relation between an object and its constituent parts;
c. TELIC: the purpose or function of the object, if there is one;
d. AGENTIVE: the factors involved in the object's origins or "coming into being".

The different aspects of lexical meaning listed in (1) and (2) can be packaged together as a set of features, illustrated below, where ARGSTR refers to the argument structure of a predicate and EVENTSTR to the event structure (cf. [13,10])

$$\begin{bmatrix} \alpha \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = x \\ \dots \end{bmatrix} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = e_1 \\ \dots \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \textbf{what } x \textbf{ is made of} \\ \text{FORMAL} = \textbf{what } x \textbf{ is} \\ \text{TELIC} = \textbf{function of } x \\ \text{AGENTIVE} = \textbf{how } x \textbf{ came into being} \end{bmatrix} \end{bmatrix}$$

When certain features (qualia) are present or absent, we can abstract away from the representation, and generalize lexemes as belonging to one of three conceptual categories [14,15].

(3) a. NATURAL TYPES: Natural kind concepts consisting of reference only to Formal and Constitutive qualia roles; e.g., *tiger*, *river*, *rock*.
b. ARTIFACTUAL TYPES: Concepts making reference to Telic (purpose or function), or Agentive (origin); e.g., *knife*, *policeman*, *wine*.
c. COMPLEX TYPES: Concepts integrating reference to the relation between types from the other levels; e.g., *book*, *lunch*, *exam*.[1]

This enriched inventory of types for the language is motivated by the need for semantic expressiveness in lexical description. We also need, however, richer interpretive operations to take advantage of these new structures. Following [15], we argue that there are four ways a predicate can combine with its argument:

(4) a. PURE SELECTION (Type Matching): the type a function requires is directly satisfied by the argument;
b. ACCOMMODATION: the type a function requires is inherited by the argument;
c. TYPE COERCION: the type a function requires is imposed on the argument type. This is accomplished by either:

[1] That is, *book* can refer both to the information contained in the book and to the physical object, *lunch* can refer both to the event and to the food, etc. For an inventory of complex types, see [16]

i. *Exploitation*: taking a part of the argument's type to satisfy the function;
ii. *Introduction*: wrapping the argument with the type required by the function.

These mechanisms will form the theoretical scaffolding with which we will perform our first-level data analysis of the argument selection phenomena in the next section. Natural types (e.g. *lion*, *rock*, *water*) are viewed essentially as atomic from the perspective of selection. Conversely, artifactual (or tensor) types (e.g. *knife*, *beer*, *teacher*) have an asymmetric internal structure consisting of a *head type* that defines the nature of the entity and a *tail* that defines the various generic explanatory causes of the entity of the head type. Head and tail are unified by a type constructor $\otimes$ ('tensor') which introduces a qualia relation to the head type: for example, *beer* = $liquid \otimes_{Telic} drink$. That is, *beer* is a kind of liquid; not all liquids are for drinking, but the very purpose (Telic) of *beer* is that someone should drink it.

Finally, complex types (or dot objects) (e.g. *school*, *book*, *lunch* etc.) are obtained through a complex type-construction operation on natural and artifactual types, which reifies two elements into a new type. Dot objects are to be interpreted as objects with a complex type, not as complex objects. The constituents of a complex type pick up specific, distinct, even incompatible aspects of the object. For instance, *lunch* ($event \bullet food$) picks up both $event$ and $food$ interpretations, *speech* ($event \bullet info$) picks up both $event$ and $info$ interpretations, etc. [17].

Type exploitation occurs when a verb selects only a part of the semantics associated with its arguments. For example, the verb *buy* selects for a physical object, which is only a part of the dot object $phys \bullet info$ in (5) below:

(5) Mary bought a book.

Type introduction is the converse, where a new structure is wrapped around a type in argument position. Consider the verb *read*, which selects for the aforementioned type $phys \bullet info$ in direct object position. When, for example, an informational noun such as *rumour* appears, it is "wrapped" with the additional type information:

(6) Mary read a rumour about you.

That is, this rumour is not just an idea (proposition) but has physical manifestation, by virtue of type introduction coercion.

3 Data informs Analysis: Two Case Studies

3.1 First-level Data Analysis: Formulating Theoretical Predictions by Introspection

When initially modeling a particular linguistic phenomenon or pattern, the typical linguistic assumption is to "idealize" the data using introspective or phenomenological data. This has become de rigueur in theoretical linguistic investigations, and we will refer to this stage as *first-level data analysis*. Corpus-oriented linguists have long criticized this approach as armchair lexicography (cf. [18], Sinclair, Hanks, and others). While they do produce a partial account of the data, reflecting valid observational tendencies, such

approaches tend to give an in-depth account of a limited set of behaviors, and typically leave unaccounted the full range of combinatorial phenomena.

Below, we present two case studies in composition: argument selection; and type coercion. We begin by giving a theoretical account of these phenomena, using synthetic data. We then examine the same phenomena using organic data taken from corpora. Finally, we show how the theoretical model of the data is enriched by accounting for a fuller range of the phenomena.

Case Study 1: Verbs Selecting for Artifactual Entities We begin our investigation with the behavior of verbs that select for artifactual arguments, as defined in the previous section. The theory makes a distinction between natural kinds and non-natural kinds, and this is realized in the types used by the lexicon and the grammar. As a result, verbs will be also be typed as natural and non-natural predicates, depending on what kind of arguments they select for. Hence, *Natural predicates* will be those properties and relations selecting for natural types, while *Artifactual predicates* will select for an Artifactual. This distinguishes the classes of verbs in (7) below.

(7) a. NATURAL PREDICATES: touch, sleep, smile
b. ARTIFACTUAL PREDICATES: fix, repair, break, mend, spoil

These classes are defined by the type assigned to the arguments. For example, the type structure for the Natural predicate *touch* is shown in (8):

(8) $$\left[\begin{array}{l}\textbf{touch} \\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = x : phys \\ \text{ARG2} = y : phys\end{array}\right]\end{array}\right]$$

An Artifactual predicate such as the verb *repair* would be typed as shown in (9).

(9) $$\left[\begin{array}{l}\textbf{repair} \\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = x : human \\ \text{ARG2} = y : phys \otimes_{Telic} \alpha\end{array}\right]\end{array}\right]$$

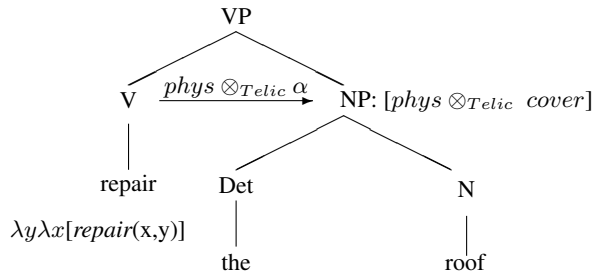
Given these theoretical assumptions, what we expect to encounter as the direct object of artifactual predicates such as *repair*, *fix*, and so forth, are entities that are themselves artifacts.

(10) a. Mary repaired the roof.
b. John fixed the computer.
c. The plumber fixed the sink.
d. The man mended the fence.

What this also predicts is the absence of verb-argument pairings with entities that are not artifactual in some sense. This would appear to be borne out as well, upon initial reflections. You do repair manufactured objects like roofs, cars, and windows; you don't repair natural kinds like boulders, rivers, trees, and pumas.

To illustrate just how this selection is accomplished, consider the sentence in (10a). The verb *repair* (under the intended sense) is typed to select only Artifactual entities as its internal argument. The NP *the roof* satisfies this constraint, as it has a Telic value (i.e., it's an Artifactual), and the verb-argument composition proceeds without incident.

(11)



What this illustrates is how verbs are strictly typed to select specific classes of arguments, in this case an Artifactual as direct object. We consider a somewhat different compositional context in the next section.

Case Study 2: Verbs Selecting for Propositions As our second study, we examine another aspect of GL's theory of selection, namely the phenomenon of *type coercion*. As we saw in the previous section, *Matching* or *Pure Selection* takes place when the type requested by the verb is directly satisfied by the argument. In this case, no type adjustment is needed. *Accommodation* occurs when the selecting type is inherited through the type of the argument. *Coercion* takes place when there is a mismatch (type clash) between the type requested by the verb and the actual type of the argument. This type clash may trigger two kinds of coercion operations, through which the type required by the function is imposed on the argument type. In the first case, *exploitation*, a subcomponent of the argument's type is accessed and exploited, whereas in the second case, *introduction*, the selecting type is richer than the argument type and this last is wrapped with the type required by the function (cf. [15,17]). The reason why two kinds of coercion operation are proposed instead of one is that the information accessed in semantic composition can be differently embedded in a noun's semantics. In both cases, however, coercion is interpreted as a typing adjustment.

To begin, consider the standard selectional behavior of proposition-selecting verbs such as *believe*, *tell*, *know*, and *realize*. This can be seen in the range of data presented below.

(12) a. Mary believes [that the earth is flat].
b. John knows [that the earth is round].
c. John told Mary [that she is an idiot].
d. Mary realizes [that she is mistaken].

Using the typing convention introduced above, the argument structure for a verb such as *believe* would be given as shown in (13).

(13) $$\left[\begin{array}{l}\textbf{believe} \\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = x : human \\ \text{ARG2} = y : info\end{array}\right]\end{array}\right]$$

While these are acceptable constructions, introspection suggests that these predicates also take non-proposition denoting expressions as arguments. For example, consider the sentences in (14) below.

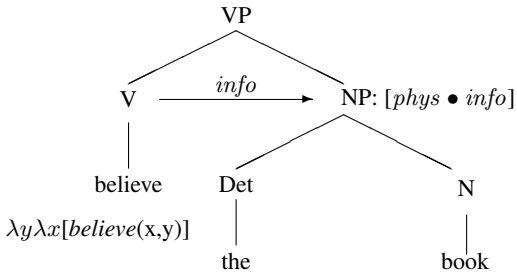
(14) a. Mary believed the book.
b. John told me a lie.
c. The man realized the truth.

Following [10], such expressions are licensed as propositional arguments to these verbs because they are "coerced" into the appropriate type by a rule of type exploitation. Specifically, as mentioned above, nouns such as *book* have double denotation. They are effectively "information containers", and can appear in contexts requiring both physical objects and information, as in (15).

(15) John memorized then burned the book.

The composition involved in a sentence like (14a) is illustrated below, where the informational component of the type structure for *book* is "exploited" to satisfy the type from the predicate.

(16)



This illustrates that predicates may have their selection preferences satisfied by exploiting the substructure associated with an argument. In this case, a propositional interpretation is construed from the type structure of the NP. Indeed, for each proposition-selecting verb in (12), there are well-formed constructions where an NP complement satisfies the propositional typing, as shown in (14).

3.2 Data Challenges Theory

In this section, we turn to naturally occurring (organic) data, equipped with the analytic framework from our first-level data analysis presented above. Using conventional and state-of-the-art tools in corpus analysis, we analyze the actual usage patterns for the predicates discussed above.

We analyze the set of complements for each verb by creating *lexical sets*,[2] a methodology first deployed in [16] and used to examine selectional contexts for complex nominals.[3] We use the Sketch Engine [20] to examine the data from the British National Corpus [21]. The Sketch Engine is a lexicographic tool that lists salient collocates that co-occur with a given target word in the specified grammatical relation. The collocates are sorted by their association score with the target, which uses pointwise mutual information

[2] Cf. [7].
[3] See [19] for more detail.

Table 1. Direct object complements for the *repair*-verbs

repair.v			fix.v			mend.v		
damage	107	42.66	pipe	9	11.83	fence	23	32.78
roof	16	20.27	gutter	4	11.45	shoe	10	19.01
fence	10	18.07	heating	5	9.66	puncture	4	18.91
gutter	5	15.87	car	19	9.43	clothes	11	18.68
ravages	4	15.76	alarm	5	9.13	net	8	18.01
hernia	4	15.61	bike	5	9.11	roof	8	16.99
car	23	15.39	problem	23	8.77	car	14	15.45
shoe	10	15.22	leak	3	8.58	way	20	14.26
leak	5	14.96	light	12	8.49	air-conditioning	2	12.71
building	17	14.02	boiler	3	7.96	damage	6	12.71
crack	6	13.99	roof	5	7.27	hole	5	11.38
wall	14	13.77	motorbike	2	7.19	bridge	4	9.68
fault	7	13.56	fault	4	6.91	heart	5	9.6
puncture	3	13.53	jeep	2	6.79	clock	3	9.45
pipe	7	12.89	door	11	6.65	chair	4	9.36
bridge	8	12.19	chain	4	5.48	wall	5	9.27
road	13	12.19	bulb	2	5.15	chain	3	8.3

between the target and the collocate multiplied by the log of the pair frequency for a given grammatical relation. Additional corpus queries were performed using Manatee, a companion concordancing engine.[4]

In Tables 1, 2, and 3, we give the salient collocates for the verbs presented in the previous section, along with frequencies and association scores for each collocate. [5] We only list the complements that activate the relevant sense of the verb. For example, for the verb *realize*, we show the frequencies for propositional complements (e.g., *mistake, truth, importance, significance, implication, futility, danger, error*) and omit the complements activating the *bring into being* sense (e.g., *potential, ambition, dream, goal, hope, fear, ideal, expectations, vision, objective, plan*, etc.)

Case Study 1 (cont) Our model predicts that the verbs *repair*, *fix*, and *mend* will select artifactual entities in direct object position, making implicit reference to the entity's Telic value. What we see in the actual data is that many of the complements do not refer to artifactual entities at all, such as: *damage*, *puncture*, *hernia*, *hole*, *crack*, *fault*, *problem*, *leak*, and *ravages* (cf. Table 1).

The problem that emerges from these data is that the same sense of each verb is being activated by semantically diverse lexical triggers, many of which are not artifactual objects. This raises the issue of the semantic relationship between these lexical items. These questions have been discussed previously in [22] and [23], and we turn to them with respect to the present case studies in Section 3.3.

[4] See `http://www.textforge.cz/products`

[5] Sketch Engine word sketches for the BNC were manually edited to correct for misparses.

Table 2. Direct object complements for the PROPOSITION/INFO-verbs

believe.v			know.v			realize.v		
luck	73	33.14	answer	389	35.17	mistake	15	20.02
ear	48	22.5	truth	219	30.92	extent	18	19.0
story	72	20.58	name	548	29.03	truth	15	18.7
word	95	19.02	whereabouts	37	24.64	importance	15	16.42
eye	74	15.19	secret	73	22.0	significance	11	16.11
hype	6	14.17	detail	142	17.77	implication	11	15.6
myth	12	14.07	story	141	17.48	futility	3	13.78
truth	19	13.31	meaning	78	16.58	value	17	13.28
lie	10	12.63	fact	159	16.28	danger	7	12.01
tale	13	12.61	reason	137	15.89	error	7	11.87
opposite	7	12.15	score	47	14.83	possibility	8	11.78
tarot	3	12.0	outcome	45	14.53	predicament	3	11.56
nonsense	7	11.6	saying	14	14.29	folly	3	10.09
propaganda	7	11.12	God	77	14.23	limitations	4	9.7
thing	47	9.12	username	7	14.02	strength	4	6.77
woman[6]	41	9.06	difference	105	13.98	need	6	6.07
fortune	8	8.82	feeling	79	13.75	threat	3	5.7
stupidity	3	8.57	word	162	13.74	benefit	4	5.31
rubbish	5	8.01	basics	10	13.53	problem	7	5.17
rumour	5	7.96	rules	99	13.03	advantage	3	5.04
evidence	19	7.81	address	42	12.74	difficulties	3	4.79
promise	7	7.78	password	10	12.4	effects	5	4.68
figures	21	7.78	identity	37	12.38	risk	3	4.68
forecast	5	7.49	joy	23	12.23	power	5	4.21
poll	7	7.48	trick	20	12.18	nature	3	3.7
gospel	4	7.45	place	171	11.88	fact	3	3.27
assurance	6	7.44	date	67	11.26	cost	3	2.94
success	14	7.35	extent	46	11.26			

Case Study 2 (cont) For the next case study, our model predicts that NPs denoting information containers have the appropriate type structure to satisfy proposition-selecting predicates through type exploitation. That is, *the book* can denote a proposition in the sentence

(17) Mary believed the book.

Furthermore, we expect to see proposition-denoting NPs as complements as well. For example, *rumour* denotes a proposition in the sentence

(18) John doesn't believe the rumour.

We see from the data that there are many non-proposition-denoting NPs, varying from verb to verb. For example, for the verb *believe*, we have: *luck*, *eye*, *ear*, *tarot*, *woman*, *success*; for the verb *know*: *name*, *score*, *address*, *rules*, *trick*; for the verb *realize*: *futility*,

Table 3. Direct object and ditransitive obj2 complements for *tell*.

tell.v/direct object			tell.v/ditransitive obj2					
story	1286	52.0	secret	36	22.42	suspicion	4	5.62
truth	600	49.48	name	122	22.21	history	13	5.34
lie	254	45.67	detail	32	12.67	answer	9	5.33
tale	274	42.04	reason	37	11.06	direction	9	5.3
fib	18	30.84	gossip	6	10.4	dream	6	5.17
joke	94	28.85	ordeal	5	9.9	thought	10	5.08
untruth	8	19.08	gist	3	9.61	legend	3	4.92
anecdote	15	17.08	fact	34	9.5	age	13	4.7
difference	108	16.82	whereabouts	4	9.09	outcome	5	4.6
parable	8	12.75	trouble	9	6.98	symptom	4	4.32
fortune	24	12.57	plan	19	6.9	position	14	4.15
news	53	12.13	date	13	6.71	fate	3	4.08
			destination	4	6.54	identity	4	3.91

folly, *threat*, *risk*, *cost*; for the verb *tell*: *history*, *ordeal*, *destination*, *suspicion*, *identity*, etc. The full list of complements, sorted by association score, is given in Tables 2 and 3.[7]

We clearly need to account for how these NPs satisfy the selectional conditions of the predicate, supposing our assumptions regarding the typing of the predicates are correct. Alternatively, we need to rethink the selectional specifications for each verb.

3.3 Theoretical Analysis of Structured Data

Case Study 1 (cont) The first observation from analyzing organic data associated with the selectional behavior of verbs like *fix*, *repair* and *mend* is that there are, in fact, two major selectional clusters, not one. One indeed involves the artifactual entities as predicted by our theoretical assumptions. The other, however, refers to a negative stative or situational description of the artifactual under discussion. Further, we observed that this latter cluster divides systematically into two classes, one a general negative situation, and the other referring to the condition of the artifact, as can be seen in lexical sets in (19), (20), and (21).[8]

(19) *fix.v*
object
a. ARTIFACTUAL: pipe, car, alarm, bike, roof, boiler, lock, engine; heart; light, door, bulb
b. NEGATIVE STATE (condition on the artifact): leak, drip
c. NEGATIVE STATE (general situation): problem, fault

(20) *repair.v*
object
a. ARTIFACTUAL: roof, fence, gutter, car, shoe, fencing, building, wall, pipe, bridge, road; hernia, ligament
b. NEGATIVE STATE (condition on the artifact): damage, ravages, leak, crack, puncture, defect, fracture, pothole, injury
c. NEGATIVE STATE (general situation): rift, problem, fault

[7] For the verb *tell*, we give both direct object complements and NPs from ditransitive constructions, as identified by RASP parser [24].

[8] Semicolon is used to separate semantically diverse elements of each lexical set.

(21) *mend.v*
object
a. ARTIFACTUAL: fence, shoe, clothes, roof, car, air-conditioning, bridge clock, chair, wall, stocking, chain, boat, road, pipe
b. ARTIFACTUAL (extended or metaphoric uses): matter, situation; relationship, marriage, relations
c. NEGATIVE STATE (condition on the artifact): puncture, damage, hole, tear

Assuming that these are all instances of the same sense for each of the verbs, how do we incorporate these observations back into the selectional properties of the verb? First, as mentioned above, there appear to be two negative states selected in many cases:

(22) a. GENERAL NEGATIVE SITUATION: "fix the problem"
b. CONDITIONS OF THE ARTIFACT: "hole in the wall", "dent in the car".

What do these clusters have in common? Does the verb select for either a negative situation or an artifact? The answer is: basically, the verbs select for a negative state of an artifactual.

When the negative relational state is realized, it can either take an artifactual as its object, or leave it implicitly assumed:

(23) a. *repair the puncture / leak*
b. *repair the puncture in the hose / leak in the faucet*

When the artifactual is realized, the negative state is left implicit by default.

(24) a. *repair the hose / faucet*
b. *repair the (puncture in) the hose / (leak in) the faucet*

This suggests that the theoretical description of the selectional properties for the verb *repair* needs modification to reflect behavior witnessed from the organic data. This can be accomplished by positing the negative state as the selected argument of a verb such as *repair*, and the artifactual posited as a *default argument*.

(25)
$$\left[\begin{array}{l}\textbf{repair} \\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = x : human \\ \text{ARG2} = y : neg\_state(z) \\ \text{D-ARG1} = z : phys \otimes_{Telic} \alpha\end{array}\right]\end{array}\right]$$

This has the effect of explaining the lexical set distribution: when the noun denotes a negative state, there is an implicit (default) artifactual quantified in the context. When the artifactual is realized, the negative state interpretation is present in a type of coercion (introduction). Hence, both patterns are accounted for by the lexical structure for the verb along with compositional principles allowing for coercion.

Case Study 2 (cont) From examination of the data on NP-complements to proposition-selecting predicates, we see that type coercions, when they exist, are distributed in very different ways for each verb. Theoretically, this means that the licensing conditions for type coercion must be distinct in each of these cases. Given the theoretical fragment we presented in Section 10, however, there are no mechanisms for explaining this distribution.

In order to understand this behavior better, let us examine the non-coerced complementation patterns of these verbs in corpora. Several subclasses of clausal complements are attested in the BNC for each of these verbs. Namely, we identify the following three complement types:

(26) a. FACTIVE: *know*, *realize*
b. PROPOSITION: *believe*, *tell*
c. INDIRECT QUESTION: *know*, *tell*

We have already encountered the syntactic behavior of propositions in (12). The class of "factives" includes verbs that presuppose the situation denoted by the complement. For example, in (27), the situation denoted by the complement is presupposed as fact.

(27) a. John realized [that he made a mistake].
b. Mary knows [that she won].

The class of "Indirect questions" includes verbs selecting a *wh*-construction that looks like a question, but in fact denotes a value. For example, the verb *know* allows this construction, as does *tell*:

(28) a. Mary knows [what time it is].
b. John knows [how old she is].

(29) a. Mary told John [where she lives].
b. John told me [how old he is].

In order to account for this data, the model must allow each verb to carry a more specific encoding of its complement's type than we had initially assumed, except for the verb *believe*. This suggests the revised argument structures [9] below.

(30) **believe**(ARG1:$human$, ARG2:$prop$)

(31) a. **tell**(ARG1:$human$, ARG2:$info$)
b. **tell**(ARG1:$human$, ARG2:$Ind\_Question$)

(32) a. **know**(ARG1:$human$, ARG2:$factive$)
b. **know**(ARG1:$human$, ARG2:$Ind\_Question$)

(33) **realize**(ARG1:$human$, ARG2:$factive$)

The question is whether these verbs have the same semantic selectional behavior when occurring with NPs as they do with clausal complements. Consider first when an NP can be interpreted as an indirect question. What we see in the corpus is that one set of arguments for the verbs *know* (and *tell*) includes nominals that denote the value of something interpreted as a varying attribute; that is, they can take on or assume the interpretation of an indirect question in the right context. For example, the noun *age* is an attribute of an object with different values, and the noun *time* in this same context can be interpreted as an indirect question.

[9] The feature structure notation is simplified for readability and space considerations.

(34) a. Mary knows the time.
b. John knows her age.

(35) a. Mary told John her address.
b. John told me his age.

This NP construction is usually referred to a "concealed questions" structure. The lexical sets for the verbs *tell* and *know*, organized by most probable semantic type, are shown in (36) and (37) below. The BNC data in these lexical sets was collected using the Sketch Engine, and manually sorted according to the complement type.

(36) *tell.v*
object
a. PROPOSITION: story, truth, lie, tale, joke, anecdote, parable, news, suspicion, secret, tale, details, gossip, fact, legend; dream, thoughts
b. INDIRECT QUESTION: name, whereabouts, destination, age, direction, answer, identity, reason, position, plan, symptoms; outcome, trouble

(37) *know.v*
object
a. FACTIVE: truth, secret, details, story, meaning, fact, reason, outcome, saying
b. INDIRECT QUESTION: answer, score, whereabouts, address, username, password, name; feeling, difference

With the verb *realize*, the data show that NPs complements can also assume a factive interpretation:

(38) John realized his mistake.

But what is interesting is that the majority of the nominals are abstract relational nouns, such as *importance*, *significance*, *futility*, and so forth, as illustrated below.

(39) *realize.v*
object
FACTIVE: importance, significance, extent, implication, futility, value, error, predicament

For the verb *believe*, all nominals are coerced to an interpretation of a proposition, but through different strategies. Those nominals in (40a) either directly denote propositions (e.g., *lie*, *nonsense*) or are complex types that have an information component which can interpreted propositionally (e.g., *bible*, *polls*). The sources in (40b) are construed as denoting a proposition produced by (e.g., *woman*), or coming through (e.g., *ear*) the named source. Finally, the last set is licensed by negative polarity context, and is a state or event; e.g., "He couldn't believe his luck.").

(40) *believe.v*
object
a. PROPOSITION: lie, tale, nonsense, myth, opposite, truth, propaganda, gospel
b. SOURCE: woman, government, bible, polls, military; ear, eye
c. EVENT/STATE: luck, stupidity, hype, success

Note also that the prediction that selectional specifications of *believe* as an information-selecting predicate could be satisfied by any information nominal is not borne out. For instance, the informational component of a complex type $phys \bullet info$ does not seem to encourage the interpretation appropriate for a complement of *believe*. While some nouns

of *phys* • *info* type, such as *letter*, do accept this interpretation, it is so infrequent that it is not attested in roughly 33,000 of occurrences of *believe* in the BNC. Other nouns of *phys* • *info* type, such as *novel*, do not seem to be capable of this interpretation altogether.

This also suggests that different information-selecting predicates in fact require different propositional structure from the complements. For example, *believe* requires the informational noun to allow either a single message interpretation (e.g. *believe the nonsense*) or a source interpretation (e.g. *believe the political blogs*).

This necessity for refinement of selectional specifications is also apparent for other information-selecting predicates, for example, for *write*. In the classic GL interpretation, this verb selects for the artifacts of *phys* • *info* type with Agentive "write" and Telic "read" – that is, they select for objects that are produced by writing and whose purpose is to be read. But consider the nouns in (41) which clearly match this specification, and yet differ in their ability to satisfy the corresponding selectional requirements.

(41) a. John wrote a novel.
b. ?John wrote a dictionary.
(but cf. "You have to love a lexicographer who had the courage, interest, and patience to write an entire dictionary by himself.")
c. ?John wrote a newspaper.
(but cf. "Sixth-grade pupils wrote a newspaper for their parents describing their experiences in different curriculum areas in the classroom."

While (41a) is acceptable without qualification, both (41b) and (41c) require a bit of context to modulate the composition, enhancing the "naturalness" of the expression, in Sinclair's sense [25]. So in fact a more refined specification is needed to explain combinatorial behavior of these nouns, one perhaps taking into account the exact manner in which information carried by each artifact is produced.

4 Concluding Remarks

In this paper, we have examined the contributing roles both corpus-based and model-based linguistics play in constructing an adequate characterization of language usage. By its very design, Generative Lexicon aims to explain the contextual modulations of word meanings in actual data. Therefore, the distributional profile presented by large corpora is not a tension so much as a necessary component to a healthy investigation of the phenomenon, namely, the infinite richness of language. While the generative notion of the *ideal speaker/hearer* of language is a powerful notion, it is empty without application and revision through data. As Sinclair has aptly stated:

> Starved of adequate data, linguistics languished. . . . It became fashionable to look inwards to the mind rather than outwards to society. [1]

References

1. Sinclair, J.: Corpus, Concordance, Collocation. Oxford University Press (1991)

2. Firth, J.R.: A synopsis of linguistic theory 1930-1955. In: Studies in Linguistic Analysis. Oxford: Philological Society (1957) 1–32
3. Hornby, A.S.: A Guide to Patterns and Usage in English. Oxford University Press (1954)
4. Sinclair, J.: Beginning the study of lexis. In Bazell, C.E., Catford, J.C., Halliday, M.A.K., Robins, R.H., eds.: In Memory of J. R. Firth. Longman (1966)
5. Sinclair, J.: Trust the Text. Routledge (2004)
6. Hanks, P.: Linguistic norms and pragmatic explanations, or why lexicographers need prototype theory and vice versa. In Kiefer, F., Kiss, G., Pajzs, J., eds.: Papers in Computational Lexicography: Complex '94. Research Institute for Linguistics, Hungarian Academy of Sciences (1994)
7. Hanks, P.: Contextual dependency and lexical sets. International Journal of Corpus Linguistics **1** (1996)
8. Sinclair, J.: The nature of the evidence. In Sinclair, J., ed.: Looking Up: An Account of COBUILD Project in Lexical Computing. London: Collins (1987)
9. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics **16** (1990) 22–29
10. Pustejovsky, J.: Generative Lexicon. Cambridge (Mass.): MIT Press (1995)
11. Moravcsik, J.M.: Aitia as generative factor in aristotle's philosophy. Dialogue **14** (1975) 622–636
12. Pustejovsky, J.: The generative lexicon. Computational Linguistics **17** (1991)
13. Bouillon, P.: Polymorphie et semantique lexical: le case des adjectifs. PhD dissertation, Paris VII, Paris (1997)
14. Pustejovsky, J.: Type construction and the logic of concepts. In: The Syntax of Word Meaning. Cambridge University Press, Cambridge (2001)
15. Pustejovsky, J.: Type theory and lexical decomposition. Journal of Cognitive Science **6** (2006) 39–76
16. Rumshisky, A., Grinberg, V.A., Pustejovsky, J.: Detecting Selectional Behavior of Complex Types in Text. In Bouillon, P., Danlos, L., Kanzaki, K., eds.: Fourth International Workshop on Generative Approaches to the Lexicon, Paris, France (2007)
17. Asher, N., Pustejovsky, J.: A type composition logic for generative lexicon. Journal of Cognitive Science **6** (2006) 1–38
18. Fillmore, C.: 'corpus linguistics' vs. 'computer-aided arimchair linguistics', Berlin, Mouton de Gruyter (1991)
19. Rumshisky, A., Batiukova, O.: Polysemy in verbs: systematic relations between senses and their effect on annotation. In: COLING Workshop on Human Judgement in Computational Linguistics (HJCL-2008), Manchester, England (2008) submitted.
20. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: The Sketch Engine. Proceedings of Euralex, Lorient, France (2004) 105–116
21. BNC: The British National Corpus. The BNC Consortium, University of Oxford, *http://www.natcorp.ox.ac.uk/* (2000)
22. Rumshisky, A.: Resolving polysemy in verbs: Contextualized distributional approach to argument semantics. Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica (2008) forthcoming.
23. Pustejovsky, J., Jezek, E.: Semantic coercion in language: Beyond distributional analysis. Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica (2008) forthcoming.
24. Briscoe, T., Carroll, J.: Robust accurate statistical annotation of general text. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, May 2002 (2002) 1499–1504
25. Sinclair, J.: Naturalness in language. In J.Aarts, Meijs, W., eds.: Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research. Rodopi (1984)

Constructing High Precision Synonym Sets

Radim Řehůřek

Seznam.cz, a.s.
radim.rehurek@firma.seznam.cz

Abstract. This article deals with the automated construction of word synonyms. Common, human created dictionaries are used to obtain coarse synonym sets. These have excellent recall, but very low precision. An automated technique is presented which improves the precision. Moreover, our approach allows more dictionaries to be used, so that the results improve as more data becomes available.

1 Introduction

The task of automatically identifying synonyms and paraphrases has enjoyed considerable attention in the past. Two fundamental approaches have been established:

1. Those that make use of monolingual information. Typically, this means extraction of synonyms from monolingual dictionaries, as well as methods based on *distributional similarity*. Here, semantically related words are assumed to appear in similar contexts. This leads to elegant, corpus-driven algorithms. However, these algorithms often run into problems with distinguishing synonyms from antonyms, hyponyms, hypernyms etc. Examples of work from this field include [1,2].
2. Those that make use of bilingual resources. An example of work that builds on aligned bilingual corpora to extract paraphrases is [3]. Other works, such as [4], try to combine both approaches in a unified framework.

In this work, we follow the second route and employ bilingual data sources. We make use of translation dictionaries to automatically construct synonym sets. The process of extraction is fully automated, language independent and has an intuitive parameter which controls the trade-off between precision and recall.

2 Proposed algorithm

Firstly, let us define the notation that will be used throughout this paper. Let $L = \{l_0, l_1, l_2, \ldots\}$ be a set of languages and let W_l be the set of words for language l. Let $D_{(l_1,l_2)} \subseteq W_{l_1} \times W_{l_2}$ be a dictionary which maps words from one language to another. As an example, $D_{(english,czech)} \supset$ *{(crimson, karmín), (crimson, karmínový), (crimson, krvavý), (crimson, purpurový), (crimson, rudý), (crimson, zbarvit), (crimson, zrudnout), (crimson, zčervenat), (crimson, červený), . . . }* .

In this notation, a pair of languages l_0, l_1 together with a dictionary $D_{(l_0,l_1)}$ can be seen as a bipartite graph $G_{(l_0,l_1)}$, with words as nodes and the dictionary mapping as edges between the nodes.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 213–219, 2009.

Table 1. Example output of the $basic$ synonym set function, for a Czech-English dictionary.

Word	Synonyms according to the basic synset function
žal	žal, bolet, bolest, bolesti, bída, trápení, politování, nářek, lítost, strast, zármutek, smutek, hoře, bol, problém, zarmoutit, litovat, zalitovat, želet
poezie	poezie, básnit, rým, rýmovat, říkadlo, říkanka, verš
pohledný	pohledný, prospěch, prospěšný, vhodný, příjemný, dobře, laskavý, půvabný, správně, hezký, pravý, milý, užitečný, blaho, užitek, sličný, výborně, důkladný, spolehlivý, čestný, značný, poslušný, slušný, dobrá, dobro, právoplatný, řádný, spořádaný, laskav, hodný, dobrý, pořádný, reprezentační, reprezentativní, vzhledný
zpívat	zpívat, zazpívat

Now let us denote by $basic^n_G(w)$, or the *basic synonym set*, the set of all words which have distance $2n$ from the node w in the bipartite graph G, $w \in l_0$, $n \geq 0$. Note that this set contains nodes strictly from l_0, the same language as w. Examples of $basic^1_{G_{czech,english}}$ for four randomly selected words can be seen in table 1. Looking at these examples, it becomes apparent that while this basic synset function allows us to discover many useful synonyms, the synonym set is quite noisy at the same time. This corresponds to high recall (the $basic$ function covers a large portion of true synonyms) but low precision (many of the words in $basic$ are not real synonyms, especially for $n > 1$).

Unfortunately, this is not caused by our poor choice of the example, or by inadvertent use of a low quality dictionary. Rather, it is a straightforward effect of the abundance of homographs in natural languages — words that are spelled the same, but have different meanings.

To remediate this problem, we extend the $basic$ synonym function. Instead of looking at the graph neighborhood of a single word for one pair of languages, we analyze neighborhoods for multiple languages. More formal notation follows, but intuitively, this corresponds to aggregating information from several dictionaries. In this algorithm, adding more dictionaries never hurts the recall, but may improve the precision.

More formally, let $w \in l_0$ be the word we wish to find synonyms for, and let $D_{(l_0,l_1)}, D_{(l_0,l_2)}, \cdots, D_{(l_0,l_n)}$ be n dictionary mappings between l_0 and n other languages, $n \geq 1$. Note that these are dictionaries between the *target language*, l_0, and n other languages — not a full set of pairwise dictionaries between all languages. Given n dictionaries represented as bipartite graphs $G_{(l_0,l_1)}, \ldots, G_{(l_0,l_n)}$, and two words $w, c \in l_0$, let

$$prevalence_{G_1,G_2,\ldots,G_n}(w,c) = |\{i \mid i \in \{1,\ldots,n\} \wedge c \in basic^1_{G_i}(w)\}|. \quad (1)$$

That is, let $prevalence$ be the number of dictionaries for which the words w and c belong to the same basic synonym set. Then we can define the $synset$ function, parametrized by $quality \in \{1,\ldots,n\}$ to be

$$synset^{quality}_{G_1,G_2,\ldots,G_n}(w) = \{c \in l_0 \mid prevalence(w,c) \geq quality\}. \quad (2)$$

Examples of output of the $synset$ function for the same words as in the previous example can be seen in table 2. The comparison favours the extended $synset$ function in all

Table 2. Synonyms for various quality levels. Results for quality 1 have been omitted for brevity — these correspond to a union of basic synsets like those from table 1, but over six dictionaries instead of a single one.

(a) žal

Quality	Synonyms
2	hoře, úzkost, bída, zármutek, chmura, mrzutost, trápení, skleslost, útrapa, muka, sklíčenost, truchlivost, běda, stesk, politování, tesknota, strast, soužení, bolest, tíseň, bol, smutek, utrpení, neštěstí, rána, žal, lítost
3	zármutek, bolest, trápení, politování, strast, soužení, hoře, tíseň, bol, smutek, utrpení, žal, lítost
4	zármutek, bolest, soužení, hoře, politování, bol, smutek, trápení, žal, lítost
5	zármutek, bolest, soužení, hoře, bol, smutek, trápení, žal, lítost
6	hoře, bol, smutek, soužení, zármutek, žal

(b) poezie

Quality	Synonyms
1	báseň, básnický, básnictví, básnička, dílo, harmonie, krása, múza, okouzlení, parnas, poezie, píseň, písňový, rytmus, rým, rýmovat, tlukot, umění, verš, veršovat, zpěv, řemeslo, říkadlo, říkanka
2	báseň, básnictví, poezie
3	báseň, básnictví, poezie
4	báseň, básnictví, poezie
5	báseň, básnictví, poezie
6	poezie

(c) pohledný

Quality	Synonyms
2	dobrý, hezky, hezký, krásný, laskavý, líbivý, milý, pohledný, pořádný, pěkný, pěkně, roztomilý, sličný, upravený, vzhledný, značný, úhledný, řádný
3	hezký, líbivý, pohledný, pěkný, vzhledný, úhledný
4	hezký, pohledný, vzhledný
5	pohledný
6	pohledný

(d) zpívat

Quality	Synonyms
2	hrát, kokrhat, opěvovat, píseň, pískat, pět, recitovat, skandovat, vrzat, vyzpěvovat, vyzradit, zapět, zazpívat, zpívat, zpěv
3	kokrhat, opěvovat, pět, zazpívat, zpívat
4	opěvovat, zazpívat, zpívat
5	zazpívat, zpívat
6	zpívat

respects — both precision and recall increased considerably. This particular $synset$ function uses six distinct dictionaries between Czech and six other languages, allowing us to construct six synonym sets with increasing precision. Note that we can tweak the output quality in two ways:

Table 3. Coverage statistics for the extended $synset$ function. The results were created using six different dictionaries, between Czech and English, German, Spanish, French, Italian and Russian.

Quality	No. synsets	No. synonyms	Average no. synonyms per synset
1	78,854	2,783,316	36.30
2	48,731	403,643	9.28
3	31,857	164,490	6.16
4	20,878	80,014	4.83
5	12,931	38,771	4.00
6	6,005	14,465	3.41

- by increasing the value of the $quality$ parameter. With the number of dictionaries held fixed, raising the quality increases precision at the cost of lowering recall. The extreme value of $quality = n$ corresponds to the condition that if two words w, c are to be synonyms, they must appear as basic synonyms in all n available dictionaries.
- by increasing the number of dictionaries used. With the $quality$ parameter held fixed, this increases recall while keeping the precision constant.

To obtain the best results, it is therefore desirable to use as many dictionaries as possible. The trade-off between recall and precision, as controlled by the $quality$ parameter, can be tuned to suit the needs of the particular application at hand.

3 Evaluation

The best way to evaluate synonym quality is by observing the benefit it gives to the task that makes use of them. Without such task, evaluation is notoriously difficult.

Since the proposed method is general and could be used in various scenarios, with various data sources and differing quality requirements, we limit our evaluation to statistical comparison of its coverage. For the Czech language, there exists a Dictionary of Czech Synonyms [5], a printed book used by the general public, high school students etc. This compilation of synonyms was created by hand and offers much added value to the user compared to our automatically derived synonyms. For example, the synonyms may be phrases as opposed to single words, different meanings of a word receive their own separate synonym sets, there are manual tags marking colloquial or archaic terms and so on. This dictionary covers about 21,600 different words for a total of 89,000 synonyms. By comparison, our automatic method has a much wider coverage — see table 3 for overall statistics. At quality 2, which is sufficient for the task of synonym suggestion to humans and is thus comparable in purpose, and with six dictionaries, our method offers synonym sets for over 48,000 words, for a total of over 400,000 synonyms.

4 Conclusion

We have presented a straightforward algorithm for synonym extraction. This algorithm is built on human-created data and as such enjoys high precision and intuitive interpretation

of its results. It does not suffer from the problems of distributional approaches of "contextual synonymy", where automated extraction often results in antonyms, loosely related or even unrelated words. Its only drawback is the need of multiple independent translation dictionaries for the target language. This is usually not a problem, as even for minor languages, such as Czech, there exist translation dictionaries to all major languages which serve well for this purpose.

Dedication

The proposed algorithm, devised and tested on data from Seznam.cz, is dedicated to doc. PhDr. Karel Pala, CSc.

References

1. Lin, D.: Automatic retrieval and clustering of similar words (1998)
2. Turney, P.: Mining the web for synonyms: Pmi-ir versus lsa on toefl (2001)
3. Barzilay, R., McKeown, K.R.: Extracting paraphrases from a parallel corpus. In: Proceedings of the ACL/EACL. (2001) 50–57
4. Wu, H., Zhou, M.: Optimizing synonym extraction using monolingual and bilingual resources. In: Proceedings of the second international workshop on Paraphrasing, Morristown, NJ, USA, Association for Computational Linguistics (2003) 72–79
5. Pala, K., Všianský, J.: Slovník českých synonym (Dictionary of Czech Synonyms) (1994)

New version of the Croatian National Corpus

Marko Tadić

University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics
marko.tadic@ffzg.hr

Abstract. This contribution presents the new version (v 2.5) of the Croatian National Corpus (HNK). In the beginning it briefly describes the history of collecting HNK and its first two versions. It continues with describing the differences and novelties introduced in this new version: 1) new text samples that bring the existing corpus structure more to the desired ideal ensemble of text types, genres and topics; 2) lemmatization and full MSD-tagging of the whole corpus. This second update is realized using lemmatizer and MSD-tagger for Croatian described in (Agić et al. 2008, Agić et al. 2009a). It achieves results at the level of state-of-art of taggers for other Slavic languages while in lemmatization it offers some novel solutions in its hybrid approach to disambiguation of lemmatization. Lemmatized, MSD-tagged and disambiguated HNK is available for querying through standard client-server architecture Manatee/Bonito. The contribution concludes with future directions for HNK.

1 Introduction: pre-HNK Croatian corpora

Although the corpus linguistics activities in Croatia has started quite early – the first Croatian computer corpus was compiled and processed in 1967 by Željko Bujas (Bujas, 1974; Tadić, 1997) –, we had to wait for a very large reference corpus of Croatian of at least 100 Mw until 2005. It was the year the Croatian National Corpus (HNK) gained the size that is expected for a decent corpus of the third generation.

In the beginning the Croatian corpus linguistics activities were situated exclusively within the projects that were run at the Institute of Linguistics, Faculty of Humanities and Social Sciences (formerly Faculty of Philosophy) at the University of Zagreb. Between 1968 and 1973, within the large contrastive project led by Rudolf Filipović, the first English-Croatian parallel corpus was compiled. The Brown corpus, that was published a year earlier (Kučera & Francis, 1967), was obtained and processed in such a way that every single text sample was cut in half thus preserving the original balance of genres and topics. This 0.5 Mw English corpus was then translated and the resulting parallel corpus was used heavily for a whole series of contrastive projects resulting in a number of contrastive linguistics publications. In fact this was the first usage of the computer corpus in the contrastive research in the whole history of linguistics and this fact is usually unknown by most of corpus linguistics chronologists.

During ’70 and ’80 the works by old Medieval, Renaissance and Baroque Croatian authors were processed and concordanced, leading to a typical literary and linguistic computing type of activities of that time. This set of results represented an important basis for a number of philological text readings, criticism and authorship detection. But it also contributed to the experience much needed for larger corpora.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 219–228, 2009.

From 1976 until 1996 a One-million Corpus of Croatian Literary Language (also know as *Moguš's corpus*) was compiled and processed resulting in the Croatian Frequency Dictionary (Moguš et al. 1999) built upon it. This corpus was the first attempt to build a representative Croatian corpus even being of the modest size at the end, but certainly not in the beginning, in the time of its design. Still, the experience gained with processing this corpus was valuable for the following project, namely compiling of the Croatian National Corpus.

2 Previous versions of HNK

Although some initiatives and arguments for building a multi-million representative corpus of Croatian existed even earlier (Tadić, 1990), the real kick-start happened after the Ramesh Krishnamurthy had a two-day course in computational lexicography with heavy exploitation of corpora developed within the COBUILD project (Bekavac, 1997). This meeting was organized by the Croatian Academy of Sciences and Arts in Zagreb in 1997-11 and a number of Croatian linguists attended it. The representatives from the Ministry of science and technology understood clearly that building a large representative corpus was of the vital importance for preservation of Croatian language by building the language technologies for it. The financial support was provided from that point on. It could be said that it was a modest one, but still important for the beginning of building the HNK in late 1998. Later, HNK was supported in three nationally funded projects: Computational processing of Croatian/130718 (1996-2000), Development of Croatian language resources/0130418 (2002-2006) and Croatian language resources and their annotation/0130618 (2007).

2.1 Version 1.0

The theoretical background for HNK was laid down in two papers (Tadić, 1996; Tadić, 1998) where the need for a Croatian reference diachronic and synchronic corpus was expressed. Also its basic structure was defined, its size, span and other parameters indicated and web accessibility suggested from the first day.[1]

The basic structure of the v 1.0 was described in detail in (Tadić, 2002) so here we can give a brief general overview. The primary structure of the HNK was actually two-folded. It existed as:

1) *30-million corpus* of the contemporary Croatian standard language where texts produced since 1990 were collected covering different domains, genres and topics. This corpus was considered to be representative for the contemporary standard.

2) *HETA* (Croatian Electronic Text Archive) where texts older than 1990 were collected or the whole series of texts that would misbalance the representativeness of 30m-corpus.

This two-fold structure actually presented a real corpus (30m) and a text-collection (HETA) as a reserve text repository that could also serve in diachronical perspective.

Since there were no recent research on text production/reception in Croatia at that time that would give us some insight into text flow in the society, we had to rely on data

[1] The HNK web-page is at the address: http://hnk.ffzg.hr.

from commercial and marketing surveys, literary critics suggestions, general statistical data on book selling and borrowing, but we also inspected the structures of other very large corpora (BNC and ČNK).

From the beginning several important technical decisions were made. We limited the sources of texts by not allowing typing or OCR. This has left us just with e-text that had to be adapted and/or converted into desired encoding. Also, since we concentrated on the written language only, we needed no transcription. In this version of HNK there were no translations (although subtitles in movies and series have the largest audience in Croatia) and there were no poetry since we wanted to have a good average starting point for the most used functional styles. One of technical decisions that was made back in 1998 has proven to be rather long-sighted and that was the immediate usage of XML encoding instead of at that time more popular SGML.

By 2002 more than 150 Mw of texts were collected, but this collection was certainly not balanced. At that time 30m-corpus was more than 17 Mw in size. The 30m-corpus was freely searchable via web interface starting from the 1998-12-05 and the first test sample of only 3 Mw in size.

2.2 Version 2.0

In 2004 the two-fold conception has been abandoned for a simple reason. During that year at the Institute of Croatian Language and Linguistics a large diachronic text collection (from 11^{th} century onward) started to be collected.[2] This has lifted the burden of diachronic text collecting and processing from our shoulders and allowed us to concentrate on the contemporary Croatian standard and build a proper representative corpus.

The beta-version 2.0 of HNK was introduced in 2004-12 with 46 Mw of newspaper texts on a new technical platform, namely the well known Bonito/Manatee client/server architecture (Rychlý 2000). Since 2005-12, the HNK has reached 101.3 Mw in size covering different genres, text types and domains, but still not completely up to the desired proportions and balance.

2.3 New text samples for version 2.5

Since 2009-04 there is a new version (v 2.5) of HNK on-line. In that version there has been a change of certain subcorpora. The subcorpus "Klasici" that actually included some literary works older than 1990 but that are extensively used in elementary and higher schools as obligatory literature, was extracted from HNK and now it functions as an independent corpus of ca 3 Mw in size.

In its place a 6 Mw subcorpus of texts from "Vijenac", a bi-weekly newspaper for culture, science and arts has been introduced thus adding to the overall size of HNK (104.3 Mw). This addition also introduced a lot of new domains and topics that are now represented in HNK and that were missing.

But the main difference from v 2.0, that has also included the additional value to HNK, is the lemmatization and full MSD-tagging of HNK that has been done in early 2009.

[2] http://riznica.ihjj.hr

3 Lemmatization and MSD-tagging

Lemmatization and MSD-tagging of HNK was done by combining the existing language resources and tools for that task, namely Croatian Morphological Lexicon and MSD-tagger CroTag, into a hybrid system for MSD-tagging and lemmatizing.

3.1 Croatian Morphological Lexicon

Croatian Morphological Lexicon (HML) was generated by the Croatian Inflectional Generator (Tadić, 1992 and Tadić 1994) that was used as a computational model for Croatian inflection. It is a classification based system that includes 614 different inflectional paradigms (or patterns) that cover all phenomena in Croatian inflection (stem or ending alternations, different stem-endings behavior etc.). It was built as a flat model that respects the linguistic units, but it is not computationally optimized in any way and there it offers a lot of space for improvement.

The result of the generator was the Croatian Morphological Lexicon (Tadić & Fulgosi 2003; Tadić 2005) which is a simple list of triples (word-form, lemma, MSD). The tagset and lexicon format are MulTextEast compliant (Erjavec et al. 2003) and were developed following the specification for Croatian that exists in MulTextEast since 1998.

The HML v 4.6 covers: 45,000+ lemmas of general language; 15,000+ lemmas of personal fe/male names; 50,000+ lemmas of surnames registered in Croatia (Boras et al. 2003). All this yielded more than 4 million entries, i.e. triples (word-form, lemma, MSD). In HML 1475 different MSD tags were detected according to the MTE v3 specification and could give us the approximate image of the tagset complexity.

The HML is stored in a database in the Croatian Lemmatization Server[3] which is functioning as a freely accessible web service that allows individual queries, but also uploading of UTF-8 verticalized text or XML document for processing.

HML coverage was tested on a 46 Mw daily newspaper corpus and 96.4% of tokens were known, while 3.6% of tokens were unknown (mostly foreign names and misspellings) to the lexicon. Also, all words that are searched are stored in server logs and the system is automatically collecting the unknown words for manual updating or for procedures of automatic updating (Oliver & Tadić, 2004; Bekavac & Šojat, 2005).

3.2 CroTag MSD-tagger

After the initial experiments with application of a well-known stochastic tagger TnT (Brants, 2000) to the Croatian language (Agić & Tadić, 2006), where accuracy on PoS only was 98.63% and on full MSD was 89.95%, we decided to build our own stochastic tagger for Croatian (CroTag). The main reason was that it would allow us to control the fine-grained parameters of the system in order to achieve several more percentages over the baseline accuracy in full MSD tagging.

The CroTag was inspired by TnT and its open source reimplementation HunPos (Halácsy et al. 2007). It features trigram/second order HMM tagging paradigm with linear interpolation, suffix trie and successive abstraction for unknown word handling. Its input

[3] http://hml.ffzg.hr

and output formats are identical to TnT so it could be combined with the existing TnT tools for pre- and post-processing.

The CroTag was trained on subcorpus CW100, a 118 Kw tagged and manually disambiguated newspaper corpus collected from texts from "Croatia Weekly". The accuracy of full MSD tagging of CroTag was tested in two different cases:

1) realistic: CroTag trained on ten folds and tested on unseen part, accuracy: 86.05% in the worst case;

2) idealistic: CroTag trained on the entire CW100 and tested on seen part, accuracy: 97.51% in the best case.

Since the accuracy in the realistic scenario was not satisfactory, we decided to combine existing resource (HML) and tool (CroTag) into a hybrid system.

3.3 Hybrid system

How to improve overall MSD tagging accuracy by a hybrid system is described more in detail in (Agić et al. 2008) and here we will give just the general overview.

HML was encoded as a minimal finite-state automaton using TMT-library tools (Šilić et al. 2007) for conversion. The basic idea was to use the inflectional lexicon in that format as a run-time handler for words unknown to stochastic tagger. For full MSD-tagging we gained on the best case side (97.97%), but we also lost on the worst case side (85.58%). But, with this system we increased accuracy on the most problematic and highly frequent PoS i.e. adjectives, nouns and pronouns. The full MSD-tagging was done on the HNK v 2.5 and this version is now available for querying with Bonito featuring queries such as [msd="Nc..g"] which are common to other MSD-tagged corpora using this platform.

Also, the same hybrid system was used for disambiguation in the process of lemmatization. Different ways of merging the HML and CroTag output for disambiguating the lemmatization is described in detail in (Agić et al. 2009a). The general idea is to use the output of MSD-tagger for selecting between several possible MSDinterpretations from the inflectional lexicon in the case of homographs. Between several possible methods of merging inflectional lexicon with stochastic tagger, the most successful one was to compare the MSD provided by the tagger with the MSD of each lemma suggested by the HML. The best results were obtained not by using complete tag equality but by using tag similarity instead. In this way possible errors, that are usually appearing at the end of the tag, are not taken into account thus yielding the higher accuracy score. In the idealistic test case this merging method yielded 98.15% in lemmatization accuracy.

The fall-back option in the case of word-form unknown to HML is to take the type as a lemma. Of course, in the case of a poor inflectional lexicon coverage this would result in a huge number of lemmatization errors, but since the coverage of HML is quite large and the list of its entries is based on a frequency counts in a corpus, number of errors is kept well under control.

The lemmatization was done on the HNK v 2.5 and this version is now available for querying with Bonito featuring queries such as [lemma="glava"].

4 Conclusion and future directions

In this contribution we have presented the new version of Croatian National Corpus (v 2.5) and its features. We have described the previous versions of HNK, discussed the additional texts (or subcorpora) that were added to this version and also presented the MSD-tagging and lemmatization of HNK using a hybrid MSD-tagger and lemmatizer.

Future directions would certainly include enlarging the HNK so it would provide us with a quantity of language data of a size that should follow the general trends in corpus linguistics nowadays. The planned and achievable size is around 200-250 Mw. It should certainly include more fiction and other so far underrepresented text types and genres. In this respect a huge textual base of PhD synopsizes from all scientific areas has became available and it is being processed for inclusion into the next versions of HNK.

Regarding HML, we are working on its enlargement with foreign names, collected unknown words, automatic generation of certain categories of words (i.e. deverbative nouns, possessive adjectives of names, names of inhabitants etc.).

Since there is no full evaluation of MSD-tagging errors and/or lemmatization errors yet, we can not exactly say which direction should be taken. We have made a preliminary typology and statistics of errors (Agić et al. 2009b) but it gave us just the general trends and it remains to be verified at the particular cases whether a rule-based error handling could be more helpful at the most frequent cases of errors (e.g. *je* could be the 3rd person, singular, present of auxiliary *biti* 'to be' and it could be accusative, singular, feminine, clitic form of a pronoun *ona* 'she'; both being highly frequent in corpora and often misinterpreted by the system).

Also there is the problem of lemmatization of unknown words that has been tackled for Slovene (Džeroski & Erjavec, 2000). This work remains valuable for any Slavic language and we are planning the development of such a module for Croatian as well.

Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grants 130-1300646-0645, 130-1300646-1776 and 036-1300646-1986.

References

1. Agić, Željko; Tadić, Marko (2006) Evaluating Morphosyntactic Tagging of Croatian Texts. In: LREC2006 Proceedings, Genoa-Paris.
2. Agić, Željko; Tadić, Marko; Dovedan, Zdravko (2008) Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis, Informatica 32(4), pp. 445-451.
3. Agić, Željko; Tadić, Marko; Dovedan, Zdravko (2009a) Evaluating Full Lemmatization of Croatian Texts. In: Klopotek, Mieczyslaw; Przepiorkowski, Adam; Wierzchon, Slawomir; Trojanowski, Krzys (eds.). Recent Advances in Intelligent Information Systems, Academic Publishing House EXIT, Warsaw, pp. 175-184.
4. Agić, Željko; Tadić, Marko; Dovedan, Zdravko (2009b) Error Analysis in Croatian Morphosyntactic Tagging. In: Lužar-Stiffler, Vesna; Jarec, Iva; Bekić, Zoran (eds.). Proceedings of the 31st International Conference on Information Technology Interfaces (ITI2009), SRCE, Zagreb, 2009, pp. 521-526.

5. Bekavac, Božo (1997) Tečaj računalne leksikografije Ramesha Krishnamurthyja, Suvremena lingvistika 23(43-44), 437-438.
6. Bekavac, Božo; Šojat, Krešimir (2005) Lexical acquisition through particular adjectival endings for Croatian, Workshop on Computational Modeling of Lexical Acquisition, Split, PPT presentation.
7. Boras, Damir; Mikelić, Nives; Lauc, Davor (2003) Leksička flektivna baza podataka hrvatskih imena i prezimena. In: Tudman, Miroslav (ed.) Modeli znanja i obrada prirodnoga jezika, Zavod za informacijske studije, Filozofski fakultet, Zagreb, pp. 219-237.
8. Brants, Torsten (2000) TnT – A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing. Seattle, Washington, pp. 224-231.
9. Bujas, Željko (1974) Kompjutorska konkordancija Gundulićeva Osmana, Sveučilišna naklada Liber, Zagreb.
10. Džeroski, Sašo; Erjavec, Tomaž (2000) Learning to Lemmatise Slovene Words. In: Cussens, J.; Džeroski, S. (eds.) Learning language in logic, LNCS 1925, Springer, Berlin, pp. 69–88.
11. Erjavec, Tomaž; Krstev, Cvetana; Petkevič, Vladimir; Simov, Kiril; Tadić, Marko; Vitas, Duško (2003) The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In: Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages, ACL, Budapest, pp. 25-32.
12. Halácsy, P.; Kornai, A.; Oravecz, C. (2007) HunPos – an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions. ACL, Prague, pp. 209-212.
13. Kučera, Henry; Francis, Nelson W. (1967) Computational Analysis of Present-day American English. Brown Unviersity press, Providence.
14. Moguš, Milan; Bratanić, Maja; Tadić, Marko (1999) Hrvatski čestotni rječnik, Školska knjiga Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu, Zagreb.
15. Oliver, Antoni; Tadić, Marko (2004) Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In: LREC2004 Proceedings, Lisbon-Paris, Vol. IV, pp. 1259-1262.
16. Rychlý, Pavel (2000) Korpusové manažery a jejich efektivní implementace (Corpus Managers and their effective implementation). Ph.D. thesis, University of Brno. (http://www.fi.muni.cz/∼pary/disert.ps)
17. Šilić, Artur; Šarić, Frane; Dalbelo Bašić, Bojana; Šnajder, Jan (2007) TMT: Object-oriented text classification library. In: Proceedings of the 29th International Conference on Information Technology Interfaces (ITI 2007), SRCE, Zagreb, pp. 559-566.
18. Tadić, Marko (1990) Zašto nam treba višemilijunski referentni korpus? In: Andrijašević, Marin; Vrhovac, Yvonne (eds.) Informatička tehnologija u primijenjenoj lingvistici, Croatian Applied Linguistics Society, Zagreb, 1990, pp. 95 98.
19. Tadić, Marko (1992) Kompjutorska obrada morfologije hrvatskoga književnog jezika na imeničnom potkorpusu, M.A. thesis, Centre for postgraduate studies, Dubrovnik.
20. Tadić, Marko (1994) Računalna obrada morfologije hrvatskoga književnog jezika, Ph.D. thesis, Faculty of Philosophy, University of Zagreb, Zagreb.
21. Tadić, Marko (1996) Računalna obrada hrvatskoga i nacionalni korpus, Suvremena lingvistika, 22(41-42), pp. 603-612.
22. Tadić, Marko (1997) Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive, Suvremena lingvistika, 23(43 44), pp. 387-394.
23. Tadić, Marko (1998) Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika, Filologija, 30-31, pp. 337-347.
24. Tadić, Marko (2002) Building the Croatian National Corpus. In: LREC2002 Proceedings, Las Palmas-Pariz 2002, Vol. II, pp. 441-446.
25. Tadić, Marko (2005) The Croatian Lemmatization Server, Southern Journal of Linguistics, 29(1-2), pp. 206-217.

26. Tadić, Marko; Fulgosi, Sanja (2003) Building the Croatian Morphological Lexicon. In: Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages, ACL, Budapest, pp. 41-46.

Bringing language technology to the masses Some thoughts on the Hungarian online spelling dictionary project

Tamás Váradi

Research Institute for Linguistics
Hungarian Academy of Sciences

The present paper is a short informal account of a project the idea of which was conceived simultaneously and (Deme 1999)independently by both Karel and myself, as it turned out at the recent FLARENET Meeting in Vienna. The aim of the project is to provide language guidance for the general public. This activity has always been an integral part of the tasks of our Institute, the Research Institute for Linguistics of the Hungarian Academy of Sciences. The Institute was established in 1949 and assigned the task among others, of maintaining and so called cultivating the Hungarian Language. Indeed, the whole Academy of Sciences was established in the mid-nineteenth century with the purpose of guarding and cherishing the Hungarian language. While the Academy was founded as part of the drive for independence and national self-assertion, long after these goals have been achieved Hungarian society still very rigidly norm following in terms of language use. The Academy is looked upon as the final arbiter of language use and the Institute is still largely perceived as the ultimate source of knowledge of how Hungarian should be spoken nice and proper.

Hungarian spelling (MTA 1985) is claimed to be following both pronunciation and the morphological structure of the words. The grapheme phoneme correspondence is certainly more transparent than in English or French but of course the conservative nature of the spelling system inevitably means that there is a fair amount of unmotivated graphic rendering.

For decades, the Institute has been running language guidance services in the form of a telephone hotline, which latterly has been complemented with the use of email. The rising cost of labour and the steady popular demand has led to the decision to use the web as the primary channel for the service. As the logs of decades of language guidance indicate, the overwhelming majority of the enquiries relate to matters of spelling. That is why we decided to implement first an interactive online spelling dictionary.

There are two major print-based spelling dictionaries available on the Hungarian market. One is published by Akadémiai Kiadó (Deme, Fábián and Tóth 1999) and represents the official publication of the Academy Commission on the Hungarian Language, the body that is entrusted with codifying Hungarian orthography. Recently, its ruling position on the market has been challenged commercially by an orthographical dictionary published by a leading publishing house Osiris (Laczkó és Mártonfi 2003) The print editions of orthographical dictionaries contain anything between sixty to two hundred thousand entries, which include regular mention of a handful inflected forms that are thought to be problematic. Both dictionaries contain a citation or discussion of the official rules of orthography with examples.

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 227–232, 2009.

Akadémiai Kiadó has recently produced an online implementation of the Hungarian Orthography Dictionary, which is offered as a bonus to the CD edition packaged with the printed version (MTA 2008). The dictionary provides little more than a lookup of word forms with an occasional reference to the relevant rule of the Academy Orthography. A pirated edition of the Akadémiai Kiadó version at `http://www.magyarhelyesiras.hu` also offers the same service. If there is no hit, users are left to their own devices.

The approach we adopted rests on two vital technologies: interactive web 2.0 technologies and language technologies. In the rest of this article, I will sum up our initial findings on the applicability of the latter to this task.

The obvious language resources and tools that could be brought to bear on the task include a corpus, a frequency dictionary, a morphological analyzer and generator, a semantic lexicon and a set of local grammars to be deployed interactively in disambiguating user queries.

A comprehensive reference corpus is a valuable source of data for the frequency dictionary and other the language technology support. Much as it is useful to be informed about prevailing frequencies in actual language use, some caution is warranted in taking raw frequency data at face value. It is, after all a fact of life that the web, particularly in this day and age of the blog culture, will contain forms that go against the rules of the official orthography. For the purposes of the present spelling dictionary project, we decided to exercise some caution. We made it a point of principle to apply a higher than usual frequency threshold level as a filter against 'deviant' forms. In addition, we have extended the frequency list of the Hungarian National Corpus (Váradi 2000) with a new collection of electronic texts from the press or general reference material published on CD's. Altogether we have generated an additional corpus of 412.6 million running words, yielding 8.77 million word forms.

At the same time as we emphasize the importance of using sources that can be reasonably presumed to have undergone some sort of editorial control, it should not be concluded that we think that corpus data cannot serve as source of guidance to orthography. There are two good reasons for adducing corpus data as evidence. Hungarian orthography professes to reflect language use at least in terms of conformity to pronunciation, which is reflected by the way people "lapse" into committing spoken language into writing. Corpus evidence is *the* prime source of information about changing language use. Secondly, rules (much as grammars) tend to "leak" in that the orthography regulation is often too general and the examples cited are too sparse to be able to adjudicate on a specific point of usage. In such a case, there is every justification for drawing deciding evidence for the correct spelling from facts of actual usage.

The last two points are important not just for the purposes of the present project. It is our conviction that the body of orthography "legislators" should draw great benefit from being informed in their constitutive activity by the evidence of actual language use furnished by corpus data.

Considering the use of *frequency data* we were confronted with the question whether the task of compiling a good spelling dictionary cannot be reduced to producing a massive frequency dictionary. While we opted to compile as large a corpus as possible, we had to realize that at least for Hungarian the use of brute force strategy will not work. Given the enormously rich morphology (any ordinary noun root may easily give rise to close to eight hundred word forms without considering the multiplicative effect of productive

derivation and compounding (Váradi & Oravecz, Morpho-syntactic ambiguity and tagset design for Hungarian, 1999)) the chances for a corpus of any practical size yielding a complete coverage of all the potential forms are very thin indeed. As a comparison, the seventy thousand lemmas in the standard desk size explanatory dictionary of Hungarian can yield about 120 million productive word forms. On the other hand, the web corpus based on a snapshot of the whole Hungarian web of the day, yielded about 8 million word forms. What is more significant, though is not just the number of missing forms but the fact that they appear to be left out by chance, they feel intuitively just as plausible as those included in the list.

On the other hand, not every word form of a paradigm is of interest when it comes to orthography. The printed dictionaries usually contain just a few forms, which readers can use as basis for analogy about all the other cases that present a problem at all. Instead of going to the lengths of generating non-attested word form lists of an intractable size, we have decided to ensure that we have full coverage of these diagnostic forms for all the lemmas be generating the forms that happen to be absent from the original frequency list.

One consideration that must guide our work throughout is the fact that expectations of the public are very high in terms of accuracy of the data. Precision and recall figures that would earn praise for most language technology project may simply prove unacceptable in the present context. This makes us wary of deploying a morphological analyzer for fear of generating an unacceptable amount of spurious analyses. As a matter of principle we want to rely on the robustness of the dictionary instead.

On the evidence of the internal log books, of the language guidance hotline service, the overwhelming majority of the spelling enquiries relate to compounding i.e. whether two words are to be written separate or together. We may cover part of the queries by assuming that our database will contain all words that are written together. This already is a tall order in the light of the extremely productive compounding rules in Hungarian. However, it does not necessarily give corroborative evidence for the cases that are to be written separately, after all, failing to find something in the database may be due to missing data. Therefore it becomes even more imperative to enrich the lexical database with as many multiword units as is feasible to collect and this is, of course, an actively researched area where language technology has a lot to contribute.

Unfortunately, the issue of word division along with quite a number of phenomena is typically unresolvable without recourse to contextual information and, indeed, semantic disambiguation. The majority of phenomena involved go beyond the reach of automated procedures. We cannot delude ourselves of providing a fully automated language guidance service anyway. Yet, before resorting to the personal helpdesk, we intend to experiment with interactive disambiguation involving the user. The idea is to tap the native speaker intuition of users to select the intended correct form through relevant examples. The difficulties we are facing stem not only from reliably predicting ambiguities but devising interactive routines with the users that are couched in terms that they find easy to understand and which enable them to make a clear choice. This is sometimes quite a challenge as some of the distinctions in the orthography rules rest on highly elusive distinctions between technical terms. In order to facilitate the recognition and possible disambiguation of ambiguous phenomena, we have collected all references to semantic classes in the orthography rules and have decided to mark the lexical database for these

semantic features. The compilation of such a semantic lexicon will be a valuable resource in its own right with great potential use in other language technology applications.

In conclusion, this brief, informal survey of the problems encountered and solutions devised presents a somewhat paradoxical picture. Typically when language technology is applied to automate a process that used to be done by hand, the problem usually is to clean up the data and make the tasks explicit enough for algorithmic treatment. Data for "direct human consumption" as it were, could be slack and ambiguous because the users instinctively interpreted them in the right manner. It was the machine that presented a challenge in this regard.

To set up an online spelling dictionary – at first blush an innocuous task – presents the opposite challenge. The amount of noise considered inevitable in machine systems may well be unacceptable and we quickly run into tasks that lie beyond current technology. Yet, this is an exciting application and because of the intense popular interest in matters of language use it is an area that, as the title suggests, can bring language technology to the masses.

References

1. Laczkó, K., & Mártonfi, A. (2003). *Helyesírás (Orthography).* Budapest: Osiris Kiadó.
2. Deme, L., Fábián, P., & Tóth, E. (1999). *Magyar helyesírási szótár (Hungarian Orthographical Dictionary).* Budapest: Akadémiai Kiadó.
3. MTA. (2008). *A Magyar Helyesírás szabályai - 3 az egyben (Rules of Hungarian Orthography – 3 in one).* Budapest: Akadémiai Kiadó.
4. MTA. (1985). *A magyar helyesírás szabályai (The rules of Hungarian Orthograph).* Budapest: Akadémiai Kiadó.
5. Váradi, T. (2000). The Hungarian National Corpus. *Proceedings of theSecond International Conference on Language Resources and Evaluation* (pp. 385-389). Paris: ELRA.
6. Váradi, T., & Oravecz, C. (1999). Morpho-syntactic ambiguity and tagset design for Hungarian. In T. Branst (Ed.), *Proceedings of EACL'99* (pp. 8-22). Bergen: Associaton for Computational Linguistics.

Reasoning About Events: the Spatio-Temporal XRCD Calculus

Gérard Ligozat[1,2], Zygmunt Vetulani[1]

[1] Adam Mickiewicz University, Faculty of Mathematics and Computer Sciences
ul. Umultowska 87,
61-614 Poznań, Poland
[2] LIMSI, Paris-Sud University, Orsay, France
{ligozat,vetulani}@amu.edu.pl
ligozat@limsi.fr

Abstract. Abstract. The main focus of this paper is on formalisms for representing events and reasoning with respect to time and space. The proposals discussed in the paper are used for the development of a system providing assistance to the security personnel during a large scale event involving a huge number of participants. Processing spatio-temporal information is crucial for such systems. The system will provide the security personnel (analysts) with visualization facilities and suggestions for decision making. We argue for the adoption of a formalism called the XRCDC (an extension of the Region Cardinal Direction Calculus) as suitable for representing and reasoning about events in this context.
Key words:spatio-temporal formalism, computer understanding of natural language, visualization

1 Introduction

The main focus of this paper is on formalisms for representing events with respect to time and space. An appropriate formalism is necessary to develop systems processing the information conveyed by text messages in terms of events. These events have various locations, and occur at definite moments. Representing the spatial and temporal aspects of the information to be processed is one of the central functionalities of the monitoring system. The general considerations of the paper are illustrated by a case study where the formalism is being applied to the POLINT-112-SMS system. The visualization module of this system directly uses the idea of active map, a concept using the XRCDC formalism discussed in this paper.

2 Visualizing events: the active map

We consider event-based systems for processing incoming information. The purpose of such a system is to assist decision making from the part of the security personnel. We focus here on the visualization module which uses what we call an active map (AM). An active map is a pictorial representation of a set of important events, which can be represented in various forms. The best way to understand the underlying concept is to think of the active map as a visualization system which constitutes a computerized animated extension of

Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): After Half a Century of Slavonic Natural Language Processing, pp. 231–247, 2009.

the traditional maps used by the military. In the traditional General Headquarters of past centuries, generals used pins, small flags or any other gadgets indicating the locations and moves of the various troops in order to evaluate the situation and to take appropriate decisions. The AM is meant as a sophisticated version of such traditional maps. The idea of building active maps is closely related to previous work on the visualization of military campaigns (Ligozat, Nowak, Schmitt, 2007). In the application discussed in this paper, the active map is based on a static background representing a soccer stadium. Events are represented by schematic images, but they may also have sounds, colored lights, blinking lights associated to them. These representations can be queried e.g. clicking on them would provide the original texts which resulted in the generation of those events. The active map also provides zooming facilities which allow the user to "take a closer look" at local situations if necessary.

2.1 Events in the active map

The input to the monitoring system is in terms of messages. Typically, a new message will trigger the generation of what we call an initial event (IE) with a set of features associated to it. As an illustrative example, assume that the system has to process the following message: *Some supporters in sector 4 are throwing stones at the security personnel*. This message will trigger the creation of an initial event. The following set of features is associated to this IE :

source (the observer)
actors (acting agents: supporters; patients: security personnel)
location (sector 4)
temporal span (time duration of the event)
time of reception (time of message reception)
action (type of action: throw; instruments: stones)
aspectual type (on-going event)

2.2 Introducing new events

As events occur in real time, the processing system receives a sequence of messages. For each new message, a new Initial Event is created. If co-references are detected to a previous initial event, the decision has to be made to graft the new information onto this pre-existing event, hence adding new information to it.

As an example of grafting, consider the following message by the same informer: *Some among them are wielding baseball bats*.

First a new IE is created, whose source, actors, location and temporal span can be recognized by the system as co-referential to those of the previous event. The remaining features are:

time of reception of the message
action (type of action: wield; instruments: baseball bats)
aspectual type (on-going event)

Hence the system will graft the new IE onto the previous one in the active map, while adding new features to it, and new information is made available for inference (here, the presence of dangerous weapons can be inferred).

3 Requirements for the spatio-temporal components

The languages used to represent the temporal and the spatial structures of the events have to allow the representation of qualitative or indeterminate information. So-called qualitative formalisms such as those derived from Allen's calculus (Allen, 1983), including the rectangle calculus or the region cardinal direction calculus, have these properties. The choice of a suitable language of representation depends on making decisions about the parameters of the temporal and spatial representation to be used. This implies answering a number of questions:

a) Ontological status: of what type are the objects to be represented? Can they be abstracted as points, lines, regions with geometrical shapes, connected regions?

b) Nature of the information: is it quantitative (the kind of information a robot receives from its sensors) or predominantly qualitative (as is mainly the case with the information carried by natural language).

c) Nature of the surrounding space: can we make use of global systems of reference (like north, south, for instance) or do we have to be content with local frames of reference (typically, the frame of reference represented by some moving person or object)?

d) What is the dimension of the surrounding space? Can we reason in 2D, in an augmented version of 2D (for instance with a finite number of 2D levels), or do we need a full 3D space?

e) Nature of the kind of spatial relations we want to represent. If one looks at the state of the art in the domain of qualitative spatial reasoning, three main types of relational information have been predominantly studied: topological information, that is relations such as containment, partial overlap, having adjacent boundaries, or disjointedness; directional information, with respect to some frame of reference; and qualitative distance information (near, far, very far).

f) Is time continuous or discrete? In the representation of linguistic data, continuity is usually assumed (since it is a property of language that it is able to open up any event and re-consider a previously punctual situation and present it in a second consideration as an extended one). It has to be remarked, however, that many formalisms can accommodate both a continuous and a discrete interpretation. This is in particular the case of Allen's calculus.

g) Questions linked to the way of anchoring the abstract model to the actual situation: for instance, a match involves absolute temporal landmarks (beginning of the game, end of the game, half-period, additional time) as well as occasional, or contingent temporal landmarks (important events of the game, such as goal, penalty, and so on).

4 Directional calculi

4.1 The rectangle calculus

The rectangle calculus considers objects which are rectangles whose sides are parallel to the axes of coordinates in a 2D Euclidean plane. Given such a rectangle as a reference, the qualitative position of any other rectangle can be described using the projections on the axes of coordinates, which are intervals. Hence those positions are described by a pair of Allen relations. For instance, in Fig. 1, rectangle A is in relation (*oi,mi*) with respect to

rectangle B, since the horizontal projection of A is overlapped by that of B (Allen's oi relation) while the vertical projection of A is met by that of B (Allen's relation *mi*).

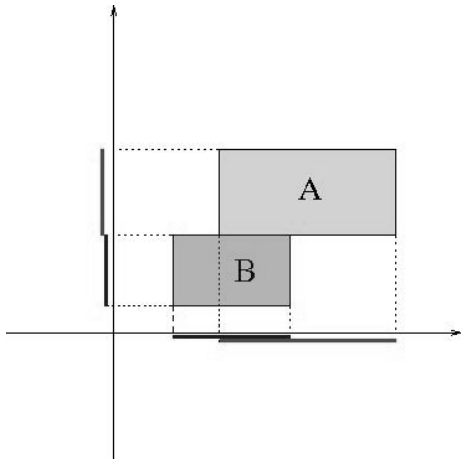


Fig. 1. The rectangle calculus

One obtains in this way a formalism whose composition table, which is the main tool for propagating knowledge, is basically Allen's table. The formal properties of the calculus have been extensively studied (Balbiani, Condotta, and Fariñas del Cerro, 1999).

The rectangle calculus can be easily extended to a calculus about rectangles, points and lines, with the restriction that the lines have to be parallel to the axes. For objects having other shapes, the simplest method consists in replacing them by their minimal bounding rectangle. For instance, in the case of Fig. 2, the two regions A and B have two minimal bounding rectangles mbr(A) and mbr(B), whose relative position can be encoded as a rectangle relation.

The drawback is that much information can be lost in this way. For instance, disjoint objects may have overlapping rectangles. This is illustrated in the same figure, where two objects A and B which are disjoint have overlapping bounding rectangles.

4.2 The Region Cardinal Direction Calculus

The basic cardinal calculus deals with points in 2D space and the eight basic cardinal directions N, S, E, W, NE, SE, NW, SW, augmented by the identity relation eq. Actually, the basic cardinal direction calculus is a 2D version of the time-point calculus, in the same way as the rectangle calculus is a 2D version of Allen's calculus. It has also been studied extensively and its formal properties are well known (Ligozat, 1998).

In order to deal more precisely with arbitrary extended regions in 2D space, an extension of the cardinal direction calculus, called the RCDC (region cardinal direction calculus) has been proposed by Goyal and Egenhofer (Goyal and Egenhofer, 2001) as well as by Skiadopoulos and Koubarakis (Skiadopoulos and Koubarakis, 2001, 2004).

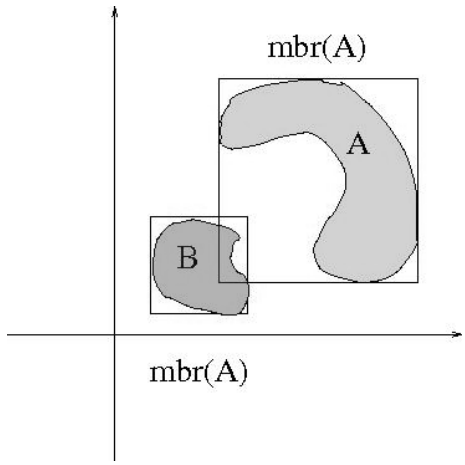


Fig. 2. Two objects and their minimal bounding rectangles

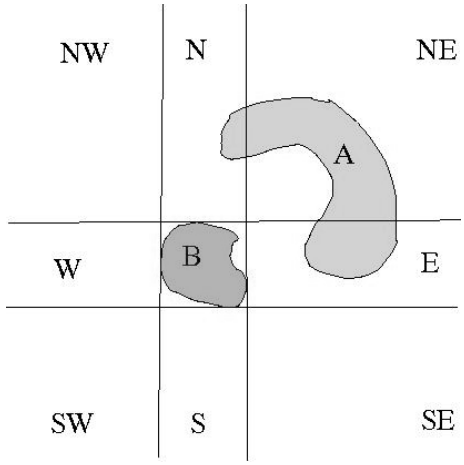


Fig. 3. The region cardinal direction calculus

The starting point of the representation consists in considering the nine regions (called tiles) defined by the minimal bounding rectangle of a reference object B (Fig. 3). Eight of them, labelled N, S, E, W, NE, SE, NW, SW, are infinite regions. The minimal bounding rectangle itself, labelled O, constitutes the ninth tile. In this way, the position of any region A can be described by listing the set of tiles which intersect the interior of A. Alternatively, the same information can be conveyed by a Boolean array of length 9 (or a Boolean 3×3 matrix) dir(A,B) listing whether A has a non-empty (denoted by 1) or empty (denoted by 0) region in common with each of the nine tiles NW, N, NE, W, O, E, SW, S, SE, in that order. For instance, in Fig. 3, the relation of A with respect to B is (N:NE:E), or, using a Boolean array dir(A,B) = [0,1,1,0,0,1,0,0,0].

The RCDC possesses very interesting properties: it deals with (connected) regions in the plane and, although still based on the same basic relations, allows a finer expression of the relative positions of more general regions.

Because of its extended base, the RCDC also implicitly encodes topological information (as does the rectangle calculus). For instance, referring back to Fig. 3, the fact that A and B do not intersect is a consequence of the stronger fact that A does not intersect the tile containing B, and this fact is encoded in the representation (O does not belong to (N:NE:E). Moreover, the RCDC may be considered as a refinement of the rectangle calculus. Recent results (Zhang et al., 2008) show that basically it allows to consider a spatial configuration as 'pixelized' by the objects present in the environment, so that those 'pixels' can be used to characterize a 'silhouette' of each object which is much finer than its bounding rectangle. This also means in particular that, if we decide to replace the objects by their minimal bounded boxes, we get in substance the rectangle calculus.

5 The Extended Region Cardinal Direction Calculus (XRCDC)

Events are spatio-temporal entities: an event has both a spatial extent and a temporal span. Hence a qualitative formalism for representing events has to include a temporal dimension besides the spatial ones. In the applications we are dealing with, we can assume as a first approximation that two dimensions are enough for representing space. Hence we use a three dimensional spatio-temporal universe. The simplest choice would be to use the equivalent of the rectangle calculus in three dimensions (called the 3-block calculus), in a spatio-temporal context. However, for reasons analogous to those discussed for space, such a simple solution has drawbacks best illustrated in a one-dimensional example**.**

5.1 A 1D example

Consider a one dimensional space (which could be the abstract representation of a portion of a road). A group of people have their vehicle (a camper) positioned near a forest. During the night, a fire starts at some place A, then progresses along the road. If the camper stays in position, it will be caught up by the fire. Fortunately, its proprietors get aware of the fire, wake up, and move their vehicle away. In the morning, the fire has died out. The camper can move back close to its original position on the road.

As entities in space-time, the fire and the camper can be represented in a 2D space with a spatial dimension and a temporal dimension, as in Fig. 4. Using minimal bounding rectangles in the situation shown in Fig. 4 would lose the information that the spatio-temporal extents of the fire and the camper are in fact disjoint, and this would lead to the erroneous conclusion that the passengers of the camper may have perished in the fire.

Using the extended cardinal direction scheme in this 2D setting means that we will both consider the relation dir(A,B) of A with respect to mbr(B), which in the case of Fig. 4 is (W:O), or equivalently the Boolean array (0,0,0,1,1,0,0,0,0), and the relation dir(B,A) of B with respect to mbr(A), which is (N:NE:E:SE:S), or in the array notation (0,1,1,0,0,1,0,1,1). Looking at the latter and observing that O is not a member of (N:NE:E:SE:S), we know that B does not intersect mbr(A), and hence, a fortiori, that A

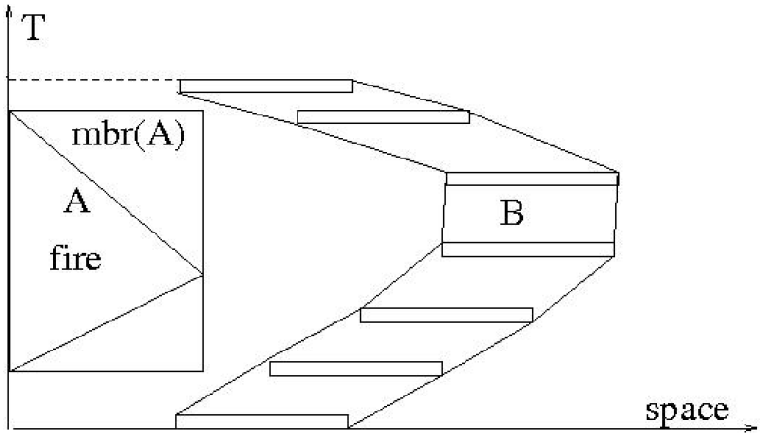


Fig. 4. A fire (A) and a camper (B) in a 2D space-time

and B are disjoint. We can conclude that the passengers of the camper may have been saved.

5.2 When is disjointness preserved?

In the general case, if A and B are two regions in the Euclidean plane which are disjoint, the rectangle calculus will represent them as intersecting if their minimal bounding rectangles mbr(A) and mbr(B) intersect. The XRCDC will only represent them as intersecting if both conditions " A intersects mbr(B)" and "B intersects mbr(A)" are met.

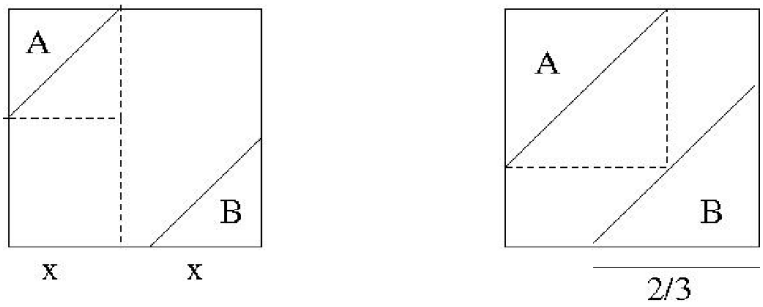


Fig. 5. Two triangles A and B

Consider the toy example of two triangular regions in a unit square represented in Fig. 5. Suppose that both triangles have two sides of length x, where $0 < x < 1$. Then it can be easily shown that their minimal bounding rectangles intersect if x is greater than 1/2. In the left part of the picture, $x < 1/2$. On the other hand, the minimal bounding

rectangle of each triangle intersects the other triangle only if x is greater than 2/3. The limit case is shown in the right part of the picture. In this particular case, then, the XRCDC formalism is at an advantage with respect to the rectangle calculus if $1/2 < x < 2/3$.

5.3 The extended region cardinal direction calculus (XRCDC)

The extended region cardinal direction calculus is the three dimensional analogue of the 2D space-time we have just considered above.

However, since dealing directly with 3 dimensions would result in 27 tiles, the XRCDC deals separately with three 2D projections: if the spatial dimensions are X and Y, and the temporal dimension T, then we consider the three pairs (X,Y), (X,T), and (Y,T). For each pair, the projections of the reference object B is considered, resulting in nine tiles in the corresponding plane. An arbitrary object A has a projection whose relation with respect to the projection of B can be described by listing the set of tiles having a non-empty 2D component with the projection of A. Hence the XRCDC uses triples of direction-relation lists dir(X,Y)(A,B), dir(X,T)(A,B), dir(Y,T)(A,B) of the region direction calculus to represent relations between events.

6 Reasoning about events using the XRCDC

The XRCDC is a qualitative reasoning which basically fits into a wide family of calculi which are similar to Allen's calculus. For all those calculi, reasoning is based on the propagation of information using the operation of composition of relations. In the context of the RCDC, the problem of computing composition has been extensively studied by Skiadopoulos and Koubarakis (Skiadopoulos and Koubarakis, 2004), who defined explicit algorithms of computation. The same methods can be used in the XRCDC, by computing composition independently for the three pairs of dimensions.

Reasoning about events involves using both explicit knowledge (such as concluding directly to spatio-temporal disjointness; cf. the fire-and-camper example), and implicit knowledge deduced by using composition. In the actual use of the formalism described below, other features of the events such as closeness are represented and may be used for reasoning.

7 Application to the POLINT-112-SMS project

7.1 The POLINT-112-SMS project

The idea of active map described above and its implementation based on the XRCDC are put into practical application in the POLINT-112-SMS project. Within this project we faced the important problem of enhancing information processing in real-life situations where the decision making process strongly depends on the quality and adequateness of the information acquired by the deciding person or team. Typical situations are those involving a large concentration of people behaving emotionally and participating in an event protracted over an extended time period. Such situations belong to the category of emergency situations where early diagnosis may decide of the efficiency of preventive measures.

In the POLINT-112-SMS project, we focus on monitoring large scale sports events, such as high-risk soccer matches, where potentially dangerous events can occur at hardly foreseeable moments. Political events, popular musical shows etc. also belong to this category.

It is the responsibility of the organizers and/or specialized services (e.g. the police) to set up appropriate security structures. Typically, such a structure involves (at least) the following three categories of agents: observers, analysts/decision makers and operating forces. The role of human observers is crucial. It consists in keeping close attention to what is happening in various parts of the scene (e.g. a soccer stadium) and to inform the analysts from the central crisis management center of any unusual or menacing event they have observed. Observers must communicate with analysts using appropriate communication channels and human communication protocols.

Concerning communication, several problems have to be taken into account:

- the capacity of the communication channel and the structure of the deciding body may be at the origin of important bottlenecks in case of critical situations (large number of messages coming in the same time from different observers).

- the quality of information processing has a direct impact on the quality of crisis management: contradictory pieces of information as well as impossibility of interpreting all of them at the same time may be at the origin of serious (sometimes critical) deformations or misinterpretation of the incoming messages.

- on-the-fly interpretation of human reports is at the origin of errors due to the lack of precision of these reports, typically made in human, uncontrolled language (including errors, abbreviations, lack of conceptual consistence).

Because of time constraints (time pressure), of the dynamic character of the environment, of the nature and quality of the data (poor quality data, contradictory elements of information, imprecision), and of the need for tracing and monitoring the decision process, it is desirable to support the decision process with automatic processing. In order to combine the human knowledge based on the human conceptualization of the world with computer technology, qualitative models of knowledge representation and processing seem appropriate.

In the POLINT-112-SMS project we assume that the observers use cell phones to communicate between them and with the crisis management headquarters. Because of the noisy (and potentially hostile) environment, a privileged way of communication is via SMS messages. The crisis management supporting system supposed to process information primarily expressed in natural language needs to be interfaced to a subsystem for understanding written natural language short messages. The tasks of the system include:

- monitoring the natural language information flow from the observers to the crisis management center (and in some cases interact with the users),

- interpreting messages, queries and requests addressed directly to the system,

- analyzing messages in order to extract information,

- interpreting and processing this information (including consistency checking and visualization).

The overall vision of the system is shown in Fig. 6.

The central idea of the knowledge management in the system is to process and represent information in terms of events. Events are located at various locations and

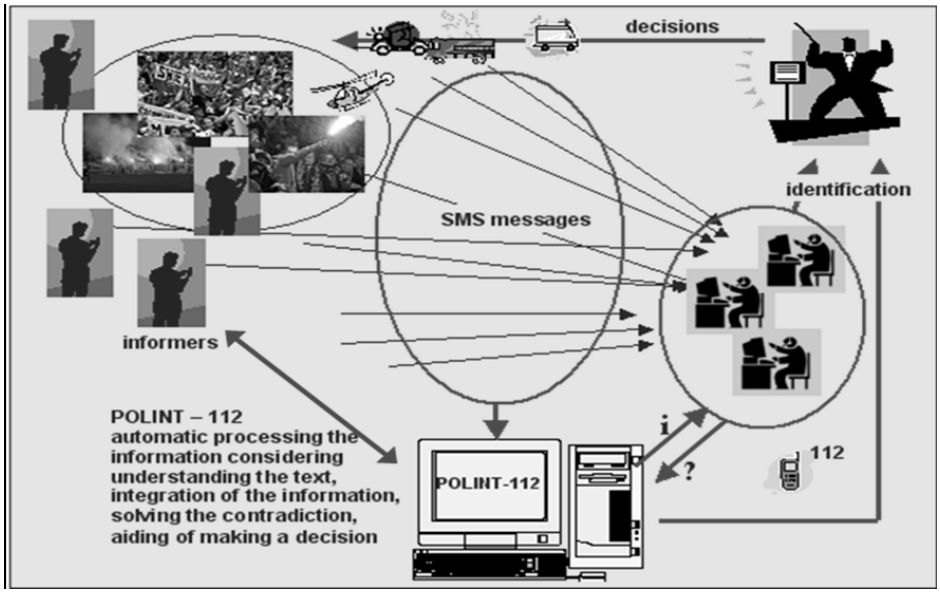


Fig. 6. POLINT-112-SMS system (Vetulani et al. 2008)

occur at definite moments in time. Hence representing the spatial and temporal aspects of the information to be processed is a central problem for the reasoning and monitoring functionalities of the system.

The system's architecture is presented in Fig. 7. The roles of the components are as follows (Vetulani et al., 2008):

1. The SMS Gate is a module allowing SMS communication with the informer by means of SMS messages. It is composed of two submodules, one of which is responsible for sending messages, the other one for receiving them. The SMS Gate communicates directly with the NLP Module.

2. The NLP Module is the main module responsible for text processing (parsing, surface understanding). The NLP Module communicates directly with the SMS Gate and with the DMM.

3. The Dialogue Maintenance Module (DMM) is responsible for dialogue with the informer. It takes into account the data controlled by the Situation Analysis Module. Thanks to the DMM, the NLP Module focuses on transforming single sentences into data structures without storing and processing these structures. The DMM communicates directly with the NLP and SAM modules.

4. The Situation Analysis Module (SAM) is responsible for reasoning. It acts as the "brain" of the system. It controls a number of subordinate modules, presented in points 5-12 (at this point some of them are integrated with the SAM). The SAM reasons about the structures without directly communicating with the informers. The SAM communicates directly with the DMM.

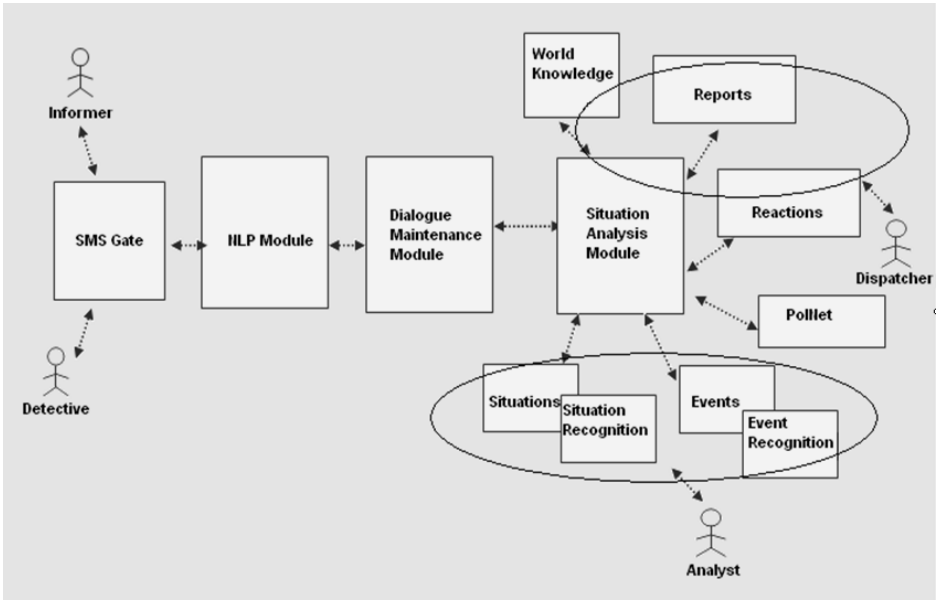


Fig. 7. The logical model for the Polint-112-SMS system

5. The World Knowledge Module stores general knowledge. It is used as the system's knowledge base. It may contain knowledge about e.g. medical emergency procedures, city maps, and other information.

6. The PolNet Module. PolNet is a WordNet-type ontology. Apart from the basic relations of hyponymy / hyperonymy it also contains relations that facilitate various forms of reasoning.

7. The Reports Module stores information obtained from users in the form of Reports.

8. The Events Module stores information about events. The information stored in this module can be directly accessed by the Analyst

9. The Event Recognition Module is responsible for creating new events in the Events Module.

10. The Situations Module stores information about Situations. The information stored in this module can be directly accessed by the Analyst.

11. The Situation Recognition Module is responsible for creating new situations in the Situations Module..

12. The Reaction Module is responsible for informing the Dispatcher at the Crisis Management Center that an action has to be taken (e.g. dispatching an ambulance).

Reasoning in the SAM. The reasoning activity involves various tasks, such as deciding the right time for a specific action, identifying the components of a complex event and characterizing it based on its temporal structure, deciding that events in a sequence of events may be causally related (an explosion followed by a movement of the crowd) or not (the same events in the reverse order). Therefore the system should be able to represent temporal and spatial knowledge gathered from various sources (observers,

built-in knowledge modules and/or event-type specific predefined knowledge) and dispose of reasoning tools for processing this information. A specific reasoning activity is also requested by the component of the system which provides the user with visualization tools (to be discussed more in detail below).

7.2 Visualization aspects of the POLINT-112-SMS project

The visualization of events and objects plays an important role in decision supporting systems such as POLINT-112-SMS (Fig. 8). It must give quick, synthetic insights into a dynamically changing situation. Since visualization is directly devised for humans, it must be in agreement with human conceptualizations: appropriate abstraction levels, especially concerning time and location of events, suitable external aspect of entities (persons, important artifacts, landmarks). This is a further reason why a qualitative approach to space and temporal knowledge representation is appropriate.

A stadium as a static element has a well-defined structure with hierarchical aspects. Most of the action, from the point of view of security, is likely to happen outside the playing field (excepting the rare possibility that groups of supporters or spectators enter it). So the main locations will refer to regions in the part of the stadium where spectators and supporters are present.

The situation of a soccer game involves a great deal of a priori knowledge about what a game is and what has to be expected from the supporters (and players) in a normal, non-disturbed game, in order to detect the abnormal, potentially dangerous situations as soon as possible.

As a process, a soccer game has a well-defined temporal structure, which implies consequences for the temporal and spatial knowledge management: the teams are supposed to play for well-defined durations at well-defined locations, one of the main distinctions being between the two periods. Each event has to be inserted in this pre-existing frame, and part of the reasoning and decisions to be taken will depend on the particular temporal environment of this event.

At the present stage (first beta tests), the system is still in development. Nevertheless the decision concerning the choice of formal methods and tools have had to be taken at the earlier design phase. Among the planned functionalities of the POLINT-112-SMS system is e.g. the visualization of hypothetical events resulting from possible decisions of the emergency staff (decision makers). It is quite clear that such hypothetical considerations about the possible future events resulting from not yet implemented decisions must take into account the partial or/and fuzzy character of the spatio-temporal knowledge about events.

As noticed above in this paper, many qualitative formalisms allow the representation of partial knowledge about relations between events, the formalism we have proposed being one of them. We are currently devising suitable ways of visually representing fuzzy or partially determined information.

8 An example of the use of the XRCDC in the POLINT-112-SMS

Let us give a short example of the way the XRCDC is applied in the POLINT-112-SMS system (for more details cf. also Osiński, 2009). We consider the soccer playing field

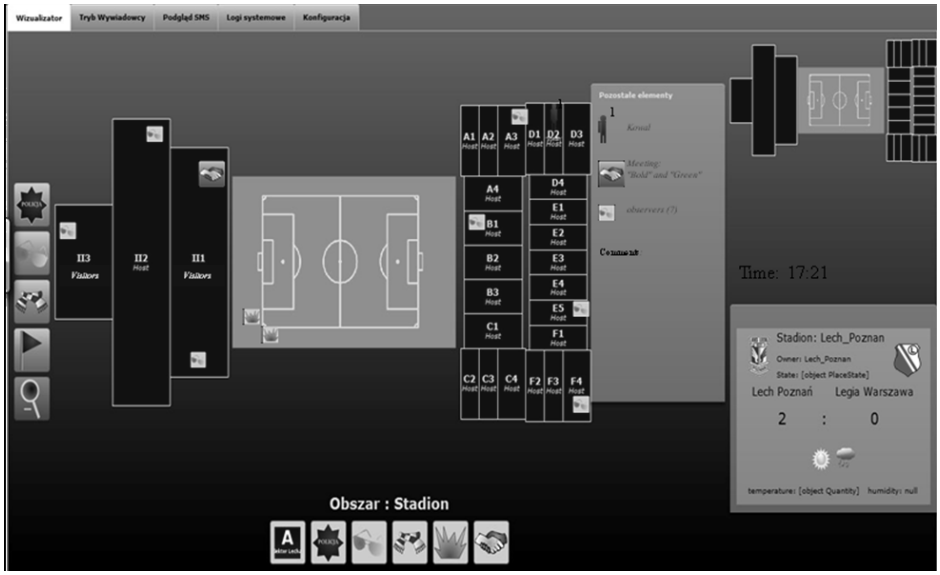


Fig. 8. The active map of the POLINT-112-SMS system (as designed by P. Kubacki and K. Witalewski).

with two adjoining zones, sector A and sector B. Here the spatial dimensions are North and East. The temporal dimension T constitutes a third dimension.

Suppose that a message sent by an informer introduces two new events, a meeting and a fight: *The meeting took place in front of sector A (between A and the soccer ground), close to sector A. The fight started just after the meeting, to the south of the meeting place, and close to it.*

Assume that, because of its knowledge of the layout of the situation, the system can interpret correctly the relative spatial expression *in front of*. It will use in its knowledge base *dir(E,N)(X,Y)*, *dir(E,T)(X,Y)* and *dir(N,T)(X,Y)* arrays for all pairs of objects X and Y.

Let us assume that another informer sends a new message asking for information: *I am in sector B. Where did the meeting take place?* In order to answer correctly, the system has to compose its knowledge about the (static) relation of sector B with respect to sector A with what it knows about the fight. After a successful computation, it could answer: *The fight took place to the north-west of sector B.*

Fig. 9 presents the spatio-temporal situation as it is represented in the system using the XRCDC. The graphical interpretation of the knowledge about the events collected in the system consists of (a) the projection on the north-east plane, (b) the projection on the time-north plane, (c) the projection on the time-east plane.

Acknowledgments

The research presented in this paper was partially covered by the on-going Polish Government research grant MNiSZW nr R00 028 02 "Text processing technologies for

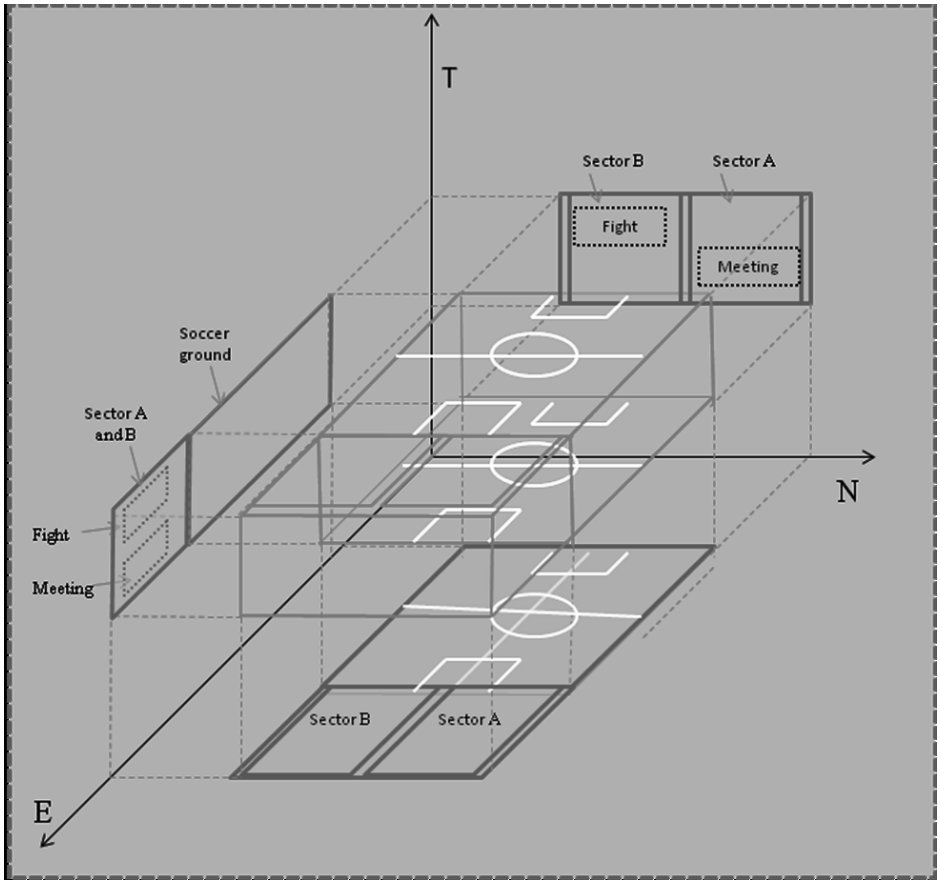


Fig. 9. Three projections of a spatio-temporal situation involving two events

Polish in application for public security purposes" (2006-2009) within the Polish Platform for Homeland Security. Thanks are due to J. Osiński, P. Kubacki and K. Witalewski for their contribution.

References

1. Allen, J. F.: Maintaining knowledge about temporal intervals. Comm. of the ACM 6(11), pp. 832–843 (1983).
2. Balbiani, P., Condotta, J.-F., Fariñas del Cerro, L.1998. A model for reasoning about bidimensional temporal relations. In: Cohn, A. G., Schubert, L.; Shapiro, S. C. (eds.) Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98), pp.124–130, Morgan Kaufmann Publishers (1998).
3. Balbiani, P., Condotta, J. F., Fariñas del Cerro, L.: A new tractable subclass of the rectangle algebra. In Proceedings of the 16th International Joint Conference on Artificial Intelligence, pp. 442–447, Morgan Kaufmann Publishers (1999).

4. Frank, A. U.: Qualitative spatial reasoning about distances and directions in geographic space. J. of Visual Languages and Computing 3:pp. 343–371 (1992).
5. Gärdenfors, P.: Conceptual Spaces: The Geometry of Thought. Cambridge/London: Bradford/MIT Press (2004).
6. Goyal, R. K., and Egenhofer, M. J.: Similarity of cardinal directions. In C.S. Jensen et al. (Eds.), SSTD 2001, LNCS 2121, pp. 36–55 (2001).
7. Guesgen, H.: Spatial reasoning based on Allen's temporal logic. Technical Report TR-89-049, ICSI, Berkeley, Berkeley,CA (1989).
8. Ligozat, G.: Reasoning about cardinal directions. J. of Visual Languages and Computing 9(1), pp. 23–44 (1998).
9. Ligozat, G., Nowak, J., Schmitt, D.: From language to pictorial representations. In Proceedings of the Language and Technology Conference (L&TC'07), Poznań, Poland, September 5-7, pp. 207–210, Wyd. Poznańskie (2007).
10. Osiński, J.: Extending the Cardinal Direction Calculus to a Temporal Direction. In: Proceedings of the FLAIRS 2009 Conference, Sanibel Island, Florida, USA, pp. 141-142 AAAI Publ. (2009).
11. Skiadopoulos, S., Koubarakis, M.: Composing cardinal direction relations. In: C.S. Jensen et al. (Eds.): SSTD 2001, LNCS 2121, pp. 299–317 (2001).
12. Skiadopoulos, S., Koubarakis, M.: Composing cardinal direction relations. Artificial Intelligence 152(2):143–171 (2004).
13. Vetulani, Z., Marciniak, J., Konieczka, P., Walkowska, J.: An SMS-based system architecture (logical model) to support management of information exchange in emergency situations. POLINT-112-SMS project. Intelligent Information Processing IV (Book Series: IFIP, Collection: Computer Science), Vol. 288/2009, pp. 240–253, Springer-Boston (2008).

Colophon

This proceedings were produced from the authors' electronic manuscripts. Following the guidelines, the authors mostly prepared their papers in Microsoft Word.

Contributions were converted into LaTeX by Adam Rambousek, then adjusted into the uniform markup of Springer LLNCS style and custom-written TeX macros prepared by Petr Sojka for the TSD conference. All articles were processed by the proceedings editors in Brno, namely Dana Hlaváčková, Aleš Horák and Adam Rambousek. Before final typesetting, the papers were sent back to the authors for checking.

The main editing, typesetting and proofreading steps were undertaken at the Natural Language Processing Center of the Faculty of Informatics, Masaryk University in Brno.

The proceedings editors thank sincerely all the authors for their contributions and the help with checking the final typesetting and thanks go to everybody who was involved in the book production. Without their hard and diligent work the proceedings would not have been in such a good shape and ready on time to be delivered to the honoree during the TSD 2009 conference in Pilsen.

Brno, August 2009 Aleš Horák

Acknowledgements

This work has been partly supported by the Ministry of Education of Czech Republic within the Center of basic research LC536.

Author Index

After Half a Century of Slavonic Natural Language Processing

D. Hlaváčková, A. Horák, K. Osolsobě, P. Rychlý (Eds.)

Published and printed by Tribun EU s.r.o., Gorkého 41, 602 00 Brno, Czech Republic

First edition, 2009

ISBN 978-80-7399-815-8