



VEDA
Vydavateľstvo Slovenskej akadémie vied

Slovenská akadémia vied
Jazykovedný ústav Ľudovíta Štúra

Computer Treatment of Slavic and East European Languages

Third International Seminar
Bratislava, Slovakia, 10–12 November 2005
Proceedings

Editor
Radovan Garabík

Reviewers
Peter Ďurčo
Jana Levická



VEDA
Vydavateľstvo
Slovenskej akadémie vied
Bratislava 2005

© by respective authors
The articles are licenced under the Creative Commons Attributions-NoDerivs
2.5 license



Slovak National Corpus
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia 2005
<http://korpus.juls.savba.sk/~slovko/>

ISBN 80-224-0895-6

Table of Contents

Opening Speech <i>Slavomír Ondrejovič</i>	7
The Role of Online Glossaries in Translating Investment Banking Terminology <i>Magdalena Bielenia</i>	9
Conjugated Infinitives in the Hungarian National Corpus <i>Gergely Bottyán and Bálint Sass</i>	27
Search Engine for Information Retrieval from Speech Records <i>Michal Fapšo, Petr Schwarz, Igor Szóke, Pavel Smrž, Milan Schwarz, Jan Černocký, Martin Karafiát and Lukáš Burget</i>	31
A Rule-Based Analysis of Complements and Adjuncts <i>Kata Gábor and Enikő Héja</i>	37
Levensthein Edit Operation as a Base for a Morphology Analyzer <i>Radovan Garabík</i>	50
Manual Morphological Annotation of the Slovak Translation of Orwell's Novel 1984 – Methods and Findings <i>Radovan Garabík and Lucia Gianitsová-Ološtiaková</i>	59
Contribution to Processing of Slovak Language at DCI FEEI TUKE <i>Ján Genči</i>	67
Towards a General Model of Grapheme Frequencies for Slavic Languages <i>Peter Grzybek and Emmerich Kelih</i>	73
DaskaL – A Web-based Application for Foreign Language Teaching <i>Kjetil Raa Hauge, Svetla Koeva, Emil Doychev and Georgi Cholakov</i>	88
Aspects of an XML-Based Phraseology Database Application <i>Denis Helic and Peter Ďurčo</i>	99
VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech <i>Dana Hlaváčková and Aleš Horák</i>	107
Orwell's 1984 – Playing with Czech and Slovak Versions <i>Jaroslava Hlaváčková</i>	116
Czech Language Parsing Using Meta-Grammar Formalism with Contextual Constraints <i>Aleš Horák and Vladimír Kadlec</i>	124

Analysis of Rule Based Phonetic Transcription Technique Applied to Slovak Language <i>Jozef Ivanecký</i>	130
Construction of Spoken Corpus Based on the Material from the Language Area of Bohemia <i>Marie Kopřivová and Martina Waclawičová</i>	137
Multimedia Reading Book – Utilization of an XML Document Format and Audio Signal Processing <i>Marek Nagy</i>	141
Morphological Idiosyncrasy in Hungarian Multiword Expressions <i>Csaba Oravecz, Viktor Nagy and Károly Varasdi</i>	147
Valency Frames and Semantic Roles in Czech <i>Karel Pala</i>	156
Question Answering in Polish Using Shallow Parsing <i>Dariusz Piechociński and Agnieszka Mykowiecka</i>	167
In Search of the Best Method for Sentence Alignment in Parallel Texts <i>Alexandr Rosen</i>	174
Word Tests for Speech Understandability Evaluation in Slovak <i>Milan Rusko and Marián Trnka</i>	186
Bulgarian and English Semantic Dictionaries for the Purposes of Information Retrieval <i>Max Silberztein and Svetla Koeva</i>	193
Slavic Text Taggers Project <i>Danko Šipka</i>	203
Multi-Word Named Entity Recognition in Polish Texts <i>Dominika Urbańska and Agnieszka Mykowiecka</i>	208
Creating of Slovak Electronic Phonetic Dictionary for Use in Speech Recognition <i>Pavol Vančo and Marek Nagy</i>	216
Russian Historical Corpora of the 18 th and 19 th Centuries <i>Victor Zakharov</i>	220
Building a Pilot Spoken Corpus <i>Jana Zemljarič Miklavčič and Marko Stabej</i>	229
<i>Appendix</i>	241

Opening Speech

Ladies and gentlemen,

I almost feel like saying that Slovko has come to Bratislava once more. In the beginning, as we all well know, was Slovo – the Word. Slovko – an event that has quickly become a tradition – can be viewed as its more familiar continuation here in Bratislava. While our inaugural linguistics conference, held in 2001, focused solely on the languages of the former Czechoslovakia (i.e. the proceedings of this conference, edited by Alexandra Jarošová, were entitled Computer Processing of the Slovak and Czech languages), the subsequent meeting was enhanced by other languages and, as a glance at the programme suggests, this trend is being maintained.

For some time now, the purpose of the conference has been not only to meet the need for mutual acquaintance and briefing with regard to each other's findings. It is also that Slovakia has succeeded in creating for itself favourable conditions within corpus and computer linguistics; Slovak linguists have made a successful entry into the international framework of these branches and have contributed to their development, so that their results – I trust you will not find my words too immodest – can be treated as substantive. My claim is to be confirmed by our scientific meeting, which holds out the promise of extremely interesting papers and discussions. I hope that I will not be divulging an official secret by saying that the outcomes and the work itself of the Slovak National Corpus team have not gone unnoticed by the higher authorities, who have decided to award them the prestigious Science and Technology Prize. The team are to receive their award from the Minister of Education of the Slovak Republic tomorrow evening.

On behalf of the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences, I wish you a stimulating discussion and I hope that you will find our “Central European” conference pleasant and fruitful. Now, it is my privilege to declare Slovko 2005 officially open.

Bratislava, November 8th 2005

Slavomír Ondrejovič
Director, Ľudovít Štúr Institute of Linguistics

The Role of Online Glossaries in Translating Investment Banking Terminology

Magdalena Bielenia

University of Gdańsk, Institute of English,
angmb@univ.gda.pl

Abstract. The aim of this paper is to show how three areas of study, such as computers, investment banking and translation are interrelated in our times. Hence, the brief characteristics of each domain is discussed, taking into consideration the features which determine the way investment banking terms are rendered into Polish. The topic of translating investment banking vocabulary has been of the author's interest for many years. However, due to the limits of this presentation, only the role of online glossaries in providing the reader with the target equivalent is presented.

Our discussion on translating investment vocabulary in the era of computers should be commenced with a brief explanation of the term *investment banking*. Although the first traces of investment banking date back to ancient times, both the term and the concept as such are quite new in modern economics. What should be stressed, however, is the diversity of what can belong to the area of investment banking. Let me quote some definitions to present this multifaceted phenomenon in a greater detail. The explanation of the term provided by *Business.gov* – the official business link to the US. Government – goes as follows: “Businesses specializing in the formation of capital. This is done by outright purchase and sale of securities offered by the issuer, standby underwriting, or “best efforts selling”.

As the purpose of the article is to discuss the role of investment banking and its terminology in Polish we will quote Polish scholars and their perception of the concept. According to Walkiewicz (11-12), there are many ways to define investment banking. The first method is to treat it as all the activities of the companies on the financial market. Others perceive investment banks as institutions which activities are connected with financial market, thus this point of view concentrates entirely on financial and capital market activities. As far as the third scope of interest is concerned, investment banks offer such services as underwriting, private placement, organizing mergers & acquisitions and corporate finance. In the fourth definition, investment banking is connected only with underwriting and private placement. The way we perceive investment banking is also dependant on the model of banking itself. As it is discussed by Pomorska (147), investment banking in Poland is rather similar to the continental model, i.e. the model popular in most countries of Western Europe. Contrary to the American Model, the modern bank in Europe, apart from offering deposits and credits can also take care of customers' financial resources by selling different investment products. With Poland in the EU, this model was also adopted by our country before the transformation process in order to abide by the rules of the community. Our banking law allows one to set up specialist banks. It should also be stressed that, according to

the *Banking Law in Transition*, Poland belongs to a group of countries with high compliance with international standards. Hence, Polish investment banking, although still different from the western concept, meets the requirements of international banking associations. The next issue which bears responsibility for the vague boundaries of the term investment banking is the lack of Polish investment banking law as such. The institutions and products are supervised in the following regulations: the Banking Law, the Bond Law or the Law on Public Trading in Securities of Aug. 21, 1997, the Warsaw Stock Exchange Statutes, to mention just a few.

However, the difficulty for investment banking in Poland is not connected entirely with the impreciseness of its definition. The more outstanding reason for the problem in comprehending its goals is the novelty of services and products offered by investment banks. For example, the Warsaw Stock Exchange was absent on the financial market for many years and started to operate again in 1991. Other banking products appeared in the late nineteen nineties, whereas many of them are still undergoing the process of implementation. The emergence of investment banking in Poland during the twilight years of the twentieth century is connected with the change of the system in Poland, namely for the market-driven economy. As Kwieciński (96) says, "In 1989, Poland has embarked on a rapid sociocultural transformation away from authoritarian statism towards liberal-democratic capitalism. Major social, cultural and political tendencies of this transformation that are characteristic of the post-communist Central European countries at large may be listed as follows: (1) the emergence and growth of market economy to replace command economy; (2) the growth of the private sector and of the middle class; (3) the emergence of a constitutional, lawful state; (4) the growth of democratic institutions, (5) a new definition of the national community; (6) growing discrepancies in the distribution of wealth". Due to the mentioned above change, after 1989 new institutions connected with banking, financial and investment services started to be available for Poles. As a result, many new investment products began their existence on the Polish market. Being a new phenomenon, investment banking was offered exclusively by the multinational leviathans of banking and financial sector. Thus, there was a need to translate the names of products for the Polish receivers. Foreign websites and product information as well as leaflets required Polish versions immediately. Hence, an immense task was set for the translators to render requested materials quickly and precisely. As has been stated above, both the terms and the concepts themselves were not known in the Polish reality and thus they were absent both in the Polish terminology as well as in people's mental lexicon. It was the translators's task to find the right equivalent or coin one in the case of its absence.

To digress, the same situation can be observed with other nations which have faced the change of systems in their countries. Let me concentrate on the neighboring ones. The role of English in shaping the current banking activity in Lithuania is discussed by Marina and Suchanova. As they (10) say, "the revival of economic and business contacts with overseas partners demanded good command of new economic terms, particularly in English as their source language". Hence, the books and other educational materials as well as commercial prospects are in the lingua franca of today. What is more, it is the English text-book which is used for teaching courses in business. Moreover, several English economic publications like *The English Dictionary*

of Economics have also been translated into Lithuanian. As Marina and Suchanova claim (10), some textbooks have caused some difficulties as they contain economic words which are unknown in Lithuanian.

Let us now turn to the second aspect of this research- translation studies. However, for the purpose of the article, we will focus our attention only on aspects connected with the symbiosis of investment banking, translation and computer science. Concentrating on Poland we should quote Piotrowska who says that "translation has been receiving more recognition in Poland recently. Some of the reasons for this expanding interest reflect the dramatic changes in the international arena of politics and sociology. Poland opened up its political, economic and cultural doors to the Western market, mindset and thought. Better and easier access to Western goods and ideas, economic exchange based on the principles of the free market and free enterprise, also in the field of publishing, are all landmarks of the new Polish reality". It should be stated that before the great transformation it was literature which had gained attention of the scholars, whereas translation of specific texts (which of course existed) had been less widespread than the one of literary works. However, with the change of the system in 1989 and with Poland in the EU the translation of LSP texts is of crucial importance both for the translators as well as for the theoreticians of translation practice.

The relation between language and economics has been of interest for many years. As Felber and Budin (67) state, this topic was very popular in 1920 among the scholars such as Fehr, Messing and Schirmer who used to teach at economic high schools in Tschechoslovakia. The scientists gathered round the Prague linguistic school opted for separating the language of economics as a type of functionally structural linguistics. Thus, the concept of LSP was researched with regard to the functional linguistics of the language of economics.

The way business and translation are interrelated is nowadays the subject of deep studies in many companies and scientific organization. For example, a language translation system for international business communication is worked on by scholars from Carnegie Mellon University. Their software with the purpose of visual representation is supposed not only to translate the utterance but also to mimic the facial expressions of the speaker. The system has a photo of the speaker and studies how he behaves during speech production, with special attention to the parts near his eyes and mouth. The scholar Jackie Fenn quotes the studies showing that when we see somebody speaking we understand him better than when we only listen to his speech. Hence, this project is very important in intercultural communication where the actual presence of the speaker is not possible due to physical distance between firms.

Many international companies devote a great deal of time and energy to make their products worldwide available. Let us choose Laplante (8) and his example of HP to present how this company, with the use of the mentioned below activities, deals with the global publishing. The process consists of three parts:

- 1) Regionalization – “adapting content to a geographic region”
- 2) Translation – “transforming content from one language to another”
- 3) Localization – “modifying content to account for differences in distinct markers, with cultural sensitiveness and legal requirements in mind”.

All the mentioned elements which are responsible for international success need translation and computerization in order to be effective. The knowledge about specific countries must be gathered by a specialist and later stored to be used when needed. The second element-translation, especially in the case of repetitive elements, is quicker and more precise when machine translation, translation memory or word benches are used. As far as localization is concerned, as in the case of regionalization, useful data can be stored or online databases can be employed to find information on legal system or cultural differences. Let us discuss briefly how some computer tools can help in doing business internationally by effective and precise translation adjusted to the needs of the target audience. Translation demand plays an important role in the quality of translation. Hutchins distinguishes four types of translation demand as far as MT and translation tools are concerned. The first one comprises texts which should be of publishable quality, no matter whether they will be printed and sold or will just circle around in the company. In the second group we have less demanded people, mainly users who only intent to understand the content of the message for their own purpose as quickly as they can thus the quality can be lower. The third group consists of one-to-one communication (via telephone or written type) or unscripted presentation as can be found in diplomacy. The last one deals with multilingual systems of information exchange. He also discusses that MT systems are used for many companies for technical documentation. In this way, for example by using the *Logos* systems many companies can carry on even outstanding translation projects. Among *Logos* users we can find Ericsson and Osram when enterprises like Berlitz, Ford or General Motors use Systran. Philips and UBS, for example, deal with the *METAL German-English system*. "Companies, in order to maintain the same terminology regardless of the flow of time and use of agency, need automated system which will store the database." It should also be stated that terminology management is a type of language technology that involves standardizing the methods how key terminology is employed in a multinational organization. "Companies that insist on good branding often develop glossaries that define how certain product, technical or medical terms should be used". As Schwartz and Toon say, "this standardization helps create a consistent brand, and is important to meet regulatory requirements and to prevent legal problems by using incorrect terminology". Stroehlein draws our attention to the speed of translation which is so important in our times. "Taking all that into account, professional-quality translation of a 1,000-word article typically takes four or five hours. That is a significant amount of time in the minute-by-minute world of online publishing". Schwartz and Toon also pay attention to the reduction of expenses in translation process. "Translation memory cuts costs by reducing the total number of words that require translation. Companies typically can save between 40-75% of their translation costs by using translation memories". Not only TM but also systematized terminology also saves costs in the long run. As we can read on the website of *dl.com* | *documentum.com*, "it has been shown to reduce customer support costs by helping to produce content that is more understandable to customers. It also prevents legal problems by avoiding the misuse of terminology or translations that can result in legal suits. Finally, good terminology management can improve search results on the Web site when customers are looking for correct information. Their ability to find what they need on the Web reduces the

number of calls to a support call center with their associated costs". What is more, as Fouzie says, "More importantly, researchers also say that people are three times more likely to buy something online when addressed in their own language". Thus, from the marketing perspective, translation is required when we want to operate internationally and have high revenues. Laplante also mentions (9) some other problems caused by terminological shortcomings. "Inconsistencies in published information reduce customer satisfaction and increase the risk of diminished customer loyalty". However, translation can be also imperfect because something else goes wrong. Laplante (9) draws our attention to the other aspect of translation within a company when it is not successful. He gives the example of HP and its problem with translation. "Translation was inefficient due to redundancies. Two factors were in play. First, some regional translation agencies were not using translation memory, starting from scratch each time a new translation cycle was required. Second, there was no leverage across the translation memories that did exist; they were isolated project silos.

- Lack of consolidated reporting tools meant that there was no visibility at the corporate level into spending and reuse. Even without metrics, HP knew that it was losing time and money.
- Corporate messaging and branding were inconsistent across regions, causing customer confusion.
- Coordinating product releases across regions was a major challenge because of the need to synchronize multiple websites in multiple languages. The demand for and interest in a solution to these problems quickly turned Toon's part-time".

To digress, computers are also important in translator's activities outside the translation process itself. Those who need to advertise their products should not forget about computer in their pursuit of offers. Lamensdorf underlines the role of computer accessories in freelance translator's job. He draws our attention to creating bilingual websites which will be available by using search engines. The last can be also very useful when looking for some translation agencies in order to offer them translation services. However, not all translators are fans of IT and computerization. As Champollion says, many translators are afraid that machines will replace them. That is why a group of translators is against any tools and prefers implementing the standard procedure.

However, in my paper I would like to concentrate on the Internet as the source of help for translators. As has been stated above, investment banking is a new phenomenon. Thus, there is not much printed specialist literature which will accompany the process of translation. What is more, as Baker (xiii) says about her own book, "all encyclopedias, this one included, are inevitably out of date before the hit the press-such is the nature and speed of intellectual progress in any field of study". In this case it is the Internet which can serve as a library for those who need explanations or equivalents. The Internet sources can be divided according to different criteria. The following typology has been adopted in this research. The gathered material has been

divided according to the following criteria: the English sources, monolingual and multilingual, and the ones in Polish or at least with terms in Polish. Due to the limit of the paper only glossaries and dictionaries have been taken into consideration. It should also be stressed that due to mainly two reasons the paper suffers from a number of potential shortcomings and provides an imperfect measure of the discussed topic. First of all, there are so many international companies that for obvious reasons all of them cannot be taken into consideration for the purpose of this study. Secondly, there are more and more banking and financial institutions appearing on the market and the ones which are already present have been changing the scope of their activities as well as their websites. However, the research can serve as a useful source of information on online glossaries. It may be an interesting idea to check how the online dictionaries and glossaries differ with the flow of time. The author has researched this topic on the basis of financial portal plus the search engine. The data represent the state of websites as for November 2005 as the research was conducted in October and November 2005. First, the difference between glossaries and dictionaries should be stressed. Let us quote some definitions of glossaries taken from the search engine *Google*.

- “An alphabetical list of technical terms in some specialized field of knowledge; usually published as an appendix to a text on that field” (www.wordnet.princeton.edu/perl/webwn)
- “Short list of words related to a specific topic, with brief definitions, arranged alphabetically and often placed at the end of a book” (www.usd.edu/library/instruction/glossary.shtml)
- “An alphabetical listing of special terms as they are used in a particular subject area, often with more in-depth explanations than would customarily be provided by dictionary definitions” (www.brochure-design.com/brochure-design-publishing-terms.html)

On the other hand, dictionaries are supposed to be larger in size. This distinction is important as many data providers given in this study tend to call a short list consisting of 14 terms a dictionary. It can be noticed, for example, by studying the table of Polish banks and their glossaries.

Let us quote some of them to see how the glossaries and dictionaries work in practice. In the first part of the research the ones provided by foreign institutions will be examined. An attempt will be made to prove that English monolingual sources as well multilingual dictionaries with no Polish entries are of great help for translators. When you visit the UBS website you can come across different glossaries. They are as follows: *UBS Dictionary of Banking, Mortgages (in Switzerland), Mortgages (in the U.S.), UBS Funds, Investment terminology, Foreign exchange, Trade & Export Finance*. Let us concentrate on *Investment terminology*. As the name suggests, this glossary consists of words connected with investment activities. They are presented in the alphabetical order and you can access them by clicking on the letter which comprises the vocabulary gathered under this heading plus the explanation which accompanies each entry. A very useful tool in translating investment banking terms is *UBS*

Dictionary of Banking. It was revised in spring 2005 and it contains the up-to-date vocabulary. It embraces definitions of more than 2400 terms. Each English term is accompanied by its equivalents in German, French and Italian plus a descriptive definition in English. You can search the dictionary by clicking on the letter and going up or down the list.

As far as glossaries are concerned, the next website worth considering for those interested in translating investment banking terminology is the one of *Glossar Trade & Export Finance*. It is especially useful for those who need deeper understanding of English terms connected with the financial aspect of trade and export. Most of the mentioned above glossaries are offered in such languages as German, French and Italian. Other bank-*La Salle* offers three glossaries in English-*Glossary of Banking Terms*, *Glossary of Investment Terms* and *Glossary of Home Equity Terms*. All the words are listed alphabetically and can be accessed by pressing the right letter. As far as the range of categories is concerned, we should draw attention to the *ADVFN glossary* which comprises 27 categories, such as Banking & Finance or Futures & Forwards, Investment Trusts & Mutual Funds, to name just a few. The terms as in the mentioned above cases can be accessed by clicking on the needed word. *Societe Generale* also has the glossary of English trade terms in the alphabetical order. *Unicredito* also offers its customers a short glossary consisting of the most important English terms connected with banking. Other online dictionaries can be viewed on the website of *Glossarist*. It can be used as a search engine for finding demanded glossaries. I have chosen, in my opinion, the most useful ones in the job of translator. *Barkley's Comprehensive Financial Glossary* is a very detailed glossary (<http://www.oasismanagement.com/glossary/>).

The same applies to the tool provided by *Money World* (<http://www.moneyworld.co.uk/dictionary>)

and the one compiled by *Investor Words* (<http://www.investorwords.com/cgi-bin/letter.cgi?a>).

Some dictionaries deal with the part of investment banking like *Offshore Banking and Trading Glossary* (<http://www.turtletrader.com/og.html>) where the words connected with offshore funds and this type of investment are taken into consideration. Unfortunately, this dictionary is not available in any languages of the countries which have recently joined the EU. The words are to be found in many different sources but are not compiled as in the case of the English glossary. *Nasdaq* also provides *Glossary of Terms* which can be found at the following internet address: http://www.nasd.com/web/idcplg?IdcService=SS_GET_PAGE&nodeId=1088&ssSourceNodeId=766. For the translators who know German other websites can be useful. One example can be *Finanzlexicon* where the terms connected with investing are given and described in German. On the other hand, French users can stick to MIEUX COMPRENDRE LE LANGAGE BOURSIER. This glossary is divided into some thematic sections like *Lexique boursier* (the lexicon of stock exchange) or *Lexique des obligations* (the lexicon of bonds). For those who work with investment vocabulary the dictionaries provided by *Investopedia* are of great help. The terms are arranged in alphabetical order plus you can browse by category (Acronyms, Active Trading, Bonds, Buzz Words, Financial Theory, Foreign Exchange, Fundamental

Analysis (Accounting), Mutual Funds, Options & Futures, Personal Finance, Real Estate & Property, Retirement Planning, Stocks, Taxes, Technical Analysis, Venture Capital and IPO's. What is more, you can view *Recently Added Terms* and *10 Most popular Terms*. The translators who specialize in certain types of investment banking instruments should use specific glossaries. One of them is *Glossary of Municipal Bond Terms* **eMuni**. The following glossary of municipal bond industry terms and jargon has been prepared for individual investors by Zane B. Mann, publisher of *California Municipal Bond Advisor*. It contains about two hundred terms dealing with municipal bonds. Another interesting online glossary is provided by *Investor Words*. This company offers thousands of terms which you can access in many ways. You can browse by category (Accounting, Banking, Bonds, Brokerages, Currency, Dividends, Earnings, Economy, Futures, Global, Insurance, Investor Relations, IPOs, Law/Estate Planning, Lending/Credit, Mergers/Acquisitions, Mutual Funds, Options, Public Companies, Real Estate, Retirement, Stocks, Strategies, Taxes, Technical Analysis, Trading, Venture Capital) or by letter. What is more, they keep track of the recently improved terms and you can receive the term of the day by email thus you can improve your knowledge without leaving your room. The next tool is prepared by *Speculative Bubble*. You can search it by typing the word you need in a search box or clicking on the right word on the list. Moreover, those interested in investment banking can read interesting articles on a particular topic where they can find useful vocabulary as well. As far as specialist knowledge is concerned, there are also dictionaries devoted entirely to one particular domain of investment banking. Let me quote some examples. The *Glossary of option terms* is prepared by the Chicago Board Options Exchange. There are almost two hundred words devoted entirely to this type of financial instrument. For those who have to translate the terminology of derivatives *The Numa Dictionary of Translating Acronyms* will prove useful. There are about six hundred acronyms of stock exchanges and other institutions as well as words connected with derivative instruments. Derivatives are also handled by *The William Margrabe Group, Inc., Consulting, THE DERIVATIVES 'ZINE™* Translators interested in stocks vocabulary should use *The ChartFilter's glossary*. There you can find different names of stocks (by clicking on the letter or inserting the needed term in a search box). As the topic of our research deals with translating investment banking terms into Polish, we should also present the resources available in this language. With Poland in the European Union, there are many tools offered by the institutions which are responsible for smooth communication between countries. There are dictionaries and glossaries prepared by the EU which can be useful for anyone dealing with specialized vocabulary. For example, *Eurodicautom* is a multilingual term bank. This online dictionary consists of five and a half million entries. You can find the needed equivalent of a specialized term in twelve languages. Unfortunately, it is not available in Polish but its successor is to contain Polish terms as well. Another useful tool for translators is the *Eurllex* base. Although in our research we should concentrate only on dictionaries and glossaries, this service is very useful in searching for the right phrase or word. There you can find the regulations concerning European Union law. It is especially useful for the translators of financial terminology as the documents are available in all the languages of the EU. The translator when trying to find the right term can use the bilingual display

which guarantees a rapid access to the right word. The translator can find the corresponding equivalent in the target language. It is possible to search this service by field or by inserting the keyword. The base, being available in Polish, can be very useful especially for those who translate legislative acts and need legal language vocabulary. On the website of *Lexicool* we can find some sources which can be of immense help for the translators of banking and financial texts. Let us examine them in our study. The first source is the website of terminological data of the *European Integration Committee*. This base comprises terminology which was compiled when different legal acts of the EU were translated into Polish. There you can find some glossaries, such as *Economy, finance, money*, which is specialized for those working with English and Polish, although certain French and German equivalents are also provided. By inserting the exemplary term *bond* in the search box the translator receives all the phrases in which this word can be found. There are also equivalents in German and French for those who translate in other languages as well. It is especially useful in such languages as Polish as the terms are quite new in our economic reality. What is more, the user can rely on materials published by such institution as the EU for precision and adequacy. The next electronic dictionary is called *Financial Terminology (EN>PL)*. In this one we can find the English terms connected with the area of finance, being arranged in the alphabetical order. The English terms are accompanied by a descriptive definition in Polish plus a Polish equivalent (if such exists) is provided in brackets. This form is especially useful as some terms are new in Polish or their equivalents are absent in Polish terminology. The third one is called *Brian Huebner's Economic and Financial Terms Glossary (PL>EN)*. Both Polish and Czech are available online. On the left side there are the Polish terms presented in the alphabetical order whereas on the right side of the column we can find the English equivalents. Unfortunately, in the version presented by *Lexicool* there is no possibility to browse the dictionary. In order to find the needed word you have to scroll down the list. The next dictionary is *Macroeconomics dictionary*. It consists of four pages of vocabulary for those who need basic macroeconomic terms. However, it may prove insufficient in case of more detailed translation needs. Unfortunately, it is not available in Polish. The next dictionary is called *APFA-40 mots-cles des affaires en 27 langues (MULTI)*. The translator can choose one of 27 foreign languages to find the right equivalents from French in the target language. On choosing Polish, the translator deals with Polish equivalents of the newest 40 financial words. Unfortunately, some of the 40 words are not translated into Polish. A very functional dictionary/translator is provided by <http://www.angool.com/>. It comprises more than 353 000 words. There are also terms belonging to banking and finance. You can insert the needed term in Polish or in English and the machine translates for you. The next source, in my opinion worth considering, is the glossary provided by the National Bank of Poland (NBP). You can insert the Polish term in the given window or click on the letter you are interested in order to read the definition of the term. Also the Warsaw Stock Exchange (GPW) offers a mini glossary with the basic terms connected with investing. As far as Polish banks are concerned, the aim of the research was to find out how many banks offer glossaries and dictionaries for the users. In my opinion banks as a source of terms are very rarely taken into consideration. An attempt will be made to prove that bank websites can

provide the reader with the term he wants to find. In the mentioned below results 63 banks were examined. The list of banks from *Parkiet*, Polish financial, portal was used.

Table 1. Banks operating in Poland and their glossaries/dictionaries

Bank	Dictionary	Title	Entries	Remarks
ABN AMRO Bank (Pol-ska) S.A.	+	<i>Glossary The definitive guide to investment banking</i>	About 170 terms. There are three ways of finding the terms-using the search box, Clicking on the letter or going down the list	
AIG Bank Polska SA	+	<i>Słowniczek</i>	About 40 terms connected with insurance	www.amplicolife.pl
BPH	+	<i>Glossary-Sustainability Report 2002</i>	40 terms connected with banking activity and ecology	www.hvb-group.com (Nachhaltigkeitsbericht 2002 der HVB Group)
Bank Handlowy	+	<i>Glossary</i>	About 600 terms. First you click on the letter, then you click on the needed term	Useful terms and financial jargon www.citigroup.com
Bank Milenium	+	<i>Słownik</i>		No access
Bank of America Polska	+	<i>Glossary</i>	The glossary is divided into 6 parts-Smart Budgeting, Investment and Retirement, Life Insurance, Education Planning, Estate Planning, Other Terms. You can click on the letter or go down the list of terms-almost 100 terms in Investment and Retirement	www.bankofamerica.com Investment and Retirement-used with permission from <i>Baron's Dictionary of Finance and Investment Terms</i>
Bank of Tokyo-Mitsubishi	-			

Bank	Dictionary	Title	Entries	Remarks
Bank Pekao SA	+	1) <i>Słowniczek</i> 2) <i>Słownik rynku finansowego</i> 3) <i>Słownik Pekao 24</i>	1) About 35 terms connected with stock exchange. You should click on the term 2) The term <i>zlecenie</i> (order) is described 3) 23 terms connected with online banking	1 & 2 www.cdmp.ekao.com.pl
Bank Pocztowy SA	+	<i>Słownik pojęć</i>	About 70 terms. You can insert the needed term or click on the letter	Terms are connected with the offer of the bank. The customer may ask to add new terms
Bank Przemysłowy Getin Bank	+	<i>Słowniczek pojęć</i>	14 terms accessible by scrolling down the list	Terms connected with the activity of the bank
Bank Rozwoju Cukrownictwa	-			
Bank Zachodni WBK	-	<i>Słownik</i>	10 terms accessible by scrolling down the list	Terms connected with the activity of the bank
Bankgesellschaft Berlin (Polska) S.A.	-			
Bank Gospodarstwa Krajowego	-			
Bank Gospodarki Żywnościowej	+	<i>Słownik pojęć</i>	About 150 terms. You can click on the letter or use the search box	The customer can add his term
Bank Inicjatyw Społeczno-Ekonomicznych SA	-			
BNP Paribas Polska SA	+	<i>Glossary</i>	About 90 terms in English available by clicking the right letter	Website available in English or French
Bank Ochrony Środowiska	+	<i>Słownik</i>	6 terms connected with the activity of this institution	www.seb.pl

Bank	Dictionary	Title	Entries	Remarks
Bank Polskiej Spółdzielczości	-			
BRE Bank SA	+	<i>Słownik pojęć</i>	About 60 terms which you can access by going down the list or clicking the group of letter for a quicker search	e-MSP (http://www.e-msp.pl/172)
BRE Bank Hipoteczny	+	<i>Słownik pojęć związanych z kredytem</i>	About 40 terms connected exclusively with mortgage -you click on the term or read the list with their explanation below	
Bank Współpracy Europejskiej SA	+	<i>Słownik</i>	There is a dictionary with terms described in a particular section. A dictionary under each page.	
BZ WBK	+	1) <i>Słownik</i> 2) <i>Słownik</i> 3) <i>Carrers glossary</i> 4) <i>AIB Tradefinance - Glossary of Terms</i> 5) <i>Capital Markets Glossary</i>	1)A list of 14 terms connected with investment fund 2)Two terms connected with bonds 3)You have to click on the letter-about 30 terms 4)About 80 terms available by clicking on the letter 5) You can click on the letter or scroll down the list	1)Customer may add his term www.arka.pl 2)Customer may add his term www.inwestor.bzwbk.pl 3) www.aib.ie 4) www.aibttradefinance.com 5) http://www.aibcm.com/
Calyon Bank Polska SA	+	<i>Glossary</i>	You can click on the letter or go down the list. Almost 100 terms.	www.credit-agricole-sa.fr
Daimler Chrysler Services (debis) Bank Polska SA	-			
Danske Bank	+			Link to Ectaco services

Bank	Dictionary	Title	Entries	Remarks
Deutsche Bank	+	<i>Deutsche Bank Banking and Stock Glossary</i> <i>Deutsche Bank Bank-und Börsen Lexicon</i>	Almost 800 terms in English which can be accessed by selecting the letter or the term on the list. When you click on the letter, the explanation in English appears plus you can choose the German version by clicking on the German flag.	
Deutsche Bank PBC	-			Glossary offered by Deutsche Bank
DomBank	+	<i>Słowniczek pojęć</i>		It belongs to Getin Bank
Dominet Bank	-			
DZ BANK Polska	-			
Eurobank	-			
FCE Bank Polska	-			Ford
Fiat Bank Polska	-			
Fortis Bank	+	1) <i>Glossary</i> 2) <i>Glossary of terms</i> 3) <i>Definicje</i> 4) <i>Glossary</i>	1)This glossary is divided into three sections (A-M,F-M,N-Z).In each section there is a list of terms in English plus their equivalents in French, Dutch, German. Altogether about 50 terms 2) 17 terms in non-alphabetical order connected with insurance 3)About 50 terms connected with investment-in Polish and English plus an explanation in Polish 4) About 40 terms connected with insurance	1)www.fortisbank.be 2)www.fortisinsurancence.co.uk 3)www.fortisinvestment.com 4)www.fortisinsurancelu
Spółdzielcza Grupa Bankowa	-			
GE Money Bank	+	<i>Glossary</i>	About 100 terms can be accessed by pressing the letter or doing down the list—there is an explanation next to the term	www.—moneybasics.ch
Getin Bank SA	+	<i>Słowniczek pojęć</i>	14 terms connected with using cards	
GMAC Bank Polska SA	+	<i>Glossary</i>	The glossary is divided into three parts :Leasing Terms, Purchase Financing Terms. And Commercial Terms-altogether about 50 terms.	http://www.gmcanada.com
ING Bank Śląski	+	<i>Glossary</i>	About 70 terms which can be accessed by pressing the letter and then all the terms with their explanation appear	www.ing.—com
Inteligo	+	<i>Słowniczek</i>	17 terms connected with bonds	

Bank	Dictionary	Title	Entries	Remarks
Investbank	+	<i>Poradnik- karty płatnicze-Słownik</i>	A list of 13 terms connected with using cards	
Izzy Bank-Bankowość Detaliczna dla młodzieży BRE Banku SA	-			Part of BRE Bank
Kredyt Bank SA	-			
Lukas Bank	+	<i>Fundusz inwestycyjny-słowniczek pojęć</i>	11 terms on the list connected with investment funds	
Mazowiecki Bank Regionalny	-			
mBank	+	1) <i>Słownik pojęć</i> 2) <i>Słowniczek pojęć</i>	1)About 120 terms connected with stock exchange investments. You can press the letter or scroll down the list where the terms and their explanation appear 2)50 terms connected with investment funds	
MHB Bank Polska SA	-			
Multibank	+	<i>Słownik pojęć</i>	a list of 32 terms connected with credits	
Nordea	+	<i>Financial dictionary Nordea</i>	You can press the letter or insert the required term in the search box (about 213 terms)	www.nordea.lu
NORD/LB Bank Polska Norddeutsche Landesbank SA	-			
Nykredit Bank Hipoteczny S.A.	-			
PKO BP	+	1) <i>Słownik</i> 2) <i>Słowniczek pojęć związanych z kredytem mieszkaniowym i pożyczką hipoteczną</i>	1)8 terms connected with Basic banking activities 2)37 terms connected with credits for housing construction	
PTF Bank	-			
Rabo Bank International Polska	-			

Bank	Dictionary	Title	Entries	Remarks
Reiffeisen Bank Polska	+	1) <i>Glossary (available in Italian and German)</i> 2) <i>Glossary</i> 3) <i>Glossary of key terms and abbreviations</i> 4) <i>Raiffeisen Centrobank Glossary</i> 5) <i>Lexicon</i>	1) About 130 terms which can be accessed by pressing the right letter 2) 15 terms connected with investment funds 3) A list of about 60 investment terms in pdf version 4) You can click on the letter and later scroll down the list of altogether 460 terms 5) Multilingual term bank (German, French, Italian)	1) www.rcm.at Raiffeisen Capital Management 2) www.raiffeisen-capital.ru 3) www.rzgroup.com 4) www.rcb.at 5) www.raiffeisen.ch
Societe Generale	+	1) <i>Glossary</i> 2) <i>Glossary and definitions</i>	1) About 120 terms-you can click on the letter, on the term or on scroll down the list 2) About 150 terms from different banking areas	1) www.sg-tradeservices.com 2) www.sgserach.socgen.com
Svenska Handelsbanken	-			
Śląski Bank Hipoteczny SA	-			Consult www.ing.pl
Toyota Bank Polska	-			
Volkswagen Bank Polska	-			
Wschodni Bank Cukrownictwa S.A. w Lublinie	-			
WestLB Bank Polska S.A.	+	<i>Glossary</i>	A list of about 120 terms connected with banking	www.westlbcarriers.com

Source: the financial portal *Parkiet* and different websites of the banks which were taken into consideration.

After the research conducted on computer resources which can make the translation process of investment terms easy, we can claim that there are many, more or less advanced, glossaries or dictionaries which serve knowledge not available in the printed version. Due to the rapid development in the area of investment banking we can

count on the Internet which offers up-to-date information. However, in the case of Polish there are unfortunately not many bilingual resources (Polish-English). Even when there are dictionaries with Polish interface very often the system does not recognize the case inflection of adjectives and nouns. Thus, the one for whom Polish is not a mother tongue can have problems in finding the right equivalent. The translation device provided by www.angool.com is an exception. After inserting the term *obligacja* (bond) in other cases, *obligacjami*, *obligacjach* – it tries to find the phrases where these words occur. We can hope that more and more resources will be available in the future as there are still new instruments entering our economic reality. The shown above table comprising the glossaries of more than 60 banks operating in Poland proves that the bank's website is important in investment banking terminology. All the biggest banks in Poland, such as PEKAO SA, PKO BP, BPH, ING Bank Śląski, BZ WBK¹, provide the reader with a glossary. The ways the sources are described are different. Some banks call them *Słownik* (dictionary), others *Słowniczek* (mini dictionary). One bank opts for *Definicje* (Definitions).

It should be also underlined that in this study only some tools have been chosen. There are also frequency lists and concordances, to mention just a few (Schneider). What is more, there are also many glossaries and dictionaries available but the author has decided to choose the ones, in her opinion, which are of great help in the translation process.

As far as the improvement in translation resources are concerned, let me quote Nogueira who reminds the readers of a very simple translation tool which task was to fasten and organize the job of the translator. We can estimate that most readers do not remember this "complicated" device. The shoebox dictionary (it was called like that) consisted of some cards, some being blank, some incomplete and some complete. In most cases they were stored at random as the translator never had enough time to categorize them. However, with the omnipresence of computers these index-card glossaries are not very popular, being a reminiscent of the past translation procedures. Probably with the growing role of technology even the glossaries which have just been described would be treated as out-of-date in the future.

References

Books

1. Felber, Helmut, and Budin Gerhard. *Teoria i Praktyka Terminologii*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego, 1994. 67
2. Kwieciński, Piotr. *Disturbing strangeness. Foreignisation and domestication in translation procedures in the context of cultural asymmetry*. Toruń: Edytor, 2001.
3. Walkiewicz, Rafał. *Bankowość inwestycyjna*. Warszawa: Poltext, 2001.

1 Polskie banki pękają od pieniędzy. 6 Nov.2005.
<http://www.biznespolska.pl/wiadomosci/prasa/?cityID=warszawa&contentID=116625>

Articles

4. <http://www.translationdirectory.com/article389.htm>, 19 Oct. 2004.
5. Champollion, Iyes. *Machine translation (MT), and the future of the translation industry*. 20 Oct.2004. <http://en.lingo24.com/home.html>
6. Fouzie, James. *Lost in the translation*. 12 Feb. 2001 <http://www.infoeconomy.com>
7. Guy, Sandra. *Translation and Treatment*. 1 June, 2005. <http://en.lingo24.com/home.html>
8. Hutchins, John. *The development and use of machine translation systems and computer-based translation tools*. 10 Oct. 2004 <http://www.foreignword.com/Technology/art/Hutchins/hutchins99.htm>
9. Lamensdorf, Jose.H. *On-line research into freelance translator recruitment processes*. 3 Nov. 2004. <http://www.lingo24.com/>
10. Laplante, Mary. "Global Content Management: Hewlett-Packard Talks the Talk of Worldwide Business HP's digital content management initiatives optimize the delivery of product content to global markets". *The Gilbane Report January 2005*. 12 Oct.2005. <http://www.gilbane.com>
11. *Law in transition online 2005-Banking law in transition*. 2 Nov. 2005 <http://www.ebrd.org/pubs/legal/OL05b.pdf>
12. Marina,Valerija and Suchanova Jelena. "The Comparative Analysis of English Economic and Business Terms and Their Lithuanian Translation Equivalents". *Kalbu Studijos – Studies about Languages* 1(2001): 10-13
13. Nogueira, Danilo *From Shoebox to SQL*. Accurapid-The Language Service. 10 Oct. 2005. <http://accurapid.com/journal/31shoebox.htm>
14. Piotrowska, Maria. *Translation studies in Poland: onward and upward*. 23 Nov. 2004. Portal Unii Europejskiej. 10 Oct. 2005. http://europa.eu.int/comm/dgs/translation/bookshelf/tools_and_workflow_en.pdf
15. Schneider, Christof. *CAT tools: A brief overview about concordance software*. 19 Oct. 2004. <http://www.translationdirectory.com/article389.htm>
16. Shwartz, Howard and Toon, Alison. *Global Content Management: The 18 Most Frequently Asked Questions*. 10 Oct. 2005 http://www.gilbane.com/ctw/Trados_gcm_whitepaper.pdf
17. Stroehlein, Andrew. *The Online Babble Barrier*. 18 July 2002 <http://www.ojr.org/>
18. Szcześ, Małgorzata and Sebastian Jakubiec." Elektroniczne usługi finansowe. Charakterystyka rynku, wyzwania i inicjatywy regulacyjne (stan na koniec 2001 r.) Feb. 2002. *Narodowy Bank Polski Materiały i Studia* (139).

Websites

<http://www.kredyt mieszkaniowy.pkopb.pl>
<http://europa.eu.int/eur-lex/lex/en/index.htm>
<http://invest.bnpparibas.com/en/glossary/glossary.asp?Letter=O>
http://membres.lycos.fr/cac7/Bourse_de_A_a_Z/index-3.html
<http://skarb.bzwbk.pl/11276>
http://www.advfn.com/money-words_.html
<http://www.aib.ie>
http://www.aigfundusze.pl/index_185.htm
http://www.allianz.pl/x_main.php?id_kategorii=1801
<http://www.amplicolife.pl>
<http://www.bankofamerica.com>
<http://www.bgz.pl/u235/template/dictionary,BgzResultsListScreen.wm/navi/30466>
<http://www.bp.com.pl/u235/template/dictionary,BpczResultsListScreen.wm/navi/30770>
http://www.business.gov/phases/launching/are_you_ready/glossary.html
<http://www.bzwbk.pl>
<http://www.cdmpekao.com.pl>
<http://www.chartfilter.com/glossary.htm>
<http://www.citigroup.com>
<http://www.credit-agricole-sa.fr>
<http://www.cu.com.pl/Portal?secId=2N4H3IX0CAJVQO84VLRJ00VV&types=3&query=wszystkie&diction ary=true&count=5>
<http://www.ecb.int/home/glossary/html/glossb.en.html>
<http://www.e-msp.pl/172>
<http://www.finanz-adressen.de/europa/lex-de/A.html>
<http://www.fortisbank.be>
<http://www.getinbank.pl/28.php>
<http://www.glossarist.com/glossaries/>
<http://www.gobcafunds.com/glossary/>
<http://www.hvbggroup.com>
http://www.inteligo.pl/infosite/oferta_dla_ciebie_slowniczek.htm
http://www.inteligo.pl/infosite/oferta_dla_ciebie_slowniczek_oblig.htm
<http://www.investopedia.com/dictionary/>
<http://www.investorwords.com/>
http://www.lasallebank.com/financial_library/glossary.html
<http://www.lexicool.com/dlink.asp?ID=0GM5OT83860&L1=23&L2=09&CA=09>
<http://www.lexicool.com/online-dictionary.asp?FSP=A23B09C09>
http://www.mbank.com.pl/inwestycje/emakler/abc_gieldy/slownik.html
http://www.millenet.pl/Millennium/pomoc/?_c=721
http://www.nbp.pl/publikacje/materialy_i_studia/139.pdf
<http://www.pko-cs.pl/SAM/index.php?id=slownik>
<http://www.pzu.pl/?nodeid=slownik>
http://www.sg-tradeservices.com/index_glossaire_ENG.php
<http://www.speculativebubble.com/terms/glossary.shtml>
http://www.thectr.com/glossary/futures_index.htm
<http://www.trados.com/solutions.asp>
<http://www.ubs.com>
http://www.unicredit.it/DOC/jsp/navigation/glossary_content.jsp?parCurrentId=0b00303980007a94&parCurrentPage=responsabilita.html&parLocale=
<http://www.union-investment.pl/vademecum/slowniczek/index.html>

Conjugated Infinitives in the Hungarian National Corpus

Gergely Bottyán^{1,2} and Bálint Sass¹

¹ Department of Corpus Linguistics, Research Institute for Linguistics,
Hungarian Academy of Sciences

² English Linguistics PhD Program, Doctoral School in Linguistic Sciences, ELTE
{bottyang,joker}@nytud.hu

1 Introduction

The infinitive is one of those linguistic forms with which nonfiniteness, i.e. the verbal feature meaning lack of tense, number and person markers, is usually associated. This is a direct consequence of the fact that we only find nonfinite infinitives in Slavic languages and in most Germanic and Romance languages. However, in languages as diverse as Hungarian, Portuguese and Welsh, for example, there are both nonfinite infinitives and conjugated infinitives, i.e. infinitives that are inflected for number and person [1].

The two types of Hungarian infinitive are exemplified in Table 1.

Table 1. The two types of Hungarian infinitive

<hr/>			
I.	Reggel	fel kell kelni.	
	morning up	must wake- <i>INF</i>	
One has to wake up in the morning.			
<hr/>			
II.	Írnia	kell.	
	Read- <i>INF</i> -[3rd sing]	must	
(S)he must write.			

Hungarian infinitives of the conjugated type (II. in Table 1) have recently attracted considerable attention from generativist syntacticians. Much effort has been made to specify the sentential contexts in which conjugated infinitives occur and the structural representation of phrases formed with conjugated infinitives [2–4]. As is generally the case with syntactic research done in the Chomskyan paradigm, the authors of these studies relied on their own linguistic intuition and no systematic data collection procedure was followed. Nevertheless, the specification of contexts provided in [2] is said to be exhaustive and based on empirical material.

The present paper reports on the investigation that we have performed on the basis of the 153.7 million word lemmatized, morphosyntactically tagged and

disambiguated Hungarian National Corpus [5]. Our principal aim was to check the validity of the claim that all linguistic items (hereafter called licensors) that take conjugated infinitival complements are identified in [2] by making a list of such items in the corpus data. A further aim was to specify which licensors occur with which conjugated infinitives in the extracted sentences. As is common in language technology, the tasks were carried out partly manually, partly automatically, with an iterative method.

2 The Procedure

First, all the sentences that contained conjugated infinitives were automatically extracted from the corpus, on the basis of the morphosyntactic annotation. Our working hypothesis was that licensors were to be found in the clause that contained the conjugated infinitive. Since the corpus is not tagged for clauses, candidates for clauses were identified with our own approximation. This approximation was based on clause-final punctuation marks and clause-initial conjunctions [6]. From the resulting set of clause candidates those members were filtered out that contained a licensor identified in [7], and the conjugated infinitive was recorded along with its licensor. In the remaining set of clause candidates new licensors were looked for manually, and the procedure was applied all over again.

3 Results Obtained

After three iterations, the following measures were obtained. The lemmatized list of licensors has 197 members. Clauses containing these licensors cover 223140 (98%) of the 228367 conjugated infinitive tokens that occur in the Hungarian National Corpus. The number of lemmatized licensor – conjugated infinitive pairs identified is 17874.

An extract from the resulting data collection can be seen in Table 2.

4 Conclusions and Further Research

On the basis of the results of our investigation, the following conclusions can be drawn. (i) A number of sentential contexts in which Hungarian conjugated infinitives occur are missing from the list in [2], thus it is not exhaustive. (ii) If at least one feature of a grammatical construction (in this case, the inflectional suffix of the conjugated infinitive) is tractable in the Hungarian National Corpus or any other richly annotated corpus of its size, it is worth the effort extracting data from the language resource partly automatically, partly manually before jumping into hasty conclusions.

Further research based on our data collection should establish whether there are semantic restrictions on the range of licensors that take conjugated infinitival complements in Hungarian. Similarly, checking the grammaticality of the

Table 2. Extract from the lemmatized licensor – conjugated infinitive pairs data collection. The headword is the lemma of the licensor, which is followed by its number of occurrences in the corpus. Then come the lemmas of the licensed conjugated infinitives in decreasing order of frequency

köteles [18 db]

3x: alávet

2x: ad megakadályoz tart

1x: átvesz biztosít gondol gondoskodik igazol marad megad megtesz visszafizet

kötelesség [138 db]

5x: biztosít

4x: lesz vesz

3x: ad hoz megjelenik megtesz szól vállal

2x: elhatárolód ellát ellenőriz elvisel értesít foglal gondoskodik ismer megvéd tájékoztat tesz visel

1x: áll átír beavatkozik beküld beszámol betart bocsát csinál eljön elmegy elvégez emel emlékezik|emlékez épít felajánl felfegyverez felismer fellép felvilágosít figyelmeztet fizet folytat fordul fölnevel gazdálkodik gyarapít hajt házasodik huny hurcol indít iszik kardoskodik kér kijelöl kikényszerít kiszab kiüresít kivesz kizeng köszön küld küzd marad megakadályoz megemlékezik|megemlékez meghallgat megismer megkérdez megkeres megőriz megszavaz megtanul megtárgyal megtart megválaszt megvív meggyőződik|meggyőződ meztelenít mozgósít néz összegyűjt politizál sorol szerez takarít támad támogat tart teljesít tisztáz törekedik tud tudat túljut tűz ügyel véd védekezik végez verekedik vet virraszt

nonfinite counterparts of the example sentences belonging to the extracted licensor – conjugated infinitive pairs in the corpus could help us specify the overlaps between the distribution of nonfinite and conjugated Hungarian infinitives. Both directions of research require a tool that enables the analyst to retrieve those sentences in the corpus that belong to a given licensor – conjugated infinitive pair in the collection. The authors of this study are planning to develop such a tool and make both the data collection and the tool available to the research community.

References

1. Miller, D. G.: Where do conjugated infinitives come from? *Diachronica*, 20, 1. (2003) 45–81
2. Tóth, I.: Inflected infinitives in Hungarian. Ph. D. dissertation. Tilburg: University of Tilburg. (2000)
3. É. Kiss, K.: Agreeing infinitives with a case-marked subject. In *The syntax of Hungarian*. Cambridge: Cambridge University Press. (2002) 210–221
4. Tóth, I.: Can the Hungarian infinitive be possessed? In Kenesei, I. and Siptár, P. (eds.), *Proceedings of the conference "Approaches to Hungarian"*. Budapest: Akadémiai Kiadó. (2002) 135–160
5. Várad, T.: On Developing the Hungarian National Corpus. In Vintar, Š. (ed.), *Proceedings of the Workshop "Language Technologies – Multilingual Aspects"*. Ljubljana: University of Ljubljana. (1999)

6. Gábor, K., Héja, E. and Mészáros, Á.: Kötőszók korpusz-alapú vizsgálata [A corpus-based investigation of conjunctions]. In Alexin, Z. and Csendes, D. (eds.), MSZNY 2003 - I. Magyar Számítógépes Nyelvészeti Konferencia [Abstracts of the 1st Hungarian Conference on Computational Linguistics]. Szeged: University of Szeged. (2003) 305–306
7. É. Kiss, K.: A személyragos alaptagú főnévi igeneves kifejezés [Verb phrases with a conjugated infinitival head]. In É. Kiss, K., Kiefer, F. and Siptár, P., Új magyar nyelvtan [A new Hungarian grammar]. Budapest: Osiris. (1999) 118–121

Search Engine for Information Retrieval from Speech Records*

Michal Fapšo, Petr Schwarz, Igor Szöke, Pavel Smrž, Milan Schwarz,
Jan Černocký, Martin Karafiát, and Lukáš Burget

Faculty of Information Technology, Brno University of Technology,
Božetěchova 2, 612 66 Brno, Czech Republic
speech@fit.vutbr.cz, <http://www.fit.vutbr.cz/speech/>

Abstract. This paper describes a designed and implemented system for efficient storage, indexing and search in collections of spoken documents that takes advantage of automatic speech recognition. As the quality of current speech recognizers is not sufficient for a great deal of applications, it is necessary to index the ambiguous output of the recognition, i. e. the acyclic graphs of word hypotheses — recognition lattices. Then, it is not possible to directly apply the standard methods known from text-based systems. The paper discusses an optimized indexing system for efficient search in the complex and large data structure that has been developed by our group. The search engine works as a server. The meeting browser JFerret, developed withing the European AMI project, is used as a client to browse search results.

1 Introduction

The most straightforward way to use a large vocabulary continuous speech recognizer (LVCSR) to search in audio data is to use existing search engines on the textual (“1-best”) output from the recognizer. For such data, it is possible to use common text indexing techniques. However, these systems have satisfactory results only for high quality speech data with correct pronunciation. In the case of low quality speech data (noisy TV and radio broadcast, meetings, teleconferences) it is highly probable that the recognizer scores a word which is really in the speech worse than another word.

We can however use a richer output of the recognizer – most recognition engines are able to produce an oriented graph of hypotheses (called *lattice*). On contrary to 1-best output, the lattices tend to be complex and large. For efficient searching in such a complex and large data structure, the creation of

* This work was partially supported by European project AMI (Augmented Multi-party Interaction, FP6-506811) and Grant Agency of Czech Republic under project No. 102/05/0278. Jan Černocký was supported by post-doctoral grant of GAČR No. GA102/02/D108, Pavel Smrž by MŠMT Research Plan MSM 6383917201. The hardware used in this work was partially provided by CESNET under project No. 119/2004.

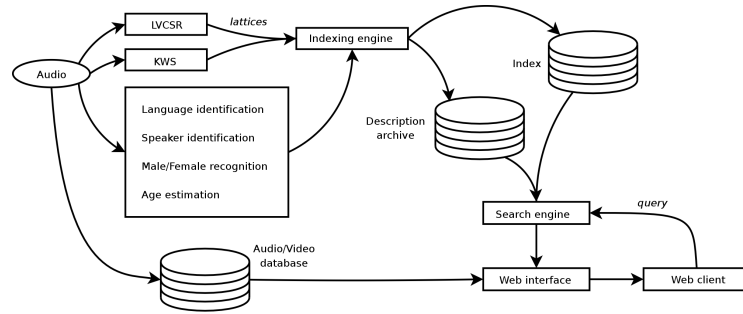


Fig. 1. The overall proposed design of the audio/speech search engine. Till now, only the LVCSR search module is implemented.

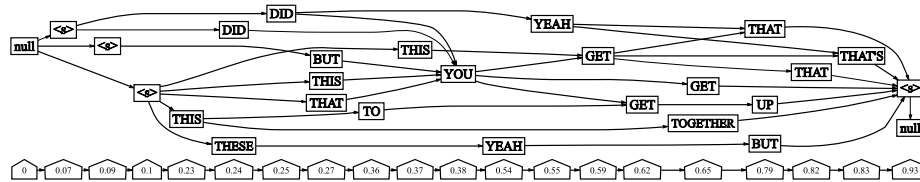


Fig. 2. Example of a word lattice

an optimized indexing system which is the core of each fast search engine is necessary. The proposed system is based on principles used in Google [1]. It consists of **indexer**, **sorter** and **searcher**.

2 Indexer

Word lattices generated by LVCSR are input to the indexing and search engine. The lattices (see example in Fig. 2) are stored in standard lattice format (SLF) [4]. The indexing mechanism (Fig. 3) consists of three main phases:

- creating the lexicon
- storing and indexing lattices
- creating the reverse index

The lexicon provides a transformation from word to a unique number (ID) and vice versa. It saves the used disk space and also the time of comparing strings (numbers need less space than words).

Lattices are stored in a structure which differs from the SLF structure. For each search result, not only the found word, but also its context has to be extracted. It means that we need to traverse the lattice from the found word in both directions (forward and backward) to gather those words lying on the best

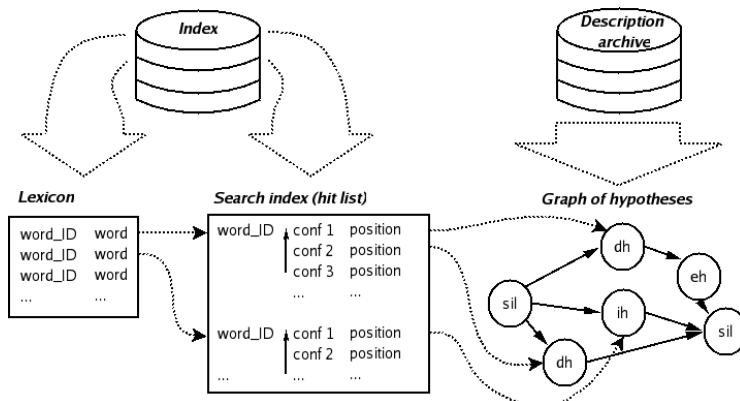


Fig. 3. Simplified index structure

path which traverses through the found word. As SLF structure keeps the nodes separated from links, lattices are converted to another structure which stores all forward and backward links for each particular node at one place. We also need to assign a *confidence* to each hypothesis. This is given by the log-likelihood ratio:

$$C^{lvcscr}(KW) = L_{alpha}^{lvcscr}(KW) + L^{lvcscr}(KW) + L_{beta}^{lvcscr}(KW) - L_{best}^{lvcscr}, \quad (1)$$

where the forward likelihood $L_{alpha}^{lvcscr}(KW)$ is the likelihood of the best path through lattice from the beginning of lattice to the keyword and backward likelihood $L_{beta}^{lvcscr}(KW)$ is computed from the end of lattice to the keyword. These two likelihoods are computed by the standard Viterbi formulae:

$$L_{alpha}^{lvcscr}(N) = L_a^{lvcscr}(N) + L_l^{lvcscr}(N) + \min_{N_P} L_{alpha}^{lvcscr}(N_P) \quad (2)$$

$$L_{beta}^{lvcscr}(N) = L_a^{lvcscr}(N) + L_l^{lvcscr}(N) + \min_{N_F} L_{beta}^{lvcscr}(N_F) \quad (3)$$

where N_F is a set of nodes directly following node N (nodes N and N_F are connected by an arc), N_P is a set of nodes directly preceding node N . $L_a^{lvcscr}(N)$ and $L_l^{lvcscr}(N)$ are acoustic and language-model likelihoods respectively. The algorithm is initialized by setting $L_{alpha}^{lvcscr}(first) = 0$ and $L_{beta}^{lvcscr}(last) = 0$. The last likelihood we need in Eq. 1: $L_{best}^{lvcscr} = L_{alpha}^{lvcscr} = L_{beta}^{lvcscr}$ is the likelihood of the most probable path through the lattice.

3 Sorting and Searching Lattices

During the phase of **indexing** lattices, the forward index is created. It stores each hypothesis (the word, it's confidence, time and position in the lattice file) in a hit list. Records in the forward index are sorted according to the document ID

(number which represents the lattice's file name) and time. The forward index itself is however not very useful for searching for a particular word, because it would be necessary to go through the hit list sequentially and select only matching words. Therefore the reverse index is created (like in Google) which has the same structure as the forward index, but is sorted by words and by confidence of hypotheses. It means that all occurrences of a particular word are stored at one place. There is also a table which transforms any word from lexicon into the start position of corresponding list in reverse index.

Searching for one word then consists only from jumping right to the beginning of it's list in reverse index, selecting first few occurrences and getting their context from corresponding lattice. The advantage of splitting the indexing mechanism into three phases is that the second phase (storing and indexing lattices), which is the most demanding, can be run in parallel on several computers. Each parallel process creates it's own forward index. These indices are then merged together and sorted to create the reverse index.

The **searcher** uses the reverse index to find occurrences of words from query and then it discovers whether they match the whole query or not. For all matching occurrences, it loads into the memory only a small part of lattice within which the found word occurs. Then the searcher traverses this part of lattice in forward and backward directions selecting only the best hypotheses; in this way it creates the most probable string which traverses through the found word.

4 Experiment

The system was tested on four AMI pilot meetings, each with four speakers and total duration of about 1.9 hours. The recognition lattices were generated using the AMI-LVCSR system incorporating state-of-the-art acoustic and language modeling techniques [2].

For testing data of 1.9 hour, the lattices consist of 3,607,089 hypotheses and 36,036,967 arcs. Searching and looking for the context of 6 hypotheses takes about 3 seconds. Although the system is not yet fully optimized, it produces search results quite fast. Approximately 95% of time is spent on looking for the context of the found word. It is possible to optimize this process with expected increase of speed by 70-80%.

5 Integration into JFerret Meeting Browser and Client/Server Architecture

JFerret [5] is a new multi-media browser for the AMI project¹ written by Mike Flynn from IDIAP Research Institute. The browser is extremely flexible, enabling almost any user interface to be composed, using a combination of plug-in modules. An XML configuration specifies which plug-in components to use, how to arrange them visually, and how they will communicate with each other.

¹ Augmented Multi-party Interaction, <http://www.amiproject.org>



Fig. 4. Search window with found hypothesis.

The JFerret plug-in for the search engine was implemented at Brno University of Technology.

On the main screen (Fig. 4), the user can see (just as in other searchers on the Web) a text field for inserting query word(s) and buttons to choose between simple and advanced search. In the advanced mode, the user can narrow the search by entering additional parameters such as name of the meeting, time interval for search etc.

The results are presented as a sorted list of hypothesis. When a user clicks on a hypothesis, a window with the particular meeting including all the available information (audio, video, slides) is opened and the particular segment is played. A list of hypothesis relevant to this particular meeting is shown as well (lower-right panel in Fig. 5) so that the user can directly browse other occurrences of the keyword. JFerret ensures the necessary synchronization of all information streams.

Although the search engine can run as a standalone application on Linux or Windows, it is more useful to run the search engine as the server. The communication is based on a simple TCP text-based protocol, so even a simple telnet client is able to send a query to the search server. One of the advantages of using JFerret as the client is that it can play audio and video data from a remote server and also synchronize several audio/video streams. All the data including audio/video files and indices can therefore be stored on the server.

6 Conclusion

We have presented a system for fast search in speech recognition lattices making extensive use of indexing. The results obtained with this system are promising

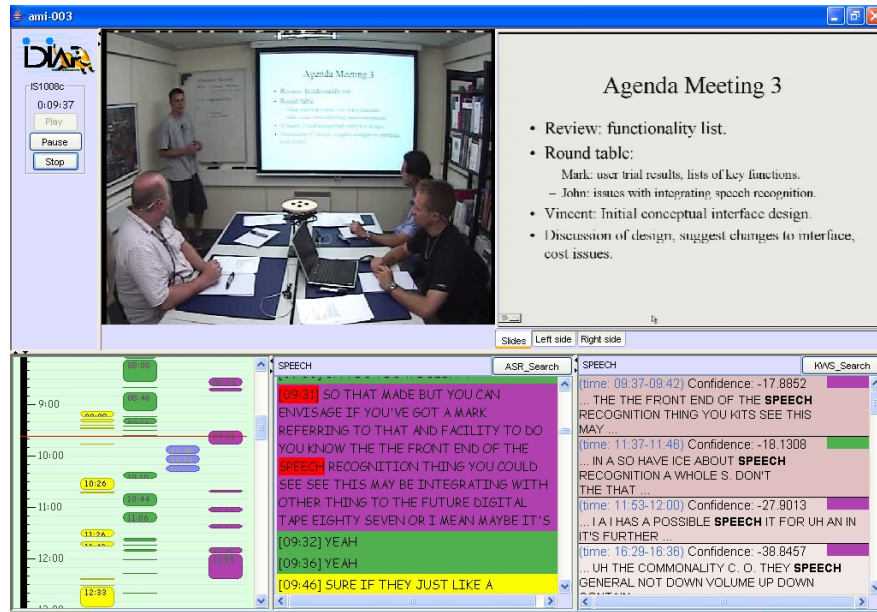


Fig. 5. JFerret replaying search result.

and the software already serves as basis for several LVCSR-keyword spotting demonstrations. The system was extended by the possibility to enter multi-word queries, and options to narrow search space (restriction for particular meetings, speakers, time intervals). It was integrated with the powerful and flexible meeting browser JFerret from IDIAP. The future work will address phoneme-lattice based keyword spotting which eliminates the main drawback of LVCSR — the dependency on recognition vocabulary [3], and methods for indexing of its results.

References

1. Sergey Brin, Lawrence Page: *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Science Department, Stanford University
2. Thomas Hain et al.: *Transcription of Conference Room Meetings: an Investigation*, In Proc. Eurospeech 2005, Lisabon, Portugal, September 2005.
3. Igor Szöke et al.: *Comparison of Keyword Spotting Approaches for Informal Continuous Speech*, In Proc. Eurospeech 2005, Lisabon, Portugal, September 2005.
4. Steve Young et al.: *The HTK Book (for HTK Version 3.3)*, Cambridge University Engineering Department, 2005, <http://htk.eng.cam.ac.uk/>.
5. Bram van der Wal et al.: *D6.3 Preliminary demonstrator of Browser Components and Wireless Presentation System*, AMI deliverable, August 2005.

A Rule-Based Analysis of Complements and Adjuncts

Kata Gábor, Enikő Héja

Research Institute for Linguistics, Hungarian Academy of Sciences
Department of Corpus Linguistics
Budapest H-1399 P.O.Box 701/518
{gkata,eheja}@nytud.hu

Abstract. In what follows we are presenting the current state of an ongoing research, which aims at differentiating between complements and adjuncts in any Hungarian text by means of a finite state transducer. The basic idea behind our method is that the case suffixes of NPs behave differently with respect to compositionality. This means that if they are attached to adjuncts, the suffixes bear always a meaning and this meaning is composed with the meaning of the verb when creating a sentence. Conversely, the suffixes on complements do not have any meaning on their own, instead the function of the whole NP is fully determined by the predicate. From this follows that verbs with complements have to be always listed in a dictionary, while adjuncts can be described by rules. In fact, the rules we applied refer both to semantic and syntactic information, and they output semantic labels attached to the NPs. These labels describe the role of the NPs in the sentence. Throughout our work we invented three types of rules: complement rules, referring only to the predicate, default-rules, referring only to the features of the NPs, and non-default rules which apply to restricted sets of predicates.

1 Introduction

The present paper describes a method for the automatic annotation of case-bearing NPs in Hungarian texts according to their grammatical function. This work is embedded in the major task of developing the Hungarian module of Intex [Silberztein, 1993], a powerful corpus processing tool which, enhanced with language-specific resources such as dictionaries and grammars, makes it possible to annotate large texts in real time. Our syntactic grammars, applied with the engine of Intex, perform a shallow syntactic analysis: they delimit the top-level constituents of clauses and identify the relations between them. For doing so, we need to be able to tell apart phrases that belong to the verbal argument structure from those which are attached to it as optional adjuncts. This paper only deals with relations between top-level NPs and the verbal predicate.

Contrary to the traditional approach which regards argument structure as a lexical property of verbs, we provide *case suffixes* with functions and derive the grammatical role of suffix-bearing NPs from the function of the suffix. The idea behind this approach is that the difference between adjunction and complementness lies in a difference of compositionality and productivity: while the occurrence of a complement is limited to the clauses which contain its governing verb, adjuncts can be productively

added to a set of (or any) predicates with the same meaning. From this standpoint, complementness becomes a possible function of case suffixes, which can only be fulfilled in a very specified context. Functions of case suffixes are captured by rules that associate a grammatical role to an NP. Accordingly, those occurrences of case suffixes which can be seized by rules without making reference to the lemma of the predicate they occur with can be considered as adjuncts.

The grammatical roles that our rules associate to NPs are mainly semantic in nature. However, the rules that attach these roles to NPs are conceived as sentence formation rules which allow the use of the NP in the given context.

This paper describes our method for isolating the different functions of case suffixes, and for creating and implementing the rules which govern their occurrences. Our work results in a criteria system for telling apart complements from free, productively used and compositional constituents. Moreover, our multilevel rule system reduces the number of verbal complements and hence the size of the lexicon by capturing more case suffix meanings by productive adjunction rules which contribute to a deeper semantic interpretation of compositional structures.

The rule system itself is implemented in Intex: it is embedded in our finite-state grammars, which, applied to a raw text, produce a simple XML output. This output can, in turn, be indexed by Xaira, a corpus query tool. This tool chain would provide means for the linguist community to annotate their own corpora and to construct and run complex queries on its elements according to their grammatical function. For instance, it becomes possible to formulate a query to retrieve NPs representing cause and to subject their internal structure to a closer examination. Moreover, we believe that rule-based NLP applications may also make use of our rule system: for example, the semantic roles that our rules output can be of significant service to IE systems.

In what follows, we will outline some of the relevant characteristics of Hungarian syntax [2], and present our viewpoint on complementness and adjunction [3]. In [4] we discuss the treatment of the case suffix chosen for testing and evaluation, while [5] describes the evaluation method and the first results.

2 Some Features of Hungarian Syntax

Hungarian is a highly inflective language with 18 cases and a (roughly) free word order: this means that almost any ordering of the verb and its complements and adjuncts is acceptable, although they yield slightly different interpretations. As described in [É.Kiss 2002], in the neutral sentence verbal complements and adjuncts follow the predicate within the VP. However, practically any of the verb's complements or adjuncts may be *topicalized* or *focused*, hence moved outside the VP. Furthermore, *verb modifiers*, i.e. verbal prefixes, adverbs or bare NP complements also precede the verb they modify. When parsing Hungarian texts, we have to face the difficulty that in Hungarian, configurationality is used to express discourse functions instead of syntactic functions, that is why it is impossible to determine dependency relations and grammatical functions on the basis of constituent order. Thus, we have to abandon the criteria of complementness relying on configurational information as in GB (on the

basis of [Radford, 1988]), for example. On the other hand, Hungarian morphology is very rich, hence we should make use of constituents' morphological features, in particular of cases, when defining complements and adjuncts.

Although we cannot gain information regarding complementness from constituent order, we can use some of the criteria proposed by GB to tell apart the NPs of different roles when inventing our rules. Even though – as we will see from the next section – we reinterpreted the notion of complementness and adjunctness, these standards suffice for our purposes. One such criterion is coordination: only two constituents of the same type can undergo coordination. For instance:

- [1] **“János beszennyezte a szőnyeget sár - ral és cipő - vel.”*
 John stained the carpet - ACC mud - INS and shoe - INS
 **“John stained the carpet with mud and with his shoes.”*

The sentence above shows that '*mud - INS*' and '*shoe - INS*' play different roles in the Hungarian sentence, in spite of the fact that they both could be interpreted semantically as the instruments of the staining event.

In accordance with the phenomenon described above, we supposed that if in a Hungarian sentence there are two distinct NPs with the same case suffix, each of them is attached to the predicate by different rules (where these rules assign different semantic roles to the NPs in tandem with their diverse syntactical function). Consider the sentence below:

- [2] *“Párizs - ban még bízta - m az apá - m - ban”*
 Paris - INE still trust - PAST - S1 the father - POSS1 - INE
“In Paris I have still trusted my father.”

In the example above the Hungarian counterparts of '*Paris - INE*' and '*father - POSS1 - INE*' own the same suffix in a well-formed sentence. According to us this is because the NPs fulfill different roles in relation to the predicate. Consider the following sentence:

- [3] *“2005 - ben Párizs - ban még bízta - m az apá - m - ban”*
 2005 - INE Paris - INE still trust - PAST - S1I the father - POSS1 - INE
“In 2005 in Paris I have still trusted my father.”

The sentence above arise a question that we will explain more elaborately in the next section, namely, if our presupposition is true, how is it possible to construct a well-formed sentence with two adjuncts that are not coordinated?

In what follows we are about to show that we need not divide the NPs into two strict categories on the ground of their relation to the verb (i.e. complementness or adjunctness). Instead we introduce a more gradual notion of complementness and adjunctness.

3 Baselines

The goal of our work is to make our syntax analyzer able to categorize all top-level NPs in a text according to their role. For doing so, instead of using configurational information we intend to use morphology, especially case feature as a marker of the syntactic role. In terms of dependency grammar the NPs' role is identical to the relation between them and the predicate. Our point of view do not contradict this approach – however, what we try to do is to reduce the references to the predicate and capture as many occurrences of NP+case structures by rules as we can. This means that we tried not to see complements in their relation to the predicates, and we conceive the predicate-argument function as one possible function that case suffixes may bear.

The case suffix can be defined as the rightmost inflectional suffix on Hungarian nouns: it cannot be followed by any other suffix, and a noun can have only one case suffix. According to these criteria, we can distinguish 19 cases in Hungarian. Our task is to examine case suffixes and enumerate their possible syntactic and semantic functions.

We found that there were two grammatical cases (nominative and accusative) that could not have a default meaning and can only occur as verbal complements. These cases have to be included in verbal valence structures. As for the other cases, we supposed that they had their own syntactic and semantic properties that can be described by general rules. Such general rules work as default rules, and they specify one or more functions for the given case suffix. They may refer to semantic or syntactic features of the constituent they appear in, but default rules may not refer to the predicate. This implies that there may be more than one default rule for a suffix: e.g. its default meaning may differ in correlation with the semantics of the NP's head noun which bears the suffix. For example, the case suffix *-ban* (inessive case) indicates the exact date if it appears on a constituent expressing time: it forms a regular adjunct of time. Otherwise, it expresses location, and also forms a regular adjunct. This rule will work as a default rule for the case suffix *-ban*, assuming that in absence of lexical rules, the case is associated to these functions independently of the context.

The structures of the form [verb + case-bearing NP] which are not covered by rules will be considered as *verb + complement* structures. They cannot be recognized by general rules as the structure is not compositional: the role of the NP in relation to the verb cannot be paraphrased without including the meaning of the verb. For example, consider the sentence:

- [4] “A közönség elhalmozta az előadó -t kérdések -kel.”
 The audience overwhelmed the lecturer[ACC] questions[INS].
 “The audience overwhelmed the lecturer with questions.”

If the structure verb + NP[INS] was compositional, we could associate an abstract label to the NP which would capture its role without making reference to the predicate of the sentence (e.g. location, manner, goal etc.). This seems impossible, as is shown by the fact that we cannot formulate natural language paraphrases for expressing the

relation between the predicate and the NP without including the verb itself or a synonym of it.

On the basis of these considerations, sentence [4] can be classified as a clear example of complementness – it has to be included in the corresponding entry of the verbal valence dictionary as a lexically coded structure.

However, some structures in which case suffixes are used seem to be midway between rule-based constructions and total lexicalism. This means that the function the case suffix has in the structure depends on the *semantic class* of the predicate. For example, the ablative case *-tól* may have two functions: with movement verbs it marks the starting point of a movement, with verbs that express a change in someone's state, it expresses the cause of the change. Thus, we can associate these functions to the case suffix by means of two rules, each of them referring to the semantic class of the verbs they occur with. We claim that adjunction is a totally productive process which works independently of the predicate of the clause (the only constraint being the *appearance* of a predicate which allows for the adjunction), while complementness is a lexical property of individual verbal lemmata. As non-default rules are only productive with a restricted set of predicates, it is not a trivial task to tell whether they relate to complementness or adjunction. On the other hand, having a wider range of constituent roles than the above-mentioned two categories has the advantage of enabling us to account for the co-occurrence constraints that apply to NPs with the same case suffix. If we affirm that a well-formed Hungarian clause cannot contain two or more NPs with the same grammatical role expressed with the same case suffix unless they are coordinated, then we have difficulties explaining the correctness of sentence [3], which has two adjuncts of the same structure (NP+inessive case). This observation led us to state as many roles for case suffix bearing NPs as there are rules for the given case suffix: each of our rules outputs a different role label. Hence, we can simply state that no clause can contain more NPs with the same role and the same suffix: in other words, each rule may apply only once in a clause¹. In sentence [3], one of the NPs in inessive case is matched by a rule which marks it as a time-adverbial constituent, while the other one is labelled as a constituent expressing location – this is why they are allowed to appear together in a clause without being coordinated.

4 Establishing Verb Classes – An Example

After introducing the general principles which serve as the background of our research, we are describing a concrete example. This example is the Hungarian instrumental case '*-val*' ('-with'). In the case of the given nominal suffix we supposed that there were two default rules – which do not refer to the predicates at all, while defining the NP's role in the clause.

One of them is an associate rule. Here the NP with the given suffix represents a participant of the event denoted by the predicate, where the participant has the same

¹ NPs which are attached to the predicate with the same rule are necessarily coordinated. Nevertheless, our syntax analyzer treats coordinated NPs as a single NP with two heads, thus they do not require two applications of the same rule to be recognized and labelled.

role in relation to the main event as the participant referred by either the subject of the predicate (if it has two arguments) or by the object (if there are three arguments). We distinguished between the two rules on the basis of semantic features: if the NP owns a “+HUMAN” feature it is supplied with the additional feature *role=ASSOCIATE* by the corresponding rule:

- [5] *'...a magyar politikusok itt véletlenül összeakadtak Berijával, [ASSOCIATE a nagy hatalmú szovjet miniszterelnök-helyetessel]...'*
 '...here the Hungarian politicians incidentally met Berija, [ASSOCIATE the mighty Soviet deputy prime minister]...'

The other rule is an instrumental-rule. This means that the appropriate NP refers to the instrument with which the relevant act is carried out. The only condition under which the NP is assigned the default instrumental role is that it has to lack any semantic feature.

- [6] *'...azok előtt, akik kezüket [INS saját kommunista elvtársaik vérével] mocskolták be.'*
 '...to those, who soiled their hands [INS with the blood of their own fellow Communists].'

It is important to emphasize that in the examples above, we really did not refer to the predicates. We made only use of the semantic features of the NPs under examination. This correlates with the fact that the traditional way of considering adjuncts really expects the meaning of the adjunct to be independent of that of the predicate.

The default-rules have also another advantage. Using such rules enables us to classify all NPs in the text with instrumental case. Considering the examples above, neither predicate is listed in our vocabulary, still the default rules enable us to classify the NPs in the sentences.

On the other hand – throughout the testing process – it turned out that we need more default rules. We had to face that in Hungarian it is quite common that the NPs with the instrumental suffix serve as adverbs. This function of the corresponding NP is absolutely productive and in fact there are some cases when it seems to be hard to tell apart adverbial NPs from instrumental ones. This is because adverbial NPs express the mode of carrying out an action, while instrumental NPs refer to the instrument of that action. Until we do not have syntactic criterion the two semantic contents can be easily mixed up. This statement is illustrated by the example below:

- [7] *“János autóval ment mozi - ba.”*
 John car - INS went cinema - ILL
 “John went to the cinema by car.”

Where the Hungarian counterpart of '*car - INS*' expresses the mode of the action and the instrument of the action at the same time. The difficulty of telling apart mode and instrument in such cases can be seen from the corresponding questions:

[8.a.] “mi - vel megy János mozi - ba?”
 what - INS go - S3 John cinema - ILL
 “With what does John go to the cinema?”

[8.b.] “*hogy megy János mozi - ba?*”
 how go - S3 John cinema - ILL
 “how does John go to the cinema?”

As the *wh*-expressions show, [8.a.] supports an instrumental reading of [7], while [8.b.] an adverbial one (i. e. one which refers to the mode of the action).

As a first step to resolve this problem we relied on the presupposition that especially nouns derived from adjectives or verbs are inclined to go in adverbial position. Hence the input of our third default rule are words ending with '-sÁg' and '-Ás' which are the most common deadjectival and deverbal derivational suffixes, respectively.

Throughout the classification of NPs we used one more default rule which assigned the attribute 'role = MEASURE' to the input text. This rule also relies on semantic features – to be exact, in this case on two semantic features – on MEASURE and TIME. An NP with such a feature and instrumental case is supposed to express the measure of a change, or a measure of time between two events. For example: '*[MEASURE Húsz évvel] ezelőtt*' ('[MEASURE Twenty years] before') or '*[MEASURE Három százalékkal] nőtt*' ('Increased [MEASURE with three percentage]').

Stating default rules we do not refer to predictable syntactic or morphological alternations, since these alternations always allude to the predicate which would contradict to the definition of default rules, according to which inputs of default rules are exclusively properties the NPs in question.

As for non-default rules we have established the following classes. The first rule we deal with is *non-default instrument*. What is the discrepancy between default and non-default instrument rules? We have called an NP non-default instrument if the rule refers not only to the NP but also to the predicate. Such a predicate is for example the above mentioned '*beszennyez*' ('stain'), where the predicate itself requires an instrumental argument. Remember the difference between '*besszennyez sárral*' ('stains with mud') '*beszennyez a cipővel*' ('stains with shoes') where 'sárral' and 'cipőjével' are different instrumental arguments. One syntactic test to verify our hypothesis was the coordination test, which in this case resulted in an ill-formed sentence (see [1]). Why do we need to distinguish between default and non-default instrumental rules? Firstly, as the syntactic tests show, there is a clear difference between them.

Secondly, when the non-default instrument rule matches a string, it excludes the applicability of default rules, especially those that would assign a different role to the NP in question.

Considering the non-default associate rule, the same questions arise as in the case of instrument rules. The answer for the first one is that the difference between default

and non-default associate rules is that while the former refers to a semantic feature of the corresponding NP (i.e. it is +HUMAN) the latter do not rely on this information. Actually, this discrepancy correlates with the fact that there is a verb class (e. g. '*veszekedik*' - "quarrel"), where the NP with instrumental suffix in their argument structure behaves always as an associate. For example:

- [9] "*János veszekszik az autó - val.*"
 John quarrel - S3 the car - INS
 "John quarrels with the car."

This sentence – as opposed to [10] – has only the meaning that the car was a participant in the act of quarrel not the instrument of it, although the Hungarian counterpart of '*car*' lacks the feature +HUMAN.

- [10] "*János Mari - val ment mozi - ba.*"
 John Mary - INS went cinema - ILL
 "John went to the cinema with Mary."

'*Mary*' in the sentence above denotes an associate only if the presupposition holds that it refers to a human being. Otherwise it would refer to an instrument.

Above we showed that the distinction between default and non-default rules is theoretically and empirically also well-motivated. In what follows we are about to describe the classification of NPs considering also the semantic and syntactic characteristics of the predicate they belong to.

The next class we have dealt with is what we called '*change in mental state*' class. This group consists of verbs such as '*megdöbbszít*' ('astonish'), '*felidegesít*' ('make sy nervous'), '*megrémít*' ('horrify'). The example below we use to represent the meaning of this group, as it is in [12]:

- [11.a.] "*János megdöbbszította Mari a hírt - rel.*"
 John astonished Mary - ACC the news - INS
 "John astonished Mary by telling the news."

- [12] CAUSE(János, E), where E<news, CHANGE(S, S')> and CAUSE(news,S')

which means that John brought(CAUSE) a situation (E) into existence, and E is a two-argument predicate, such that there is an x (news), which causes(CAUSE) a change in Mary's *mental* state, namely a change from S into S'. The next question is how could we verify syntactically these three semantic components (i.e. CAUSE, MENTAL, CHANGE)?

The test we applied looked like this:

- [11.b.] "A hírt megdöbbszította Mari - t."
 The news astonished Mary - ACC
 "The news astonished Mary."

- [11.c.] "Mari megdöbbszített a hírt - től."

Mary astonished the news – ABL
 “Mary was astonished by the news.”

We supposed that a verb belongs to this class if and only if it can undergo systematically the syntactic alternations represented in [11.a.], [11.b.] and [11.c.].

As [11.a.] and [11.b.] show, the predicates belonging to this verb class have to have at least one interpretation where the subject is non-agentive. Otherwise [11.b.] should be ungrammatical, since the denotata of such subjects cannot carry out an action voluntarily. This requirement is responsible for the fact that most verbs in this class – not all, though – are mental verbs. (Note that all mental verbs with this argument structure have a non-agentive interpretation.)

[11.c.] illustrates the necessity of the metapredicates CAUSE and CHANGE. According to us one default meaning of the ablative case is the CAUSE of CHANGE, where the change has to be a transition from a state *into* another state. There are two arguments to support this thesis. The first relies on the English translation; the elements of this verb class are inclined to be translated into English by perfective verb forms. The structure in sentence [11c] cannot even be put in an imperfective form with the same argument structure. This fact is in accordance with our expectation that sentences with the perfective forms of these structures involve the complete transition between two states, while imperfective forms express the process of transition, but do not imply the end of this process. The other argument takes the observation as its starting-point that there is a verb class with verbs such that the argument with instrumental suffix represents the CAUSE as in the instances above, but there is no transition between definite states which means that CHANGE predicate cannot apply:

- [13a] “Az igazgató János - t terhelte a feladat - tal.”
 The director John - ACC burden the task - INS
 “The director burdened with the task”
- [13b] “A feladat János - t terhelte.”
 The task John - ACC burdened
 “The task burdened John.”
- [13c] “János terhelve van.”
 John burdened is
 “John is burdened.”
- [13d] **“János terhelve van a feladattól.”*
 John burdened is the task – ABL
 “John is burdened by the task.”

This semantical intuition is caught by the explicit criterion of the syntactical ill-formedness of the sentence [13d]. As the counterexample demonstrates the metapredicate CHANGE is distinctive, that is why we need it independently of CAUSE.

An other predicate class consists of *factive* verbs. In Hungarian factive verbs are morphologically marked. They are derived by means of the '-At', '-tAt' suffixes.

- [14] “*János levág - at - ja a haját a fodrász - szal.*”
 John cut - FAC - S3 the hair the hairdresser - INS
 “John makes the hairdresser cut his hair.”

In semantical terms such a verb could be described as:

- [15] CAUSE(John, E), where E <hairdresser, hair, ...> and SUBJ2(hairdresser, E)

This means that John brings an event (E) into existence, and this event has at least two arguments, and the hairdresser is the subject of the verb which describes E.

Consequently, the instrumental argument of the predicate is the subject of the factive predicate's base verb.

At the present stage of our work we have one more class, which we called *theme*. We did not rely on additional syntactical information while defining this group, instead we listed here those verbs that are typical in this respect.

The remainder verbs which allow for the appearance of a NP with instrumental case but cannot be correctly described by any of the elaborated rules are considered as verbs lexically governing a NP with instrumental case. This property has to be marked in the corresponding entry of the verbal dictionary.

5 Implementation

The work flow started with the choice of the verbal vocabulary. It consists of the 2.800 most frequent verbs from the Hungarian National Corpus [Váradi, 2002]. This vocabulary is being used for defining case suffix functions, forming the predicate groups and coding valence. The way we go through the verbs is according to case suffixes: we have chosen four frequent cases: instrumental (*-vAl*), dative (*-nAk*), ablative (*-tÓl*) and sublative (*-rA*) on which detailed rule systems were elaborated. First we determined default rules, because the creation of non-default rules presupposes the existence of the default meaning. After that we examined and classified all the verbs of the list according to the role that an NP with the given suffix may play in relation to them. This process yielded the predicate groups which proved to be semantically motivated and, as we expected, we found that many of them could be used for more than one case suffix. As a last step, structures which are not matched by any of the rules were coded as verbal complements in verbal dictionary entries.

The rule system was implemented as a part of the complex linguistic analysis module for Hungarian within Intex [Váradi, T. and Gábor, K., 2004]. Intex is a language processing tool which enables its users to provide large texts with linguistic annotations at several levels of linguistic descriptions, such as morphology, syntax and semantics. Lexical and morphological information are coded in dictionaries, while syntactic patterns are represented by graphs. Intex compiles both dictionaries and graphs into finite state transducers, hence its speed and efficiency in pattern matching. Although its core is a finite state engine, it is enhanced by functions which give it the descriptive power of a Turing machine.

The input of the automatic processing is raw Hungarian text, while the output contains syntax-related annotation of the sentences. Lexical and morphological

analysis relies on dictionaries. A dictionary entry is composed of a word form, its corresponding lemma and morphological code, possibly enhanced by additional semantic or syntactic information. The dictionaries we use for text processing contain the 900.000 most frequent word forms of the Hungarian National Corpus, analyzed morphologically by the Humor analyzer [Prószéky, G. and Tihanyi, L., 1996].

Linguistic analysis includes tokenization, sentence splitting, lexical and morphological analysis, recognition of multi-word expressions and named entities (as a part of the tokenization), and finally shallow syntactic analysis. Syntactic analysis in turn is composed of several steps of grammar applications, and the cascaded grammars refer to the output of previously applied ones. As the finite verb representing the predicate is considered to be the central element of the clause and we assume that it enters in dependency relations with (heads of) fully built phrases, the first task of the analysis is to recognize these phrases. After phrase recognition, finite verbs are marked as predicates. As we need to identify predicates' argument structures, the domain which contains verbal arguments and free adjuncts has to be identified. This step is performed by our clause boundary detection grammars [Gábor, Héja, Mészáros 2003]. Hereupon, each step of the analysis is accomplished within the domain of the clause.

The implementation of the rule system requires as a preprocessing the annotation of NPs and verbal predicates with their relevant properties used by the rules. First of all, Intex dictionary entries are enriched with nouns' semantic and morphosyntactic features (e.g. time, human, measure). On the basis of dictionary entries, recognized NPs receive additional semantic attributes. In addition, verbal predicate classes were marked up in the dictionary.

Among the elaborated case suffix rule systems, we have chosen the instrumental case for implementation and testing. The rules apply to all NPs in the text in the order corresponding to their degree of specificity. The three specificity degrees we distinguished were 1) complements (the highest specificity is due to the reference to verbal lemmata in the rules), 2) non-default rules (which make reference to predicate classes), 3) default rules. Application order of rules within the same group is not significant as they cannot apply to the same NP.

6 Evaluation

The implemented rules were tested on a text which consists of two novels: “Abigél” (1978) by Magda Szabó and “Leírás” (“*Description*”, 1979) by Péter Nádas. These texts are parts of the Hungarian National Corpus. The testing process was followed by the evaluation. The input text is “Nagy Imre élete és halála” by Tibor Méray (“*Thirteen Days that Shook the Kremlin: Imre Nagy and the Hungarian Revolution*”, 1958), a 12545-sentence novel which contains 191373 tokens: 130027 (26475 different) word forms, among which 24914 were morphologically recognized and 1561 were unknown.

In the text we found 29855 NPs. As the main point of the evaluation is to examine the efficiency of our approach and to decide on the possibilities of the future improvement, we concentrated on precision values. Since our approach is a rule-based

one, stating the reasons of improper hits enables us to improve systematically our system.

The evaluation was made using a corpus query tool called Xaira (XML Aware Indexing and Retrieval Architecture). Xaira was developed by Lou Burnard and Tony Dodd and it is distributed by the Research Technologies Service at Oxford University Computing Services.

As its name implies, Xaira was designed to enable queries on corpora consisting of well-formed XML. This tool renders possible to compose complex queries, furthermore to attach stylesheets to the results of those queries. Using these two capacities of the software, we created a human-readable output which facilitated the evaluation process.

Results

	<i>PRECISION</i>	<i>NUMBER OF HITS</i>
complement	55,00%	127
rule factive	100,00%	1
rule cause	36,00%	11
rule theme	92,30%	52
rule associate	65,00%	76
default associate	61,60%	129
instrument (rule and default)	42,37%	573
default mode (-Ás), (-sÁg)	54,76%	168
default measure+time	88,09%	42
lexicalised (örömmel, valósággal...)	100,00%	59
	59,57%	1238

Notes on results

Throughout the evaluation process we distinguished between default and non-default associate rules, while the outputs of the default and non-default instrument rules we did not count as separate values. This is because what we found is that the main source of

the improper results is the high frequency of noun phrases with instrumental case which serve as adverbs. Thus, the very difficulty is to distinguish between the adverbial NPs and the instrumental ones. However, this problem do not arise, or at least is less common among NPs categorized as associates. This is because the default associate rule is applicable only if the relevant NP owns the feature +HUMAN, and the adverbial NPs obviously lack this semantic feature. In the case of default and non-default instrumental NPs we do not rely on any special semantic feature. That is why it is hard to tell apart instrumental NPs from adverbial ones. In this respect they do not differ from non-default rules, since the non-default rules have no obligatory arguments either and the predicates they refer to can be also modified by adverbial NPs.

From the results we had to conclude that our future work has to concentrate on the improvement of the vocabulary on one hand. On the other hand, since our rule-system is a relatively high-level application, we have to exclude the mistakes originating from the lower-level rules.

References

1. É. Kiss, K.: The syntax of Hungarian. Cambridge University Press, 2002
2. Gábor K., Héja E., Mészáros Á.: Corpus-based Examination of Hungarian Conjunctions. In: Alexin Z., Csendes D. (eds.): *Proceedings of the First Hungarian Conference on Computational Linguistics*, Szeged University Press, 2003. Szeged, pp. 305-306.
3. Gábor, K.: *Syntactic Analysis and Named Entity Recognition for Hungarian with Intex*. to appear in the Proceedings of the 6th and 7th Intex workshops
4. Prózék G., Tihanyi L.: Humor – a Morphological System for Corpus Analysis. *Proceedings of the first TELRI Seminar in Tihany*. 1996. Budapest, pp. 149-158.
5. Radford, A.: *Transformational Grammar*. Cambridge University Press, 1988. Cambridge
6. Silberztein, M.: *Dictionnaires électroniques et analyse automatique de textes: Le système Intex*. Masson, 1993. Paris
7. Váradi, T.: The Hungarian National Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002. Las Palmas pp. 385-389
8. Váradi, T., Gábor, K.: A magyar INTEX fejlesztésről. (Developing the Hungarian Module of Intex). In: (Alexin Z., Csendes D. eds): *Proceedings of the Second Hungarian Conference on Computational Linguistics*, Szeged University Press, 2004. Szeged, pp. 3-11.
9. Xaira: www.xaira.org

Levenshtein Edit Operations as a Base for a Morphology Analyzer

Radovan Garabík

Ludovít Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia
korpus@juls.savba.sk
<http://korpus.juls.savba.sk/>

Abstract. Levenshtein distance between two strings is defined as the minimum number of operations needed to transform one string into the other, where an operation is a character insertion, deletion, or substitution. Sequence of edit operations needed to transform lemma into an inflected word form can be applied to a broader class of words belonging to the same paradigm template and can be used as a base for a word form generator, providing an alternative for commonly used approach based on word stem and suffixes conforming to an appropriate inflectional paradigm.

1 Levenshtein Distance and Some Definitions

Levenshtein distance[1] is a metric defined on the space of strings as a minimum number of Levenshtein edit operations needed to transform one string into the other, where by a Levenshtein edit operation we understand insertion, deletion or a substitution of a character. Levenshtein distance is commonly used in fuzzy string comparisons and in evaluating word similarities.

Let \mathbb{S} be a set of all strings. A Levenshtein edit operation e can be formally described as $e = (o, s, d)$ – a triple of operation type o , position in the source string s and position in the destination string d , where operation type o is one of *replace*, *insert* or *delete*. For *replace* or *insert*, the replacement/new character is taken from the destination string. For *delete*, only the source position is relevant.

Sequence of edit operations $q = (e_1, e_2, e_3, \dots)$, together with the destination string D , when applied to a string $S \in \mathbb{S}$ defines a mapping function $f : \mathbb{S} \mapsto \mathbb{S}$. Empty sequence corresponds to identity function.

Let \mathbb{W} be a set of all the word (i. e. all the word forms) in a given natural language – be it a controlled (codified) subset of a language, language attested in a corpus or an ambitious project of describing the “complete language”.¹

The elements of \mathbb{W} can be conveniently grouped into *lexemes* – subsets according to (intuitively defined) grammar categories and semantic identities.

¹ Even if in in this situation the relation of “belongs to” would not be clearly defined and therefore we could not talk about a proper set.

Words belonging to one lexeme have the same semantic meaning and differ only in grammar categories. From each lexeme $\mathcal{L} \subset \mathbb{W}$ we choose one word form $l \in \mathcal{L}$ and call this word form a *lemma*².

To each word form $w \in \mathbb{W}$ we can assign a set of *grammar categories* $G_w = \{g_1, g_2, g_3, \dots\}$ (two or more such sets can be assigned to the same word form, in case of *homonymy*). The exact categorisation into lexemes and grammar categories is subject to different grammar theories and linguistic opinions and is by no means fixed for a given language.

Let us define a bijective mapping $G_w \mapsto T_w$ from these sets of grammar categories into short strings called *tags*. The set \mathbb{T} of all defined tags is called the *tagset*.

For a lexeme $\mathcal{L} = \{l, w_1, w_2, w_3, \dots\}$, each element of which has been assigned one or more tags we define a tagged lexeme as a set of tagged word forms (tuples consisting of a word form and a corresponding tag, $w_i^T = (w_i, t_i)$):

$$\mathcal{L}^T = \{(l, t_l), (w_1, t_1), (w_2, t_2), \dots\}$$

Now for each tagged word form w_i^T there exists a mapping function f_i consisting of Levenshtein edit operations such that $f_i(l) = w_i$. Set of mapping functions $\{f_0, f_1, f_2, \dots\}$ belonging to one lexeme defines a *paradigm template*. Conveniently, mapping function f_0 maps lemma to itself. Let us take another lemma l' . By applying each of mapping functions f_i to the lemma l' we get a set of strings $w'_i = f_i(l')$. If these strings w'_i are meaningful words from our language $w'_i \in \mathbb{W}$ and the set $\{w'_i\}$ forms a lexeme \mathcal{L}' (or a subset of a lexeme), we say that lemma l' is *inflected* by the paradigm template of lemma l . Sometimes, in order to get the full lexeme \mathcal{L}' , we have to inflex lemma l' by several paradigm templates l^j :

$$\mathcal{L}' = \bigcup_j \mathcal{L}'^j$$

In natural languages, we can expect that the paradigm templates as described above correspond to *paradigms* as used in common linguistic theories, and that if the set of grammar categories is selected according to commonly used grammar, inflections (as defined by our Levenshtein edit operations) of almost all the lemmas in the language can be described by a small number of basic paradigm templates. This can be an alternative to commonly used approach based on word stems or root morphemes, suffixes and rule based inflections[2, 3]. It is obvious that we need not to limit ourselves only to the inflectional morphology – the description above can be applied to any word form changes that can be described in terms of basic forms, their changes and formal tags, for example it can be used for derivational morphology, if we can find (derivational) inflection paradigms and formalise the morphology categories.

² Strictly speaking, a lemma might not be a part of language vocabulary, but be just a potential form – see Slovak word *pošiel* with a formal lemma *pôjst* for a nice example.

2 Technical Implementation

Our system is really just a morphology generator – for each lemma known to it, it is able to generate all the forms, together with their respective tags. By putting all the forms and tags with information about lemma into a database, the system is able to work as a morphology analyzer – we just look up the analysed form in the database and find out corresponding morphology tag and lemma.

The system consists of two logically disjunct parts. One part is responsible for creating tables of paradigm templates and lists of mapping of all the lemmas into appropriate paradigm templates. This contains also helper programs used by linguists to create, evaluate and modify these tables and lists.

The second part is meant for end users queries and is nothing more than a simple wrapper around the database query library, to facilitate the lookup.

The software is published under GNU General Public License[4] version 2 and can be obtained from the Slovak National Corpus WWW page³.

3 General Principles

All the texts, input and output in our system is done in UTF-8 encoding[5]. While the whole system could in principle work in an 8-bit encoding, in order to evade eventual problems with encodings we decided to keep all the data exclusively in UTF-8. This means that all the files parameters and the output of all the commands mentioned hence are in UTF-8.

Since it is a suffix morphology we are interested in, we need to count the position for Levenshtein edit operations from the end of the words, so that words of different lengths but sharing the same suffix inflections can be declined by the same paradigm template, in order not to inflate unnecessarily the number of paradigm templates. This is easily realised by reversing the input strings before applying the edit operations, and by reversing the output obtained as the result – all done transparently to the users.

Another little twist used to keep the number of paradigm templates down is based on the observation that, at least in some Slavic languages, the orthography often marks certain phonetic features implicitly. For example, palatalisation in Western Slavic languages is marked either by special diacritics, or not marked if a certain vowel follows. Since inflection suffixes often start with a vowel, the overall visual effect is that of stripping diacritics from the last consonant of the root morpheme during inflection. This means that we need at least as many additional paradigm templates as there are different possible palatalised consonants at the end of lemmas, because for each of the consonants, the change “*consonant with diacritics*” → “*consonant without diacritics*” is a separate Levenshtein edit operation. Consequently, if we encode the diacritic sign as a separate character, all the edit operations would converge to one, deletion of the diacritics. Fortunately, this is exactly what the Unicode normalization NFD does[6]. Therefore,

³ <http://korpus.juls.savba.sk>

we designed our system to work in NFD normalization internally, normalizing user input into NFD before processing, and normalizing the output to NFC – again, completely transparently to the user.

4 Format of a Paradigm Template

Paradigm templates are described in separate files, one file for one paradigm template. The files have `.par` extension and are scanned recursively down to an arbitrary deep directory structure – this makes it possible to conveniently group paradigm templates in subdirectories, for example according to commonly used part-of-speech categories or first letters of a template name, or any combination thereof.

Each paradigm template file is a simple text file in UTF-8 encoding. Any line beginning with `# U+0023 NUMBER SIGN` is a comment and is ignored, empty lines are ignored too. First non-ignored line contains either a single word – lemma of the paradigm, that serves as a paradigm template name, or it contains two words separated by a whitespace – first one is lemma, second one template name (more templates can exist for the same lemma, in case of highly homonymous lexemes). Template name is unique for a given template, two templates cannot have the same template name. All the following lines have to begin with tag, followed by colon, followed by a specific inflected word form for the given tag – or two or more word forms separated by a whitespace, in case of several possibilities.

```

ucho ucho_2
# ucho: orgán sluchu, arch. tvar G pl.

SSns1: ucho
SSns2: ucha
SSns3: uchu
SSns4: ucho
SSns5: ucho
SSns6: uchu
SSns7: uchom

SSnp1: uši
SSnp2: ušú uši
SSnp3: ušiam
SSnp4: uši
SSnp5: uši
SSnp6: ušiach
SSnp7: ušami

```

Table 1. Example of a paradigm template, with lemma `ucho` and paradigm template name `ucho_2`. Note the stem change in plural and double form in genitive plural.

5 Working with Paradigm Templates

Contrary to common trends, we have not designed our system to be a monolithic application with a graphical user interface, but rather as a set of command line utilities with a clearly defined functionality.

Paradigm template can be created either fully manually, with an ordinary text editor, by entering all the tags and corresponding word forms, or by inflecting the new paradigm template lemma by another, already existing template and manually fixing the discrepancies.

Following commands are used to work with paradigm template tables and lists:

mlv_decl lemma [template]

Inflex *lemma* and print all the inflected forms, using either given paradigm template, or a default one if not given.

mlv_addpar new_template old_template

Create a new paradigm template, using *old_template* to inflex lemma given as *new_template*. *new_template* can be optionally given as a full path inside the data directory, such as nouns/masculine/lemma.

mlv_learn

Read all the tables and prepare internal pickled dictionaries for further use. It is necessary to run this command in order for any changes in the paradigm templates or lists to take effect.

mlv_maketables

Prepare constant database tables.

6 Format of Paradigm Lists

A paradigm list maps all the lemmas from the language into paradigm templates. File containing paradigm list has `.list` extension, and similarly to the paradigm templates, multiple files with paradigm lists are possible, in an arbitrary directory structure. Again, empty lines and lines beginning with `# U+0023 NUMBER SIGN` are ignored. Any non-ignored line contains lemma, followed by a colon, followed by a name of paradigm template the lemma should be inflected by. If a lemma can be inflected by two or more paradigm templates, it should be specified more times.

7 Software Needed

As our preferred programming language is Python[7], the whole system was implemented in Python and is a bit Python-centric. A reasonably recent python version is needed, the system was developed with version 2.3. We used GNU/Linux as our development platform, but the system should work on any reasonably modern Unix OS.

To create and test paradigm templates, following software libraries and python modules are needed:

- python-levenshtein extension module,
<http://trific.ath.cx/resources/python/levenshtein/>
- cdb-compatible library[8], such as tinycdb,
<http://www.corpit.ru/mjt/tinycdb.html>
- python-cdb module,
<http://pilcrow.madison.wi.us/>

For end users, in order to access the database tables, only the cdb library is needed for the C interface, and python-cdb for the python interface.

8 Structure of Constant Database Tables

There are four constant database tables created. First one `lemma2forms.cdb` contains lemmas as keys, with inflected word forms as values. Second table `lemma2tagforms.cdb` has again lemmas per keys, but the values contain tags together with inflected forms (as one string, joined by `\t` tabulator character). The third table `form2lemma.cdb` contains inflected word forms for keys, with all possible lemmas as values for a given key, and the fourth table `form2taglemma.cdb` has inflected word forms for keys and tags joined with lemmas as values.

9 C API

C-based searching is provided by a convenient library `mlv_libquery`. The library intentionally mimics the usage of cdb library and provides following functions:

- ```
int mlv_init (char *table_file, struct cdb *cdb);
```
- Initialises structure `cdb`, using `table_file` as a file name of table that should be initialised. `table_file` should be one of "lemma2forms.cdb", "lemma2tagforms.cdb" or "form2lemma.cdb", optionally with a path specification. Returns 0 on success or a negative value on error.
- ```
void mlv_free (struct cdb *cdb);
```
- Releases internal structures holding information about an open table *and closes* the file associated.
- ```
int mlv_findinit (struct cdb_find *cdbf, struct cdb *cdb,
char *key);
```
- Initialises the searching structure `cdbf` to search for key string Returns positive value on success, negative on failure.
- ```
int mlv_findnext (struct cdb_find *cdbf, struct cdb *cdb,
char *val, int maxlen);
```
- Finds next (first if called right after `mlv_findinit`) matching key. Returns positive value if a given key was found, zero if there are no more such keys, and negative value on error. If the key was found, the value is put into `*val`, up to the `maxlen-1` characters, and the trailing `'\0'` is added to the string.

Code using the C library should include "mlv_libquery.h" and `<cdb.h>` headers.

```

#include <cdb.h>
#include "mlv_libquery.h"

void main(void) {
    struct cdb cdb;
    struct cdb_find cdbf;

    char val[255];
    int i;

    char *key = "mier";

    mlv_init("form2lemma.cdb", &cdb);
    mlv_findinit(&cdbf, &cdb, key);
    while (mlv_findnext(&cdbf, &cdb, &val, sizeof(val)) > 0) {
        printf("%s\n", val);
    }

    mlv_free(&cdb);
}

```

Listing 1.1. Get all the possible lemmas for a word *mier* – *miera*, *mier*, *mierit*

10 Python API

Python API is contained in the `mlv_query` module. The module contains one class, `MlvQuery`, providing a dictionary-like interface. The class' constructor takes one parameter, file name of a constant database table. Class instance supports `get`, `has_key`, `__getitem__`, `__iter__`, `__len__` and `__contains__` methods. The `__getitem__` method returns a generator iterating through all the values bound to the key.

11 Limits

The system, as described here, can conveniently handle suffix changes. While the prefix morphology can be in principle handled by Levenshtein operations, in practice it means creating a new paradigm template for each lemma with different length (since the positions of Levenshtein edit operations are counted from the end of the word), and therefore vastly increasing the number of paradigm templates. Of course, for a hypothetical language with prefix-only morphology, the system works well, if we remove the word reversing. However, for natural languages with mostly postfix morphology and only limited prefix morphology⁴,

⁴ For example, prefix morphology in many Slavic languages is limited to creating superlatives with the use of prefix *naj-*, *naŭ-* and verb and adjective negation with *ne-*, *ne-* or similar prefixes, often even masked by orthography and written separately.


```

from mlv_query import MlvQuery

q = MlvQuery("form2lemma.cdb")
print len(q) # number of entries in the table

for lemma in q["mier"]:
    # output is mier miera mierit
    print lemma.encode("utf-8"),

# test if word "mierou" is in the table
print q.has_key("mierou") # True or False
print "mierou" in q # the same as above

for word in q: # print all the word forms in the table
    print word.encode("utf-8")

```

Listing 1.2. Example of the Python API

the recommended way is to use a separate rule-based algorithm to deal with prefixes.

The system is suitable especially for languages that have reasonably complex suffix morphology with a reasonably large set of basic paradigm templates – a prime example of this are Slavic languages. In fact, we are deploying this system for Slovak language.

The system does not handle compound morphology. For languages having rich system of compound morphology (e. g. German), the system is suitable only to describe morphology of core vocabulary, and the compound word analysis has to be taken care of separately by other means.

For languages with mostly template morphology (Arabic, Hebrew), the system could be conveniently used if the word stem changes can be regularly described in terms of positions of changed graphemes, counted from the end of the word – both Arabic and Hebrew probably satisfy these requirements.

For agglutinative languages (Hungarian, Finnish, Turkish), the situation is a bit different. These languages tend to have very regular morphology, but thanks to the agglutinative nature, each lemma has a huge number of possible forms. Therefore, contrary to the situation of fusional languages, agglutinative languages would require to write huge paradigm templates, but on the other hand, the number of paradigm templates would be quite low – so the system might be quite usable here. However, thanks to the regularity of agglutinative morphology, it might be in fact less demanding to use rule-based algorithmic approach to the morphology analysis.

References

1. Левенштейн, В. И.: Двоичные коды с исправлением выпадений, вставок и замещений символов, Докл. АН СССР, 163, 4, (1965) 845–848.

2. Hajič, J., Hladká, B.: Czech Language Processing - POS Tagging. In: *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain: (1998) 931–936
3. Sedláček, R.: Morfologický analyzátor češtiny. PhD. thesis. Faculty of Informatics, Masaryk University Brno, (1999)
4. Free Software Foundation, Inc. (1989, 1991)
5. The Unicode Consortium. The Unicode Standard, Version 4.0 Boston, MA, Addison-Wesley Developers Press, ISBN 0-321-18578-1 (2003)
6. The Unicode Consortium. Unicode Technical Report #15: Unicode Normalization Forms. <http://www.unicode.org/unicode/reports/tr15/>
7. <http://www.python.org/>
8. <http://cr.yp.to/cdb.html>

Manual Morphological Annotation of Slovak Translation of Orwell's Novel 1984 – Methods and Findings

Radovan Garabík¹
Lucia Gianitsová-Ološtiaková²

¹ Eudovít Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia
korpus@korpus.juls.savba.sk
<http://korpus.juls.savba.sk>
² University of St. Cyril and Methodius
Trnava, Slovakia
gianitsova@zoznam.sk

Abstract. Manual morphological text annotation is indisputably an important part of building a framework of NLP tools used in corpora construction. From 2004 to 2005, the complete text of Orwell's 1984 novel, some Slovak Wikipedia texts and some newspaper articles have been annotated. In the paper we present the methodology used in manual annotation and correction of annotated data, and the discussion of obtained results.

Manual morphological text annotation of the Slovak National Corpus is a part of an intense work made as a part of constructing a corpus. It represents another processing of corpus data, providing rich information about language and its usage. The importance of exact manually annotated data for subsequent computer processing of morphology is indisputable. For that reason during the years 2003 – 2005 a great attention has been given to this phase of corpus construction.

During the introductory phase (in 2003) after the first theoretical discussions[1] about morphological tagging a tagset described in [2] has been designed.

The second phase, a manual annotation (2004 – 2005) started after confrontation with a real text material, using the annotation rules described in [3]. From February 2004 to June 2005 a manual lemmatization and tagging have been carried out using the complete texts of the Orwell's novel *1984*, samples from *InZine* (internet magazine), *Wikipedia* (internet encyclopædia) and *SME* (daily newspaper). The annotation was done by students of the Faculty of Philosophy, Comenius University, Bratislava. The number of students varied from 2 at the beginning to 11 at the end. Though manual annotation is a time-consuming work, following texts containing 215 000 tokens have been annotated: Orwell's *1984* (102 000 tokens), Slovak *Wikipedia* (50 000 tokens), *SME* daily (about 21 000 tokens), internet magazine *InZine* (more than 42 000 tokens).

The paper deals with our experiences with manual annotation acquired during annotation of Orwell's novel *1984*. Attention is also paid to the description of some fundamental methods applied to correction and finalization of manual annotations.

The files being annotated are conforming to XML TEI XCES standard[4]. We have created a GUI program written in python-gtk, using ElementTree library to parse and modify the XML files[5], used to manually annotate the files, called *Anno*. The program displays list of words (tokens) in the file, and for each selected token a list of possible lemmas and tags. The user either selects a corresponding pair of lemma and tag (disambiguation), or if none are provided or suitable, he/she can enter or fix the lemma and tag directly.

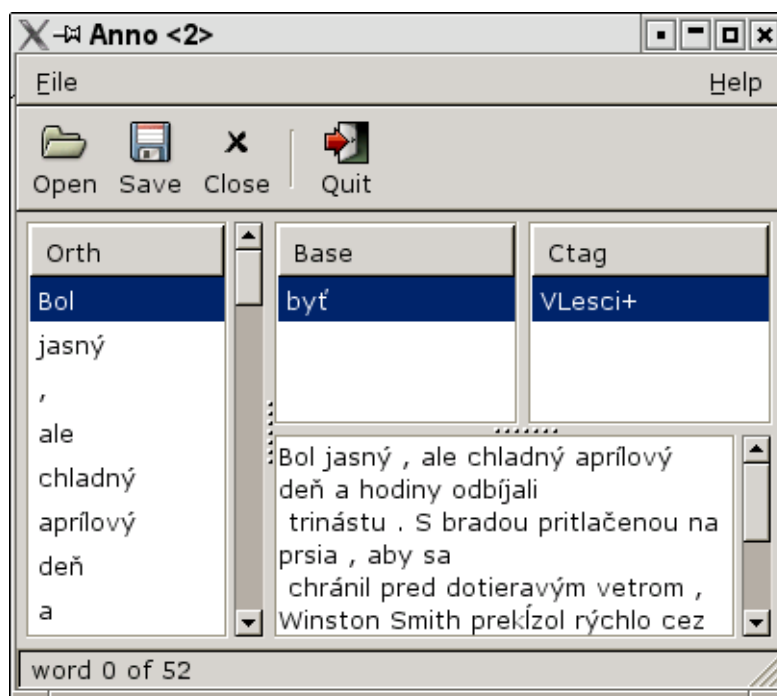


Fig. 1. Anotation tool Anno

At the beginning the annotation speed was about 80 tokens per hour, but after the annotation tool has been tuned according to the requirements of the annotators, a remarkable acceleration of the annotation (up to 200 – 250 tokens/hour) occurred. The possibility to work with automatically pre-tagged texts has been extremely advantageous. Pre-tagging has been done by using the morphological tagger described in [6], with combination with the TNT statistical tagger[7] trained on already annotated texts. Consequently, the annotators needed to focus their attention only at verification and correction of lemmas and tags. This significantly increased the accuracy of a manual annotation. At the beginning it was 84 per cent (partly due to changes being made in the tagset specification and annotation principles). After transition into a new annotation tool the accuracy of manual annotation increased to 92 per cent.

Each annotator was given a part of the novel 1984, at the beginning in chunks of text containing about 100 – 200 words, later expanded to 500 and more words. Completed parts have been gathered into a single-unit text, which became a subject of verification and unification of various token interpretations. During the first annotation phase (first 20 000 tokens) each annotator obtained one chunk of text, and then the whole text was checked by a linguist responsible for morphological annotation. Out of the corpus with 21 500 tokens there were 2 952 mistakes (14 per cent of the annotated text) detected. However, this method turned out not to be very effective because of its slow speed (1000 – 1200 tokens/day) and a high demand put on the linguist. Moreover, with increasing the text length the percentage of unspotted mistakes was increasing. Later, it was found out that 350 mistakes and incorrect interpretations (1.6 per cent of the annotated text) have not been detected.

However, later we used the method used when annotating the texts in the Prague Dependency Treebank – each file is annotated by two persons and results are automatically compared and subsequently checked and disambiguated only by one annotator[8]. Consequently, in the following phase the text samples have been given to two annotators and we focused on correcting just the differences in the annotation, with the use of command line utility `diffxc.es.py`, providing a diff(1)-like comparison of two XCES files.

1	OdkiaIsi OdkiaIsi	odkiaIsi odkiaIsi	Dx PD
17	Kávy Kávy	káva káva	SSfs2 SSfs2x:r
21	sa sa	sa sa	R Z
47	akoby akoby	akoby akoby	OY O
85	byť byť	byť byť	VIe+ VKe+
107	jedine jedine	jedine jedine	Dx T
124	na na	na na	Eu6 Eu4
153	aj aj	aj aj	O T
160	* *	* *	# Z
186	váčšmi váčšmi	váčšmi veľa	Dx Dy

Table 1. Example of output of comparing two XCES files

Unfortunately, this method also turned out to be inconvenient, because often the detected errors had origin in an insufficient practical morphological skill of one of the annotators (it is represented by 1118 tokens out of 15 061 tokens and it makes 58 per cent of detected differences). We needed just to have the text annotated by a different annotator and then verify only his/her annotation. Moreover, a comparison of controversial cases and correction of those which really required it (from original 1921 differences only 803 ones required correction) has been a time-consuming work. Accordingly, results indicated the effectiveness of manual annotation up to 95 per cent. On the other hand, many mistakes have not been detected because annotators often made the same mistakes. These misinterpretations occurred especially in cases of part-of-speech homonymy – conjunctions and particles, adverbs and particles, in cases of wrong indication of homonymous nominative and accusative, genitive and accusative and similar grammar categories. In the sample the number of non-detected mistakes was 387, i. e. about 3 per cent of all the tokens. This kind of mistakes is typical of Slovak language grammar analysis, regardless of the linguistic level of the person doing the analysis – from elementary school pupils up to the university students. This reason reduced the annotation accuracy down to 92 per cent and was the main reason for the fact that a comparison of two annotations eliminates on average only 67 per cent out of all mistakes, the remaining 33 per cent is not detected.

Third phase of a final verification after various experiments started in January 2005. We made use of additional semi-automatized verification tools, to check out the annotated files. However, these tools have to be supplemented by a manual correction anyway. Each tool was designed to check out one specific class of mistakes. Our verification process included three phases:

1st phase, using tool named `checkxcestags.py`: removal of superfluous whitespace in lemmas (in the table below tokens number 267 and 2932), automatic checking of correct tag length and correct combination of characters in tags, e. g. a missing tag (token number 11637), an unknown tag (token number 7818), a missing tag for the level of adjectives (token number 1333), a missing tag for the congruence in gender of -l- participle (token number 503), an unknown tag for the category (token number 5856), a redundant tag (tokens number 126 and 3057), missing tags for categories (token number 2990), inappropriate gender for the given pronoun type, or person of verbs (tokens number 651 and 2500), wrong type of paradigm (token number 3332):

126	nejakej	nejaký	PAfs6x	Bad length
267	"	"	Z	Spaces in lemma/orth
503	bola	byť	VLesc+	Bad length
651	mi	ja	PPms3	Bad gender
1333	tučné	tučný	AAfp4	Bad length
1456	sú	byť	VKefp+	Bad number
2500	zažili	zažiť	VLdpbm+	Bad gender
2932	-	-	Z	Spaces in lemma/orth
2990	niekoľko	niekoľko	PU	Bad length
3057	deviatich	deväť	NUip2w	Bad length
3332	ich	on	PPmp4	Bad gender
5856	pohrá	pohrať VKmsc+		Bad aspect
6057	služia	služiť VKepci+		Bad length
7818	II	II	C}-----	Bad POS
11637	,	,	None	Not string

This method made it possible to eliminate some repeating mistakes and obvious incorrect interpretations of tagging manual. The tool is based on some general properties of certain grammar categories encoded in the tag. Out of all the mistakes, 28% were corrected in this phase.

2nd phase: We generated lists of unique triplets (token, lemma, tag) from the text, using tool named `cesstat-tab.py`. We then sorted the list either by lemma or by the tag, thus making it possible to easily spot any discrepancies. This phase decreased significantly tag assignment inconsistency, most notably mistakes with wrong indication of paradigm, gender, case or number. Overall, we corrected about 31 % of all the mistakes using this method.

Lemma	Token	Tag	Correction
akoby	Akoby	O	tag = OY
akoby	akoby	OY	
blízky	bližšie	AAfp1x	tag = AAfp1y
blízky	bližšie	AAns4y	
byť	bude	VBesc+	
byť	bude	VKesc+	tag = VBesc+
celý	celé	AAns4x	
celý	Celý	AAns4x	tag = AAis4x
čo	čo	PD	tag = PFns1
čo	čo	PFns1	
dav	dav	SSis4	
dav	dav	SSms1	tag = SSis1
do	do	Eu2	
do	do	Eu4	tag = Eu2
hľadieť	hľadelo	VLescn+	
hľadieť	hľadieť	VId+	tag = VIE+
indický	Indický	AAis4x	
indický	Indického	AAis2x:r	tag = AAis2x
iný	iných	AAmp2x	tag = PAmp2
iný	iný	PAms1	
katharine	Katharine	SSfs4:r	tag = SUfs4:r
katharine	Katharine	SUfs1:r	
každý	Každé	PAms4	
každý	každého	NAns2	tag = PAns2
nedefinovateľný	nedefinovateľnými	Gtip7x	tag = AAip7x
nedefinovateľný	nedefinovateľného	AAns2x	
niekoľko	niekoľko	NUns4	tag = PUns4
niekoľko	niekoľko	PUns4	
otvorený	otvorený	Gtis4x	
otvorený	otvorenou	AAfs7x	tag = Gtfs7x
predtým	predtým	Dx	
predtým	predtým	PD	tag = Dx
prsia	prsia	SSfp4	tag = SSnp4
prsia	prsia	SSnp1	
winston	Winstona	SSms2:r	
winston	Winstonom	SSms7	tag = SSms7:r

Table 2. List of triplets sorted by the lemma

Tag	Lemma	Token	Correction
AAmplx	bezpečný	bezpeční	
AAmplx	mladý	mladí	
AAmplx	mladý	mladý	tag = AAmsslx
Dx	celkom	celkom	
Dx	celok	celkom	lema = celkom
Eu6	v	V	
Eu6	v	vo	tag = Ev6
Gtfs1x	preťať	preťatá	lema = preťatý
PAis7	ktorý	ktorý	tag = PAisl
PAis7	nejaký	nejakým	
PAis7	niektorý	niektorým	
PAis7	nijaký	nijakým	
PFmp1	ten	tí	
PFmp1	ten	tých	tag = PFmp2
SSfs4	zákonnosť	zákonnosť	
SSfs4	záležitosť	záležitosť	
SSfs4	záležitosť	záležitosťi	tag = SSfp4
SSms1	pán	pán	
SSms1	pán	pána	tag = SSms2
VIe-	necivieť	necivieť	
VIe-	neexistovať	neexistoval	tag = VLescm-
VIe-	nemyslieť	nemyslieť	
VKdpc+	pobiť	pobijú	
VKdpc+	podariť	podarí	tag = VKdsc+
VKdpc+	pokaziť	pokazia	
VKdsc+	vybrať	vyberie	
VKdsc+	vyčistiť	vyčistím	tag = VKdsa+
VKdsc+	vydobiť	vydobije	

Table 3. List of triplets sorted by the tag

After the final text sample (about 40 000 tokens) has been corrected, this verification method has been replaced by a method of generating only a list of those triplets (token, lemma, tag) that did not occur in previously corrected texts. A pair of tools have been used for this purpose. The first one, `make3.py`, makes a pickled list of all existing triplets from the XCES files given as parameters to the program, and subsequently the second program, `check3.py` loads the pickled list and prints the triplets that are present in a XCES file given as a parameter but that are not present in the pickled lists. The annotator then verifies only these suspicious triplets. This phase had significantly reduced the number of inconsistencies including token interpretations and some mistakes not detected during a routine check.

3rd phase: quick visual check of annotated text, using annotation tool.

In this phase the attention was focused upon mistakes where the correct grammar categories can be found out only taking into account context of the word (case, person,

part-of-speech homonymy). The speed of this checking was about 1500 – 2000 tokens per hour and remaining 41 per cent of all mistakes were removed.

After implementation of the presented verification model from January to June 2005 more than 102 000 tokens were checked and corrected, i. e. the whole Orwell's novel 1984. Currently, the Slovak National Corpus uses this methodology for verification of further manual morphology annotation. In our opinion, this system proved to be able to provide positive results and improved texts verification. Its advantages could be seen especially in an implementation of semi-automatised methods that interactively (along with the manual control) participate in detecting ambiguities of manual annotation.

Acknowledgements

Publication of this article has been a part of the grant Morphosyntactic research in the Slovak National Corpus, MŠ SR VEGA 1/3149/04

References

1. Forróová, M., Horák, A.: Morfológická anotácia korpusu. In: *Proceedings of the Conference Slovenčina na začiatku 21. storočia*. Prešov, Fakulta humanitných a prírodných vied PU (2004) 174–183
2. Forróová, M., Garabík, R., Gianitsová, L., Horák, A., Šimková M.: Návrh morfológického tagsetu SNK. In: *Proceedings of International Conference Slovko 2003 – Slovanské jazyky v počítačovom spracovaní*. Bratislava (2003). To be published.
3. Garabík, R., Gianitsová, L., Horák, A., Šimková M.: Tokenizácia, lematizácia a morfológická anotácia Slovenského národného korpusu. (Current version of May 4, 2004) SNK JÚLŠ, Bratislava.
<http://korpus.juls.savba.sk/publikacie/Tagset-aktualny.pdf>
4. Ide, N., Bonhome, P., Romary, L., XCES: An XML-based Encoding Standard for Linguistic Corpora. In: *Proceedings of the Second International Language Resources and Evaluation conference*. Paris, European Language Resources Association (2000)
5. Garabík, R.: Processing XML Text with Python and ElementTree – a Practical Experience. Bratislava, E. Štúr Institute of Linguistics (2005). To be published.
6. Hajič, J., Hric, J., Kuboň, V.: Machine Translation of Very Close Languages. In: *Proceedings of the ANLP 2000*. Seattle, U.S.A. (2000)
7. Brants, T.: TnT – a statistical part-of-speech tagger. In: *Proceedings of the the ANLP 2000*. Seattle, U.S.A. (2000)
8. Hajič, J., Hladká, B., Pajas, P.: The Prague Dependency Treebank: Annotation Structure and Support. In: *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia, University of Pennsylvania (2001) 105–114

Contribution to Processing of Slovak Language at DCI FEEI TUKE

Ján Genčí

Department of Computers and Informatics
Technical University of Košice
genci@tuke.sk

Abstract. Paper describes some works in the area of computer linguistics at the Department of Computers and Informatics (DCI), Technical University of Košice (TUKE). It presents algorithm and set of regular expressions for adjectives compare. More precisely papers presents database model of word database with morphological information, suitable for stemmer implementation, including actual content of database. As a last part, paper presents project Synset Builder – experimental tool for building English-Slovak synsets based on the WordNet database, its theoretical foundation and future developments.

1 Introduction

There were some works regarding computational linguistics done at the DCI TUKE during several last years, as presented at the workshop Slovko '03. In the last two years we have started several new projects both based on the previous experience and some external requests and inspiration.

Among such projects are – comparison of adjectives based on the regular expressions, project of morphological database of Slovak language, transcription of information presented in the numerical form to textual form and project Synset Builder. Mentioned projects will be presented more deeply in the next sections.

2 Comparison of Adjectives Based on Regular Expressions

Comparison of adjectives for Slovak language is specified in the comprehensive form in several linguistics sources, i.e. [PSP92], [KSSJ97]. In the literature dealing with morphology (i.e. [Dvonč66]) set of rules for adjective compare can be found, but these rules cannot be expressed algorithmically.

To building morphological database of Slovak language (see bellow), we decided to store full words in the database, without using algorithmic approach. However, to fill the database, we use, if possible, algorithmic approach to generate word, together with their attributes.

Generally, according literature, 2nd declension of Slovak adjective is formed by adding suffixes -ší and -ejší to the root of the word. However, there are plenty of exceptions and irregularities in this process (i.e. word “biely” which is compared to “belší”).

The decision was to use regular expressions to select group of adjectives, which compare in the same way and specify the way of their compare.

Regular expressions. Regular expressions are pattern-matching expressions. They can be quite complex, but for our purposes small set of rules were sufficient. For readers, to understand specified rules, we briefly summarize them.

- List of characters is a regular expression (i.e. “expression”).
- Square brackets specify any character stated in the brackets can be at the position of brackets (i.e. “a[bc]z” specifies strings “abz” or “acz”).
- Character “^” as a first character in the square brackets mean “any other characters, as stated” (i.e. “a[^b]c” specifies any tree-letter character, except “abc”).
- Character “\$” means end of line.
- Character “^” as the first character means the beginning of the line.

Designed set of regular expressions consists of ordered 3-tuples (*RE*, #, *suffix*), where *RE* is regular expression for matching the word, # is the number of characters, which have to be stripped from the end of the word and *suffix* is the suffix added to stripped word. At this time we develop the set of regular expressions for adjective ended by “-y/ý”. The most of the words is matched by the expression

```
(“[bcčdfghjklmnňpqrsštřvxzž][cčdfghjkmnňpqrsštřvxzž][yý]$”,1,“ejší”)
```

which is stated at the end of the set. In front of it, there are several groups of 3-tuples which serves for treatment of exceptions (i.e. adjectives ended by “-py”, “-dy”, “-ry” etc.):

```
// ry
("skorý$",1,“ší”),
("starý$",1,“ší”),
("r[yý]$",1,“ejší”),
("r[yý]$",1,“ejší”),
// my
("strmý$",1,“ší”),
("m[yý]$",1,“ejší”),
// dy
("mladý$",1,“ší”),
("[eur]dý$",1,“ší”),
("d[yý]$",1,“ejší”),
...
```

Beginning of the set consists of the 3-tuples given mainly for full-words exceptions:

```
("^jeden-jediný$",0,“”),
("^krásny$",0,“”),
("^primalý$",0,“”),
("^dobrý$",0,“”),
```

```

("[ps]laný$",1,"šf"),
("^zdravý$",1,"šf"),
("^živý$",1,"šf"),
("^plný$",1,"šf"),
("^poloplňý$",1,"šf"),
("^hrôzyplňý$",1,"šf"),
("plňý$",1,"ejšf"),

```

Generated compared adjectives were validated by the MS-Word spellchecker. Next step validation is planned to be carried out by linguists.

3 Morphological Database of Slovak Language

For reduction of dimensionality of various type of classification tasks for Slovak texts stemmer of Slovak words is required. There are several projects, which less or more publicly declare availability of Slovak morphological database, which can provide data necessary for stemming. We can mention IStemSK [www02] provided by Forma s.r.o., project AJKA [www03] adopted to Slovak language and used in the project of Slovak National Corpus, or project MORFEO [www04]. But no of these databases is publicly available. This situation leads to decision to build publicly available morphological database [www06].

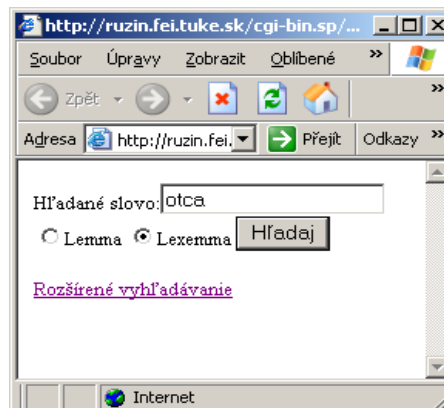


Fig. 1. Search option in the morphological database

Because of lack of linguistic support at the time the decision was made, we decided to design and implement flexible data model of our database to have an opportunity to store additional attributes of words in the future.

The system was implemented as a semestral work. Data was used from the previous projects. The illustrative output from the database is presented in the figures 1 and 2.

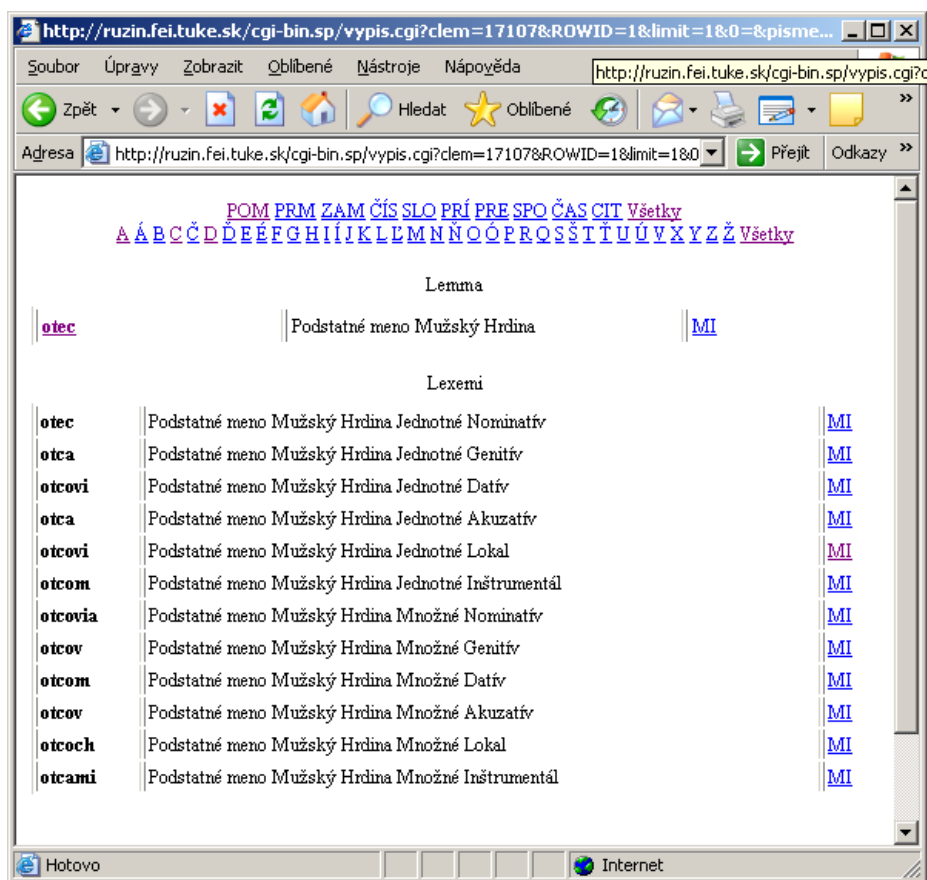


Fig. 2. Result of search

4 Synset Builder

Project EuroWordNet [www05] covers several European languages. However, it lacks Slovak language.

The aim of the Synset Builder [www06] project was to evaluate possibility to find Slovak equivalents to foreign words using available on-line dictionaries. During the project we decided to use the WordNet database, where the senses of the words are documented. Use of several on-line dictionaries for translation was also proposed. Our approach use intersection of translation of words from synsets stated in the WordNet database to find out relevant Slovak words equivalent to the synset sense.

Proposed approach is language independent, what we demonstrated by including Czech on-line dictionary to the project. It gives us opportunity to build (to some extent) Slovak-Czech dictionary based on WordNet senses.

Project was implemented as a diploma thesis and illustrative output is presented in the figures 3 and 4.

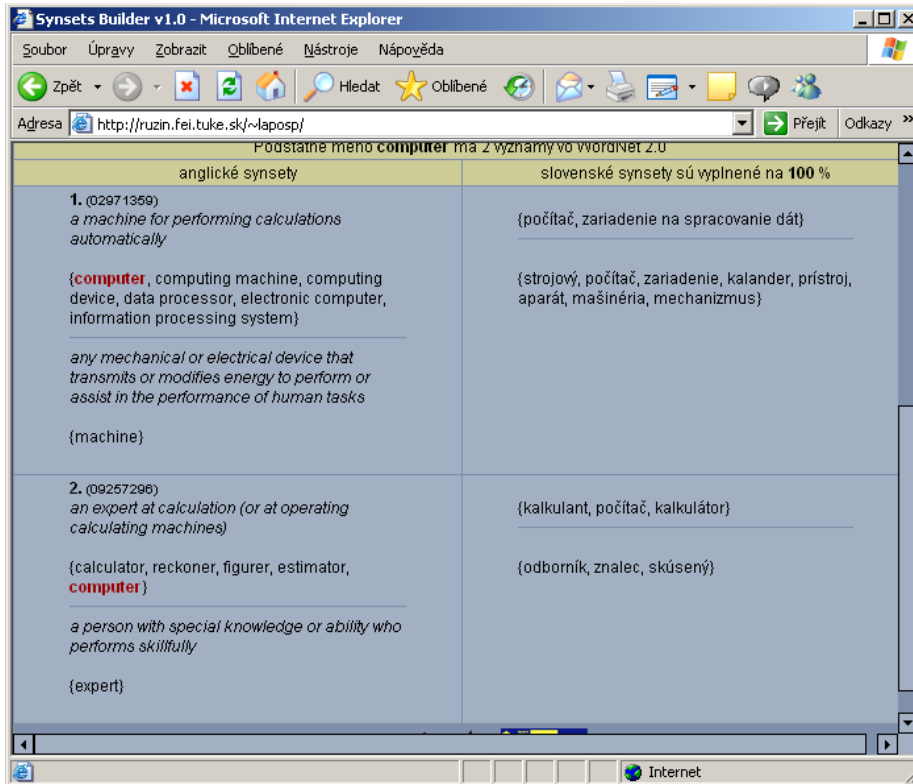


Fig. 3. Synsets for word “computer”

5 Conclusion and Future Work

Presented paper is a short description of the current projects carried out at the Department of the Computers and Informatics at Technical University of Košice.

The project “Comparison of Adjectives” needs linguistic validation. After that, we plan to use it as a source of data for the morphological database.

The project “Morphological Database” needs linguistics validation also. But because of large database we plan to use some automated validation with subsequent human validation.

The last project “Synset Builder” suffer from quality of data in the on-line dictionaries and/or even more from lack of availability of various different dictionaries.

We plan to evaluate our approach using “multilayer” dictionaries – dictionaries generated using one or more “internal” translations to get final translation (i.e. to get English-Slovak translation, we plan to use i.e. English-German-Slovak translations).

The screenshot shows a web browser window titled "Synsets Builder v1.0 - Microsoft Internet Explorer". The address bar shows "http://ruzin.fei.tuke.sk/~laposp/". The main content area displays a table comparing synsets for the word "computer" across three languages: English, Slovak, and Czech. The table is titled "Podstatné meno **computer** má 2 významy vo WordNet 2.0".

anglické synsety	slovenské synsety sú vyplnené na 100 %	české synsety sú vyplnené na 100 %
<p>1. (02971359) <i>a machine for performing calculations automatically</i></p> <p>{computer, computing machine, computing device, data processor, electronic computer, information processing system}</p> <p><i>any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks</i></p> <p>{machine}</p>	<p>{počítač, zariadenie na spracovanie dát}</p> <p>{strojový, počítač, zariadenie, kalander, prístroj, aparát, mašineria, mechanizmus}</p>	<p>{počítač}</p> <p>{počítač, mašinerie}</p>
<p>2. (09257296) <i>an expert at calculation (or at operating calculating machines)</i></p>	<p>{kalkulant, počítač, kalkulátor}</p>	<p>{kalkulačka, kalkulant, počítáň}</p>

Fig. 4. English, Slovak and Czech Synsets

References

- [Dvonč66] Dvonč, L.: Morfológia spisovnej slovenčiny. 1966
- [PSP92] Kačala, J.: Pravidlá slovenského pravopisu. Bratislava, 1991
- [KSSJ97] Doruľa, J., et al.: Krátky slovník slovenského jazyka. Bratislava, 1997
- [www01] <http://wordnet.princeton.edu/>
- [www02] <http://www.forma.sk/prod-jm-istem.asp>
- [www03] <http://nlp.fi.muni.cz/projekty/ajka/>
- [www04] <http://www.morfeo.sk>
- [www05] <http://www.let.uva.nl/~ewn>
- [www06] <http://ruzin.fei.tuke.sk/~spireng/menu.html>
- [www07] <http://ruzin.fei.tuke.sk/~laposp/>

Towards a General Model of Grapheme Frequencies for Slavic Languages

Peter Grzybek¹ and Emmerich Kelih²

¹ Graz University, Austria, Dept. for Slavic Studies,
peter.grzybek@uni-graz.at,

² Graz University, Austria, Dept. for Slavic Studies,
emmerich.kelih@uni-graz.at,

WWW home page: <http://www-gewi.uni-graz.at/quanta>

Abstract. The present study discusses a possible theoretical model for grapheme frequencies of Slavic alphabets. Based on previous research on Slovene, Russian, and Slovak grapheme frequencies, the negative hypergeometric distribution is presented as a model, adequate for various Slavic languages. Additionally, arguments are provided in favor of the assumption that the parameters of this model can be interpreted with recourse to inventory size.

1 Graphemes and Their Frequencies

The study of grapheme frequencies has been a relevant research object for a long time. From a historical perspective, only a small part of the studies along this line have been confined to the mere documentation of grapheme frequencies, considering this to be the immediate object and ultimate result of research. Other approaches have considered the establishment of grapheme frequencies to be the basis for concrete applications. In fact, relevant studies in this direction have often been motivated or accompanied by an interest in rather practical issues such as, for example, the optimization of technical devices, the structure of codes and processes of information transfer, cryptographical matters, etc.

A third line of work on grapheme frequencies has been less practically and more theoretically oriented. In this framework, research has recently received increasing attention from quantitative linguistics. As compared to the studies outlined above, the focus of this renewed interest has shifted: In a properly designed quantitative study, counting letters (or graphemes), presenting the corresponding absolute (or relative) frequencies in tables, or illustrating the results obtained in figures, is not more and not less but one particular step. In this framework, data sampling is part of the empirical testing of a previously established hypothesis, motivated by linguistic research and translated into statistical terms. The empirical testing thus provides the basis for a decision as to the initial hypothesis, and on the basis of their statistical interpretation one can strive for a linguistic interpretation of the results (cf. Altmann 1972, 1973).

Providing and presenting data thus is part of scientific research, and it is a necessary pre-condition for theoretical models to be developed or elaborated. As

far as such a theoretical perspective is concerned, then, there are, from a historical perspective (for a history of studies on grapheme frequencies in Russian, which may serve as an example, here, cf. Grzybek & Kelih 2003), two major directions in this field of research. Given the frequency of graphemes, based on a particular sample, one may predominantly be interested in

1. comparing the frequency of a particular grapheme with its frequency in another sample (or in other samples); the focus will thus be on the frequency analysis of individual graphemes;
2. comparing the frequencies of all graphemes in their mutual relationship, both for individual samples and across samples; the focus will thus be on the analysis and testing of an underlying frequency distribution model; this approach includes – if possible – the interpretation of the parameters of the model.

In our studies, we follow the second of these two courses. We are less interested in the frequency of individual graphemes. Rather, our general assumption is that the frequency with which graphemes in a given sample (text, or corpus, etc.) occur, is not accidental, but regulated by particular rules. More specifically, our hypothesis says that this rule, in case of graphemes, works relatively independent of the specific data quality (i.e., with individual texts as well as with text segments, cumulations, mixtures, and corpora). Translating this hypothesis into the language of statistics, we claim that the interrelation between the individual frequency classes is governed by a wider class of distributions characterized by the proportionality relation given in (1):

$$P_x \sim g(x)P_{x-1} , \quad (1)$$

relating a given class to previous classes, or by a partial sums relation, thus relating a class to the subsequent classes.

Thus, as opposed to studies focusing on the frequency of individual graphemes, the accent is on the systematic relation between the frequencies of all graphemes (or rather, the frequency classes) of a particular sample. Research thus is interested in the systematic aspects of frequencies, concentrating on the (relative) frequency of the most frequent grapheme, as compared to the second, third, etc. It is thus the study of the rank frequency distribution of graphemes in various texts and languages, which stands in the focus of attention. The objective is the theoretical modeling and mathematical formalization of the distances between the individual frequencies, irrespective of the specific grapheme(s) involved. Consequently, the procedure is as follows: If one transforms the raw data obtained into a (usually decreasing) rank order, and connects the data points with each other, one usually obtains not a linear decline, but a specific, monotonously decreasing (usually hyperbolic) curve. The objective then is to model the specific form of this curve, and to test, if the frequencies in different samples (i.e., the specific decline of the frequencies) display one and the same form, or not.

Thus far, convincing evidence has been accumulated to corroborate this hypothesis for three of the Slavic languages: Slovene, Russian, and Slovak. The

basic results have been presented in detail elsewhere – cf. Grzybek, Kelih, & Altmann (2004) for Russian, Grzybek & Kelih (2003) for Slovene, and Grzybek, Kelih & Altmann (2005a,b) for Slovak. The present contribution is a first attempt to arrive at some synopsis and to develop some generalizing conclusions. Therefore, it will be necessary to briefly present the results hitherto obtained by way of some summary, before we turn to a synopsis of these results, which will ultimately lead to some hypothesis for further studies.

2 A Model for Grapheme Frequency Distributions in Slavic Languages

In our endeavor to find an adequate theoretical model, we have concentrated on discrete frequency distribution models, rather than on continuous curves – for methodological reasons, which need not be discussed here. In order to test the goodness of fit of the models tested, we have employed χ^2 tests. This traditional procedure is problematic, however, since the χ^2 value linearly increases with sample size, the χ^2 value thus becoming sooner significant – and in case of grapheme studies, we are almost always concerned with large samples. Therefore, we have relativized the latter by calculating the discrepancy coefficient $C = \chi^2/N$, considering a value of $C < 0.02$ to be a good, a value of $C < 0.01$ a very good fitting.

As to the models tested, we did not expect that one and the same model would be universally relevant, i.e. would be able to cover all languages of the world. We did not even assume that one model would be sufficient to cover all those (Slavic) languages which were the objective of our study. Therefore we have tested all those models which have been favored as successful rank frequency models in the past. Specifically, we tested the following distribution models (for details, cf. the studies mentioned above):

1. Zipf (zeta) distribution;
2. Zipf-Mandelbrot distribution;
3. geometric distribution;
4. Good distribution;
5. Whitworth distribution;
6. negative hypergeometric distribution.

It would be beyond the scope of the present paper to discuss the mathematical details of these distribution models, or the theoretical interrelations between them (cf. Grzybek, Kelih & Altmann 2004). Rather, it should be sufficient to summarize that for all three languages mentioned above, we found that the organization of the grapheme frequencies followed none of the traditionally discussed models. Rather, it was the negative hypergeometric distribution (*NHG*) – and only this model¹ – which turned out to be adequate; quite unexpectedly, all

¹ It should be noted that the allegedly exclusive validity of the *NHG* distribution as a theoretical model claimed here relates only to the data we have analyzed thus far.

other models did not fulfill the above-mentioned criteria and thus had to be ruled out as adequate models.² Therefore, the *NHG* distribution should briefly be presented here. It may be derived in different ways; here, it may suffice to interpret it with recourse to Wimmer & Altmann's (2005a,b) *Unified Derivation of Some Linguistic Laws*, namely, in the form of equation (2):

$$P_x = \left(1 + a_0 + \frac{a_1}{(x + b_1)^{c_1}} + \frac{a_2}{(x + b_2)^{c_2}} \right) P_{x-1} \quad (2)$$

Inserting in (2)

$$\begin{aligned} a_0 &= b_2 = 0, \\ a_1 &= (-K + M + 1)(K + n - 1)/(-K + M - n), \\ a_2 &= (n + 1)(M - 1)/(K - M + n), \\ b_1 &= -K + M - n, \\ b_2 &= 0, K > M \geq 0, n \in \{0, 1, \dots\}, \quad c_1 = c_2 = 1 \end{aligned}$$

one obtains equation (3):

$$P_x = \frac{(M + x - 1)(n - x + 1)}{x(K - M + n - x)} P_{x-1} \quad (3)$$

from which the *NHG* results (with $x = 0, 1, \dots, n$, $K > M > 0$, and $n \in \{1, 2, \dots\}$), as given in equation (4):

$$P_x = \frac{\binom{M + x - 1}{x} \binom{K - M + n - x - 1}{n - x}}{\binom{K + n - 1}{n}} \quad (4)$$

Since in case of rank frequency distribution, the first class is $x = 1$, the *NHG* has to be used in its 1-displaced form, as displayed in equation (5), with $x = 1, 2, \dots, n + 1$, $K > M > 0$, and $n \in \{1, 2, \dots\}$,

$$P_x = \frac{\binom{M + x - 2}{x - 1} \binom{K - M + n - x}{n - x + 1}}{\binom{K + n - 1}{n}} \quad (5)$$

This does not principally rule out all other models as possibly being relevant, and this is not to be misunderstood as a claim for a single universal model. Rather, there may be transitions between various model, or covergencies between them, and it is a matter of boundary conditions to be controlled in each single study, if one of the above-mentioned model, or eventually even other models not mentioned here, are more adequate.

² Only in case of Russian, the Whitworth distribution which, under particular conditions, is a special case of the *NHG* (in its partial sums form), turned out to be an adequate model, too.

3 Three Case Studies: Russian, Slovene, Slovak

Thus far, the results of four case studies have been reported which were conducted to test the model described above. In the case study involving Russian (Grzybek, Kelih, & Altmann 2004), 37 samples composed of different genres were analyzed. The text corpus included literary texts by A.S. Puškin, L.N. Tolstoj, F.M. Dostoevskij, and A.P. Čechov, as well as a number of scientific texts. In order to control the factor of text homogeneity, all texts were individually analyzed as homogeneous texts. Additionally, text segments, mixtures, and cumulations were artificially formed on the basis of these texts and analyzed in this form, as well. Finally, they were put together and to build a complete corpus of ca. 8.7 million graphemes and analyzed as such.

As a result, the *NHG* distribution turned out to be an adequate model for all 37 samples, with a discrepancy coefficient of $C < 0.02$ for each of them. Figure 1 illustrates the result for the complete corpus, where fitting the *NHG* distribution resulted in a discrepancy coefficient value of $C = 0.0043$.

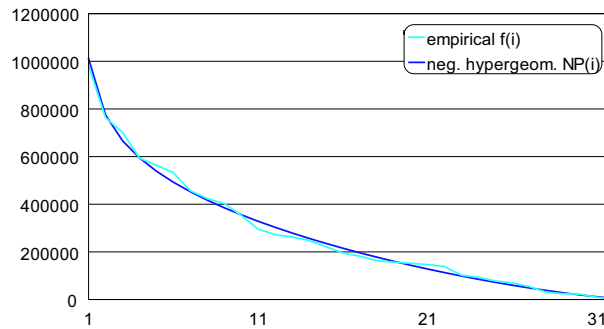


Fig. 1. Fitting the *NHG* Distribution to Russian Corpus Data

In the Russian study, a first interesting observation was made as to the parameters of the *NHG* distribution: Apart from parameter n – which, with $n - 1$, directly depends on the inventory size and thus is for all cases is constantly $n = 32 = 31$ in the case of Russian with its 32 different graphemes³ –, also

³ If one counts the Russian letter ‘ë’ as a separate letter, instead of realizing it as an allograph of the letter ‘e’, the inventory size of the Russian alphabet increases to 33, of course. It is evident that, as soon as inventory size comes into play as an influencing parameter when fitting a given distribution to particular data, this question may turn out to be relevant for the results obtained. Therefore, in order to control this factor systematically, Grzybek, Kelih & Altmann (2006) have re-run their analysis of Russian material under three different conditions in thirty homogeneous texts: (a) texts in which the Russian letter ‘ë’ does not occur ($n = 32$), (b) texts containing the letter ‘ë’ ($n = 33$), and (c) the same texts as in (b), thus in principle containing the letter ‘ë’, but the latter a posteriori being transformed to ‘e’ ($n = 32$) for the

parameters K and M seemed to display a relative constancy across all samples (with $K \approx 3.16$ and $M \approx 0.82$), K ranging from $2.95 \leq K \leq 3.42$, and M ranging from $0.77 \leq M \leq 0.85$. Figure 2 illustrates the observed constancy of the results obtained, with $0.043 \leq C \leq 0.0169$.

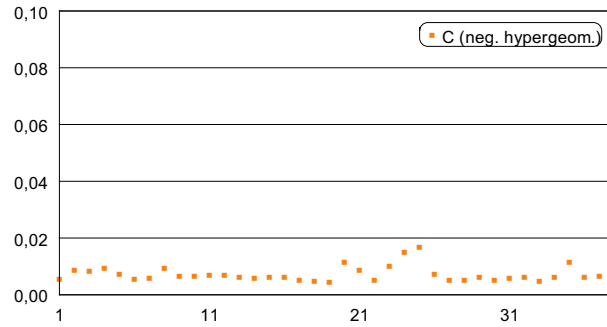


Fig. 2. C Values for Fitting the NHG distribution to Russian data)

Given these findings of the Russian case study, the idea was born to study the problem systematically for other Slavic alphabets, too. In this respect, Russian with its 32 (or 33) letters, has to be considered as having a medium inventory size as compared to other Slavic languages. Slovene, in turn, with its 25 letters, represents the minimum inventory size, and Slovak, with its 46 letters, is located at the upper end of the scale.⁴

In the Slovene study (Grzybek & Kelih 2003), twenty samples were analyzed, including literary texts and letters by Ivan Cankar, France Prešeren, Fran Levstik, as well as journalistic texts from the journal *Delo*; again, in addition to homogeneous texts, cumulations, segments and mixtures were artificially created and analyzed, as well as the complete corpus consisting of ca. 100.000 graphemes. As a result, the NHG distribution turned out to be the only adequate model for all samples: the discrepancy coefficient was $C < 0.02$ in all cases (with $C = 0.0094$ for the corpus).⁵

Again, for the Slovene data, too, the values of the parameters K and M of the NHG distribution turned out to be quite stable across all samples, with

analytic purpose described above.– Since these data have not yet been published, the present article is based on the results reported in Grzybek, Kelih, & Altmann (2005).

⁴ In case of Slovak, the inventory size decreases to 43, if one does not consider the digraphs ‘ch’, ‘dz’, and ‘dž’ to be separate letters in their own right.– Here, too, Grzybek, Kelih, & Altmann (2005a,b) conducted systematic studies to control the factor of defining the basic graphemic units.

⁵ For Slovene, too, Grzybek, Kelih, & Altmann (2005) have re-run their analyses, extending the data basis to thirty homogeneous texts. As in case of Russian, the present study is based on the results reported by Grzybek & Kelih (2003).

$K \approx 2.89$ and $M \approx 0.81$), K ranging from $2.79 \leq K \leq 3.01$, and M ranging from $0.80 \leq M \leq 0.83$. Interestingly enough, no significant difference was observed between the group of homogeneous texts, on the one hand, and the artificially composed text samples (segments, cumulations, mixtures), on the other hand, as far as the parameter values of K and M are concerned (the mean values being $\bar{K} = 2.89$ and $\bar{M} = 0.81$, for both groups of texts as well as for all samples jointly). Thus, on the level of graphemic organization, text heterogeneity does not seem to play a crucial role.

A comparative inspection of Figure 3 shows that for each of the languages, parameters K and M are relatively constant, but that the constancy of parameter K is realized on different levels, being slightly higher for Russian.

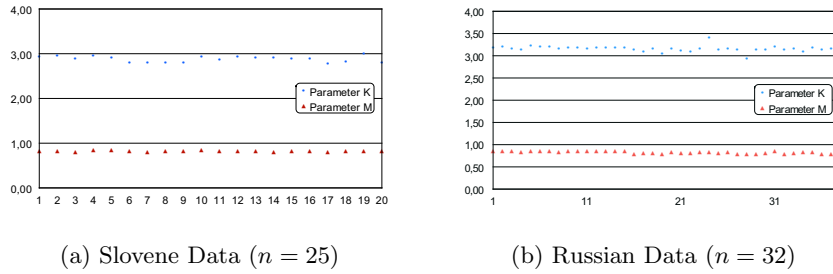


Fig. 3. Constancy of Parameter Values K and M (NHG distribution)

Given this observation, the hypothesis brought forth that not only parameter n of the NHG distribution, but also parameter K might be particular function of the inventory size. In this case, the analysis of Slovak data, should yield additional arguments in favor of this assumption. Consequently, two studies were conducted, based on thirty Slovak texts, summing up to a corpus of ca. 150.000 letters. In the first of these two studies (Grzybek, Kelih & Altmann 2005a), Slovak grapheme frequencies were analyzed without taking into consideration the above-mentioned digraphs, the inventory size thus being $n = 43$; in the second study (Grzybek, Kelih & Altmann 2005b), the same material was analyzed, this time counting digraphs as a category in its own right, the inventory size thus rising up to $n = 46$.

As a result, the NHG distribution once again turned out to be the only adequate model, under both conditions, with K and M displaying a relative constancy in either case. In case of the first study (with $n = 43$), the discrepancy coefficient was $C < 0.02$ in 28 of all 30 samples (with $C < 0.01$ in ten of the samples, and $C = 0.0102$ for the whole corpus); as to an interpretation of the finding that no good fitting was obtained for two of the samples, the authors referred to the fact that these two samples were extremely small with $N = 562$,

and $N = 446$ graphemes, respectively. Once again, the values of the parameters K and M of the NHG distribution were relatively constant across all samples, with $K \approx 4.07$ and $M \approx 0.85$, K ranging from $4.46 \leq K \leq 3.69$, and M ranging from $0.78 \leq M \leq 0.94$.

In case of the second study (with $n = 46$), the results were slightly worse, with a discrepancy coefficient of $C < 0.02$ in 25 of all 30 samples (with $C < 0.01$ in five of the samples, and $C = 0.0139$ for the whole corpus). Yet, with $K \approx 4.31$ and $M \approx 0.84$, K ranging from $4.86 \leq K \leq 3.81$, and M ranging from $0.76 \leq M \leq 0.92$.

Figure 4 illustrates the observed constancies of parameters K and M for both conditions.

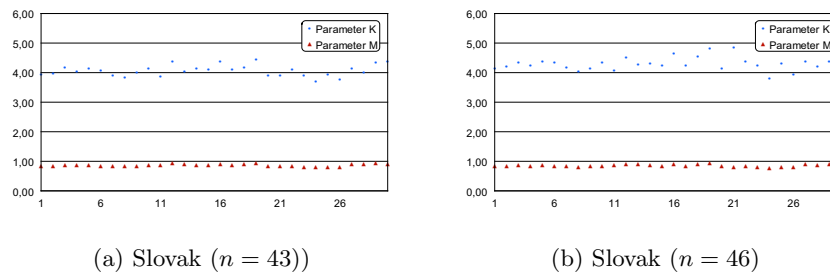


Fig. 4. Constancy of Parameters K and M (NHG distribution; Slovak data)

By way of a preliminary summary, one can thus say that the two Slovak studies yield two important results: first, the K values of the first study (with $n = 43$), is indeed lower as compared to those of the second study (with $n = 46$); and secondly, the Slovak K values, taken on the whole, are clearly higher as compared to those from the Slovene (with $n = 25$) and Russian (with $n = 32$) studies.

4 Consequences of the Single Case Studies

The four case studies reported above thus not only corroborated the initial hypothesis that the grapheme systems of the languages under study are systematically organized. Additionally, the findings clearly showed that the grapheme frequencies can be modelled with recourse to one and the same frequency distribution, namely, the NHG distribution. Furthermore, the results obtained gave rise to further hypotheses as to a possible interpretation of at least one of the parameters of this model, namely, parameter K .

Taking into account the results for each language separately, it first seemed that the two parameters K and M are both relatively constant within a given

language. However, as soon as data for all three languages were available, it could be seen that parameter K is definitely higher for a language with a larger inventory size, parameter M not displaying such a direct increase. Grzybek, Kelih, & Altmann (2005a) therefore assumed this to be a hint at the possible (direct or indirect) dependence of parameter K on inventory size, whereas parameter M still seemed to be relatively constant across languages. The assumption of a direct dependence of K on inventory size was therefore directly tested in Grzybek, Kelih, & Altmann's (2005a,b) studies on Slovak: For the sake of simplicity, they considered parameters K and M to be random variables with finite mean values and finite variances, and then compared the mean values of the parameters for both Slovak conditions ($n = 43$ vs. $n = 46$) by way of a t -test. As the results showed, parameter K is significantly higher for $n = 46$ as compared to $n = 43$ ($t_{FG=56} = 4.53$; $p < 0.001$). However, a comparison of the mean values of parameter M by way of a t -test showed that in this case, for both conditions ($n = 43$ vs. $n = 46$), there is no significant difference ($t_{FG=58} = 1.07$; $p = 0.29$).

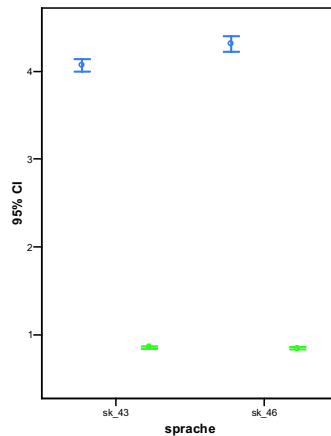


Fig. 5. Mean Values and Confidence Intervals for K and M (Slovak Data)

Fig. 5 illustrates the tendencies of both parameter values in form of a 95% confidence interval within which the relevant parameter may be expected with a 95% probability. It can easily be seen that parameter K clearly differs for both conditions ($n = 43$ vs. $n = 46$), whereas parameter M does not seem to vary significantly. The detailed Figure 6 additionally shows that the confidence intervals of K do not overlap, whereas they do for parameter M .

Whereas there is thus some evidence that parameter K may be directly related to inventory size, there is no such evidence with regard to parameter M . However, in their second study on Slovak graphemes, Grzybek, Kelih, & Altmann (2005b) found some other evidence of utmost importance, hinting at a direct relation between the two parameters, within a given language: under this

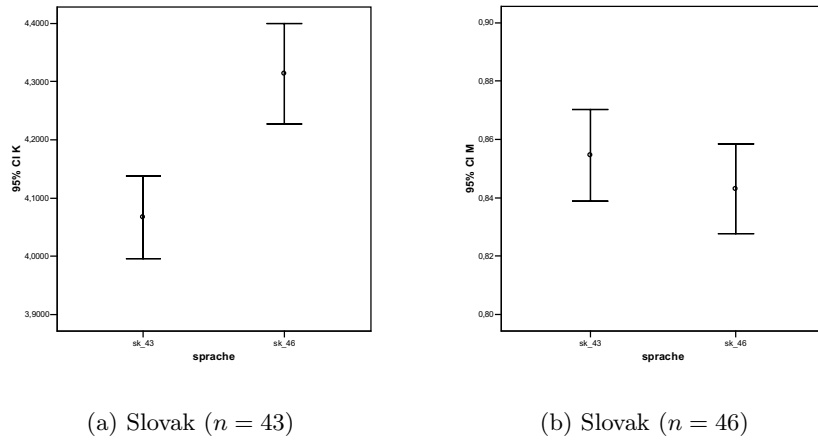


Fig. 6. 95% Confidence Intervals for Parameters K and M (Slovak Data)

condition (i.e., with $n = 46$), they found a highly significant correlation between K and M ($r = 0.59$, $p = 0.001$). In a re-analysis of the Slovak data with $n = 43$, the very same tendency could be found, the correlation even being more clearly expressed ($r = 0.83$, $p < 0.001$).

The interpretation arising thus is that one of the two parameters (K) is dependent on inventory size (and thus particularly relevant across languages), whereas the second parameter (M) is relevant within a given language. As Grzybek, Kelih, & Altmann (2005b) state, we are concerned here with a highly promising perspective: if the findings obtained could be corroborated on a broader basis, an interpretation of both parameters K and M would be at hand.

This assumption needs further testing, of course, and the present study is, as was said above, a very first step in this direction. As was said above, it would be too daring to utter far-reaching conclusions at this time, and if so, only with utmost caution. The four case studies reported above do not allow for solid generalizations; first, they imply some methodological problems, and second, the number of languages is too small for any extrapolation of the results obtained. Yet, the impression arises that not only the grapheme frequencies of each language per se are systematically organized, but also, in addition to this, the organization of the graphemic systems in general. One argument supporting this assumption is the fact that the grapheme frequencies of all three languages studied follow one and the same model; this is only a minor argument, however, since a model may well be a special case of a more general one, or it may converge to a related model. A major argument in favor of the assumption brought forth, then, is the possible interpretation of the parameters.

Yet, there seems to be sufficient evidence to generalize the results obtained in form of the derivation of some working hypotheses for future research.

5 From Case Studies to Systematic Research: Towards a Theory of Grapheme Frequencies

A first step in the direction outlined might thus be a comparative analysis of the four studies reported above. Table 1 presents the results obtained in a summarizing manner.

Table 1. Mean Parameter Values and Confidence Intervals

Language	n	Parameter K			Parameter M		
		\bar{K}	K_{\uparrow}	K_{\downarrow}	\bar{M}	M_{\uparrow}	M_{\downarrow}
Slovene	25	2.89	2.86	2.92	0.8115	0.8062	0.8168
Russian	32	3.16	3.14	3.19	0.8186	0.8105	0.8267
Slovak	43	4.07	4.00	4.14	0.8546	0.8389	0.8703
Slovak	46	4.31	4.23	4.40	0.8430	0.8276	0.8584

As a closer inspection of Table 1 shows, there seems to be a clear increase of parameter K with an increase of inventory size (n), whereas parameter M does not display a corresponding tendency; rather, parameter M seems to be rather constant across languages. Fig. 7 illustrates these two tendencies.

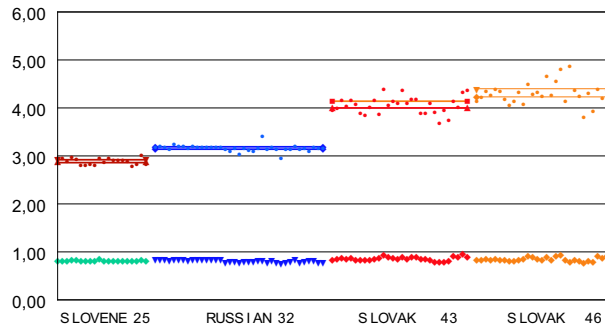


Fig. 7. Parameters K and M (With Confidence Interval) For Four Slavic Languages

Yet, as a statistical analysis shows, facts are more complex than it seems at first sight: Thus, calculating a bivariate correlation between the inventory size and the parameter values for K and M , results in a correlation coefficient of $r = 0.956$ (for K) and $r = 0.424$ (for M), both correlations being highly significant ($p < 0.001$), the correlation for K being more clearly expressed as compared to M . Figure 8 displays the result of regression analyses with inventory size as independent variable, K and M , respectively, as dependent variables.

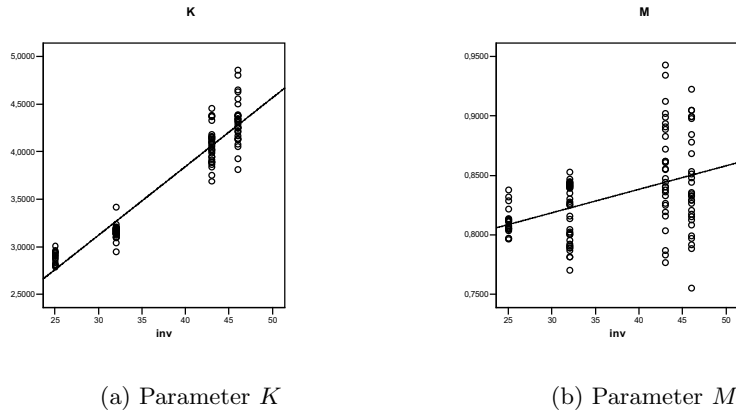


Fig. 8. Dependence of Parameters K and M on Inventory Size

The impression arising thus is that both K and M might depend on inventory size; this interpretation is weakened, however, or specified, by a closer analysis of the relation between both parameters. Given the finding that the correlation between parameter K and inventory size is expressed more clearly (see above), it seems reasonable to take into consideration the possibility that M is only indirectly dependent on inventory size, and directly on K . In fact, the correlation between K and M is highly significant ($r = 0.57$, $p < 0.001$). Figure 9 illustrates this tendency; as a closer inspection shows, however, the dependence seems to be much more clearly expressed not across languages, but within a given language.

This observation may then be interpreted in terms of a direct (linear) dependence of parameter K on inventory size n , and a direct (linear) dependence of parameter M on parameter K . Consequently, parameter M may be interpreted in terms of an indirect dependence on n . At this point, two perspectives emerge as possible orientations for future studies:

1. The first perspective is directed toward the study across languages; if in this respect, inventory size (n) is directly relevant for K , then it seems reasonable to concentrate on the mean values of K for each language (\bar{K}).
2. The second perspective concentrates on processes within a given language; if M indeed depends rather on K , within a given language, and less on n , then K must be studied for each language individually (K_i).

As was shown above, \bar{K} seems to be a linear function of n , thus being characterized by the equation $\bar{K} = h(N) = u \cdot N + v$. Furthermore, it now turns out that in fact M_i seems to be a linear function of K_i , within a given language,

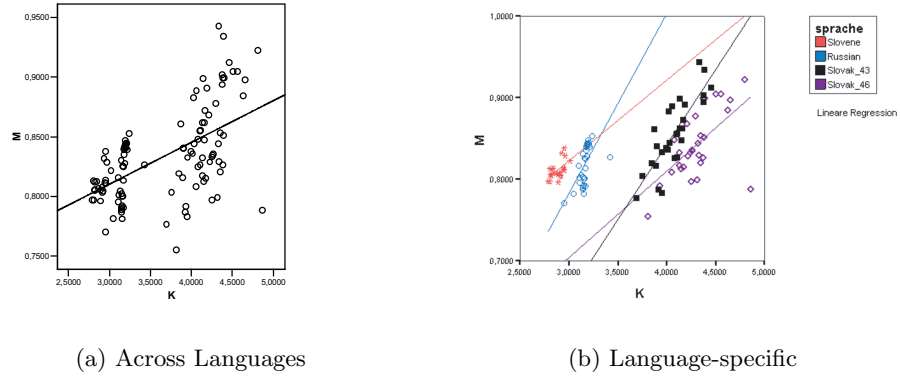


Fig. 9. Dependence of Parameter M on M

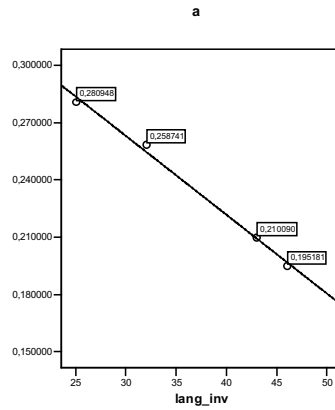
characterized by the linear function $M_i = a_i \cdot K_i$. Applying this formula to the data described above, one obtains the values represented in Table 2.

Table 2. Linear Dependences Between K and M

Language	n	\bar{K}	\bar{M}	a
Slovene	25	2.8874	.8115	.280948
Russian	32	3.1636	.8186	.258741
Slovak	43	4.0666	.8546	.210090
Slovak	46	4.3137	.8430	.195181

As a closer inspection of Table 2 shows, we are not yet at the end of our interpretations: quite obviously, a_i stands in a direct (linear) relation with n , which may be expressed by way of the formula $a_i = f(N) = c \cdot N + d$, the regression being almost perfect with a determination coefficient of $R^2 = .99$.

The observed tendency is illustratively presented in Figure 10), from which the linear decline of a with increasing inventory size becomes evident.

Fig. 10. a and n

6 Perspectives

It goes without saying, and it has been emphasized repeatedly, that at this moment these interpretations are rather daring. More material, and more systematically chosen material, must be analyzed to put our assumption on a more solid ground. Still, if additional evidence can be gathered for the plausible interpretations outlined above, a scheme as depicted in Table 3 might be derived to describe this situation.

Table 3. A General Schema of Dependences

$$\begin{array}{l}
 \bar{K} = h(N) = u \cdot N + v \\
 M_i = g(K_i) = a_i \cdot K_i \\
 a_i = f(N) = c \cdot N + d
 \end{array}$$

If the assumptions and hypotheses outlined above would indeed receive further support, we were in a lucky situation, which is highly desirable in quantitative linguistics, since we would be able to interpret all parameters of the theoretical distribution and thus have a qualitative interpretation. If the hypothesis brought forth above can be corroborated on a broader and more solid basis, including further (Slavic) languages, this might be relevant not only for linguistics. Ultimately, this would be a highly tricky mechanism from a broader perspective as well, relevant for systems theory and synergetics, in general: from this point of view, we are concerned with a low-level system of units relevant for the formation of higher-level units; on this low level the system's behavior is

determined merely by the inventory size of the units involved, and any variation on this level would be “corrected” by a second parameter, thus guaranteeing the system’s flexible stability.

Only thorough research can show if our assumptions stand further empirical testing – the fate of science, though. . .

References

1. Altmann, G.: Status und Ziele der quantitativen Sprachwissenschaft. In Jäger, S., ed.: *Linguistik und Statistik*. Vieweg, Braunschweig (1972) 1–9
2. Altmann, G.: *Mathematische Linguistik*. In Koch, W., ed.: *Perspektiven der Linguistik*. Kröner, Stuttgart (1973) 208–232
3. Grzybek, P., Kelih, E.: Grapheme Frequencies in Slovene – a Pilot Study. In Benko, V., ed.: *Slovko 2003, Bratislava (2003)* (to appear)
4. Grzybek, P., Kelih, E.: Grapheme Frequencies in Slovene. *Glottometrics* **12** (2006) (to appear)
5. Grzybek, P., Kelih, E.: Häufigkeiten von Buchstaben / Graphemen / Phonemen: Konvergenzen des Rangierungsverhaltens. *Glottometrics* **9** (2005) 62–73
6. Grzybek, P., Kelih, E.: Graphemhäufigkeiten (Am Beispiel des Russischen). Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. *Anzeiger für slawische Philologie* **31** (2003) 131–162
7. Grzybek, P., Kelih, E., Altmann, G.: Graphemhäufigkeiten im Slowakischen (Teil I: Ohne Digraphen). In Nemcová, E., ed.: *Philologia actualis slovacica*. UCM, Trnava (2005) (to appear)
8. Grzybek, P., Kelih, E., Altmann, G.: Graphemhäufigkeiten (Am Beispiel des Russischen). teil III: Systematische Verallgemeinerungen. *Anzeiger für slawische Philologie* **33** (2005) (to appear)
9. Grzybek, P., Kelih, E., Altmann, G.: Graphemhäufigkeiten im Slowakischen (Teil II: Mit digraphen). In: *Sprache und Sprachen in Mitteleuropa*. GeSuS, Trnava (2005) (to appear)
10. Grzybek, P., Kelih, E., Altmann, G.: Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für slawische Philologie* **32** (2004) 25–54
11. Wimmer, G., Altmann, G.: Towards a Unified Derivation of Some Linguistic Laws. In Grzybek, P., ed.: *Contributions to the Science of Language: Word Length Studies and Related Issues*. Kluwer, Dordrecht (2005) (to appear)
12. Wimmer, G., Altmann, G.: Unified Derivation of Some Linguistic Laws. In Köhler, R., Altmann, G., Piotrowski, R.G., eds.: *Handbook of Quantitative Linguistics*. de Gruyter, Berlin (2005) (to appear)

DaskaL – A Web-based Application for Foreign Language Teaching

Kjetil Raa Hauge¹, Svetla Koeva², Emil Doychev³, and Georgi Cholakov³

¹ University of Oslo

² Bulgarian Academy of Sciences

³ Plovdiv University

Abstract. The DaskaL web application is a framework for the creation and interactive use of foreign language exercises for beginners and advanced students. It supports the construction of different types of tasks: grammar drills; element order drills, and free-style drills (essays, dialogues etc.). Several options for interaction (keyboard entry, menu choice, single or multiple correct answers) are provided. Inflectional dictionaries may be plugged in to speed up development of exercises.

1 Introduction

The DaskaL web application is a framework for the creation and interactive use of foreign-language learning exercises for beginners and advanced students (<http://daskal.net>).⁴ It is being developed with language data from Bulgarian, Serbian, Czech and Polish, but other languages can easily be added, as the system architecture is designed to be language-independent. DaskaL provides possibilities for educators to make several types of exercises and present them to students. The exercises may relate to different levels of language units – sound, morpheme, word, phrase, sentence, or text.

The application is based on three conceptual units labelled: task, exercise and test. A **task** is seen as the actual linguistic knowledge to be attained by an exercise. **Exercises** are the concrete realizations of the tasks and are the basic units of the system. A **test** is an ordered group of two or more exercises.

DaskaL offers several exercise patterns covering different types of tasks: **grammar drills**; **word order** (element order) **drills** (where scrambled units on different levels should be put into correct order); and **free-style** (semantic) **exercises** (essays, dialogues etc.). Different types of exercises may be combined into tests. The exercises can be designed to accept either type-in answers or a selection from a menu.

When several exercises are combined into a test, they may be assigned varying

⁴ The DaskaL application is developed by the Department for Computational Linguistics at the Institute for Bulgarian language at Bulgarian Academy of Sciences. The project is financed by the Program for Cooperation between the Faculties of Letters at the University of Oslo and Gothenburg University, and by the Program for Flexible Learning at the University of Oslo.

point weights. The weights may also be assigned by default and adjusted by the educator after real-life testing. The student may be given the opportunity to show and hide the correct answers. For essay-type exercises a model answer or teacher's hint may be shown and hidden again in a similar fashion.

DaskaL's advantage over similar systems is that the educator may either specify correct answers on an ad hoc basis (by typing or pasting) or may specify that the correct answer(s) should be retrieved automatically from a database according to given criteria. The database will include lexica of the above-mentioned languages (presently only Bulgarian and Czech are available) in which each word form is associated with its lemma and its grammatical features. Thus the educator may specify that a given slot in each question in an exercise should be filled with masculine singular animate noun, generated automatically from different lemmas. The results from these queries may be used for populating a drop-down menu, for providing the correction to a typed-in answer from the student, or for providing part of the stimulus.

2 Application Functionalities

DaskaL is a web-based application. It distinguishes three classes of users:

- Educator – with permission to create, edit and delete exercises and tests and specify access levels for them.
- Student – with access to a number of exercises and tests specified by his or her educator.
- Guest – with access to a number of exercises that are made available to any user.
- Administrator – who manages the users.

2.1 Functions for Educator Users

A user logged in as Educator will meet a window for selecting tests and a window for selecting exercises. Making a selection in either of these windows will start the Test Editor or the Exercise Editor respectively.

Developing and Editing Exercises Selecting exercises

Checkboxes and menus in the Exercise Browser allow the Educator to select a list of exercises according to a number of parameters:

- Exercise Type – with the following alternatives: All Types, Grammar Exercises, Element Order Exercises, Semantics Exercises, and Mixed Exercises.
- Exercise Level – with the values: All levels, Beginners, Intermediate, and Advanced.
- Status – with the following alternatives: All, Active, and Inactive (active exercises are exercises that are accessible to Student and Guest users as part of a test, while inactive exercises are accessible only to the Educator (and Administrator).

- Language – the choice relates to the language that is taught (not the language of the menus) with the following alternatives: All Languages, plus an option for every installed language.
- Name - option for search by name (full or in part) of a given exercise.

Exercise list

The result from the exercise search appears in the Exercise browser as a list with three columns:

- The column "Exercise" contains a short description of the exercise. For inactive exercises, the text is greyed out.
- In the "Status" column active exercises are indicated by a green icon, while the column is empty for inactive exercises.
- The third column "Operations:" provides two options: Edit exercise and Delete Exercise.

At top and bottom of the list are links named "New Exercise", which lead to the form for creating exercises.

Creating an exercise

New exercises are created by filling in the fields of the form of the Exercise Editor (Figure 1). The filter fields reflect the basics of the chosen methodology. They are: **Type** (with the options Grammar, Element Order, Semantic, and Mixed), **Level** (with the options Beginners, Intermediate, and Advanced), **Language** (the language under study), **Name** (max. 255 characters), **Description** (a brief description of the task of the exercise), **Example** (an example of the type of questions the exercise contains), **Contents** (the content of the exercise itself), **Points** (the number of points awarded for a correct answer), **Answers in Text** (this option indicates that the exercise contains text with gaps to be filled in by the student and that the Contents field consequently should contain special codes for specifying the positions of the gaps), **More Than One Valid Answer** (this option allows the educator to specify several correct answers), **Active** (active exercises are those that are included in some test, while inactive exercises are not assigned to any test). When the fields are filled in, clicking the button Add will enter the exercise into the database.

From the same window the link **Fields and Answers** opens a tab page with options for further specification of answer fields. The answer fields are shown in a table with the following columns providing different options for specification: **Field** an explicit link to the gap position code already assigned in the Contents field on the general descriptive page; **Type** the answer field can be specified either as a Type-in field (keyboard entry) or Choose from List; **Operations** – edit and delete operations for the given field.

If the field is of the type Choose from List, the full list of alternative answers from which the student should select the correct one should be entered (Figure 2).

The Educator may specify the order of the possible answers (applicable only for the type Choose from List) together with indication of whether they are

The screenshot shows a web form titled "Exercise Editor" with a sub-header "Add New Exercise". The form contains the following elements:

- Three dropdown menus: "Type" (set to "Grammar"), "Level" (set to "Beginners"), and "Language" (set to "Croatian").
- A text input field for "Name".
- A large text area for "Description".
- A large text area for "Example".
- A large text area for "Contents" with the instruction "enter each question on a separate line".
- A "Points" input field.
- Three checkboxes: "Answers in text", "More than one valid answer", and "Active".
- Two buttons at the bottom: "Add" and "Back to Browser".

Fig. 1. The Exercise Editor

correct or not. Lists of answer options may be made either manually or by database search.

Answer options through database search

If an inflectional dictionary for the given language is present in the database, answer options can be retrieved from it. The system presently contains Bulgarian [3] and Czech dictionaries. The form for dictionary access is shown in Figure 3.

In this form selection parameters for answer option retrieval are specified. The three tables: Part of Speech, Class, and Flexion, are hierarchically linked, so that a change of selection in a table on the left will cause the table on its right to be populated with new content. For instance, if a noun is selected from the Part of Speech list, the database relations will allow selection among different classes of nouns: common, proper, masculine, feminine, neuter, singularia tantum and pluralia tantum. The class specification in its turn interrelates with the grammatical features linked with the chosen class, so that for instance Bulgarian common feminine nouns will be specified as either singular or plural and either definite or indefinite. Selection of at least one option in the leftmost table is compulsory. Selecting options from the two other tables is optional and will narrow down the number of word forms returned. Multiple selections are allowed in the two leftmost tables.

Clicking the button Load Words will load the word forms corresponding to the selected criteria and will show them in the table at the bottom. If, for example, Bulgarian common feminine noun, singular, and definite are the selected features,

Order	Answer Contents	Correct?	Operations
0	на	Yes	
0	пък	No	
1	за	No	

[New Answer](#) [Back to Fields List](#) [Add From Dictionary](#)

Fig. 2. Fields and Answers

word forms like *rozata* (*the rose*); *zhenata* (*the woman*), etc. will appear. The Educator can select one or several word forms from the list and mark them as correct or incorrect candidates for the answers.

For any given field in the exercise, answer options may be added by retrieval from the database any number of times, with different criteria each time. The method may also be combined with manual editing of answer options. **Editing an exercise**

Editing an exercise works in the same way as creating an exercise, with the exceptions that fields are already filled in and the link Fields and Answers is active.

Developing and Editing Tests **Selecting tests** Possible search criteria for extracting tests from the database are Language, Status, and Name, corresponding to the same criteria for exercises.

Test list

The list of tests is a table similar to that used for exercises, with search criteria options and a list of test names (Figure 4). Links for "New Test" are at the top and bottom. **Creating a test**

Tests are created with the Test Editor. It contains fields for Language, Name, Description, and Active. The Educator fills in these fields and enter the test into the database. He can then continue editing it in the Test Editor, and the link Exercises will be active. It opens a page from which exercises are added to the test. This page shows the exercises in a view similar to that of the Exercise Editor, but with some added features (Figure 5). This page allows viewing all exercises (which will be most convenient when creating a new test) or only those that are included (for reviewing the contents of an already existing test) and provides buttons for excluding, including or updating an exercise. As it is assumed that educators would not wish to make multilingual exercises, there is no option for Language here, that choice having already been made. The List of exercises has the following columns:

The screenshot shows the 'Exercise Editor' window with the 'Fields and Answers' tab selected. It features three main sections for selecting word properties:

- Select Part of Speech:** A list box containing options like 'Съществително', 'Съюз', 'Частица', 'Наречие', 'Прилагателно', 'Глагол', 'Числително', 'Местоимение', 'Предлог', and 'Междуметие'. A 'Load >>' button is to its right.
- Select Class:** A list box containing morphological classes such as 'см, нарицателно, мъжки род', 'сж, нарицателно, женски род', 'сс, нарицателно, среден род', 'сж, нарицателно, мъжки род, женско склонение', 'сс, нарицателно, мъжки род, средно склонение', 'смг, мъжки род, сингулария тантум', 'ссг, среден род, сингулария тантум', 'сжг, женски род, сингулария тантум', 'сп, нарицателно, плуралия тантум', and 'смжб, собствено, мъжки род'. A 'Clear <<' button is to its right.
- Select Flexion:** An empty list box with a 'Load >>' button to its right.

Below these sections is a 'Load Words' button. At the bottom, there is a table with the following structure:

Select	Word (found 0)	Correct?
<input type="checkbox"/>		<input checked="" type="checkbox"/>

A 'Back to Browser' button is located at the bottom left of the interface.

Fig. 3. Answer options added from the Bulgarian morphological dictionary

- Number – allowing the Educator to change the ordering of exercises within the test:
- Included – a green icon indicates an exercise that is active and included in the test. Exercises that are not included are greyed out. Exercises can be marked as inactive and will not be shown to student users even if included in the test.
- Points – the number of points awarded for a correct answer. If the exercise has been assigned a default value by the educator, this value will be shown here, but it may also be altered if the purpose of the test calls for it.
- Operations – with buttons for the functions Include (includes exercises in the test) and Exclude (removes the exercise from the test).
- Update – this button is used if the values for numbering and points have been changed.

Editing a test

Editing a test works in the same way as creating a test, except that fields are filled in and the link Exercises is enabled.

2.2 Functions for Student Users

Users of the category Student may take any specified active test. Each exercise in the test will be shown on a separate page (Figure 6). These web pages are generated on the fly from the database according to the parameters set by the Educator. For instance, exercises may be generated with a selection of possible answers, or with a field for keyboard entry of the student's answer. Navigation

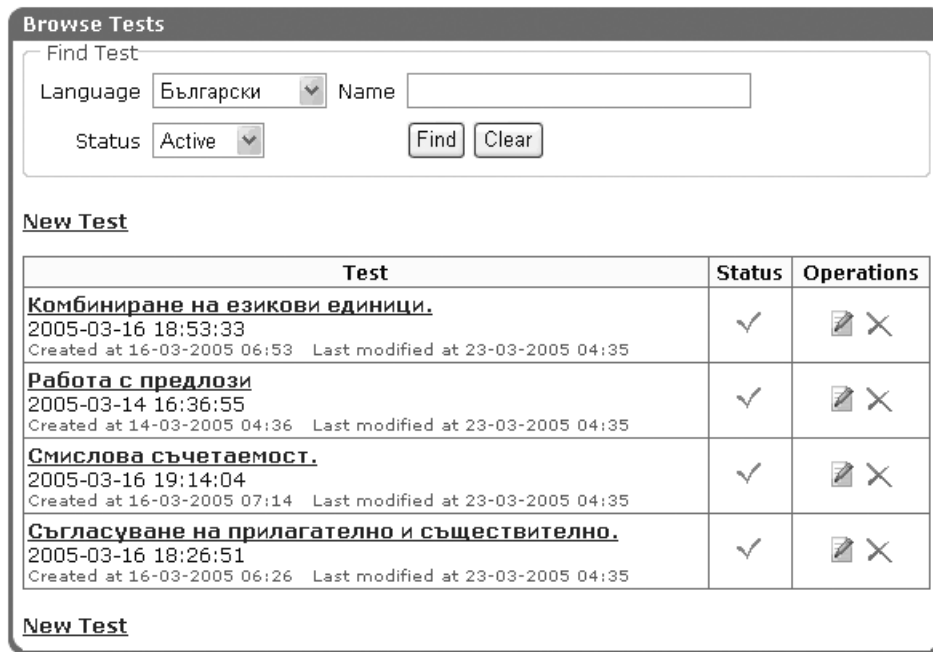


Fig. 4. The test selection window

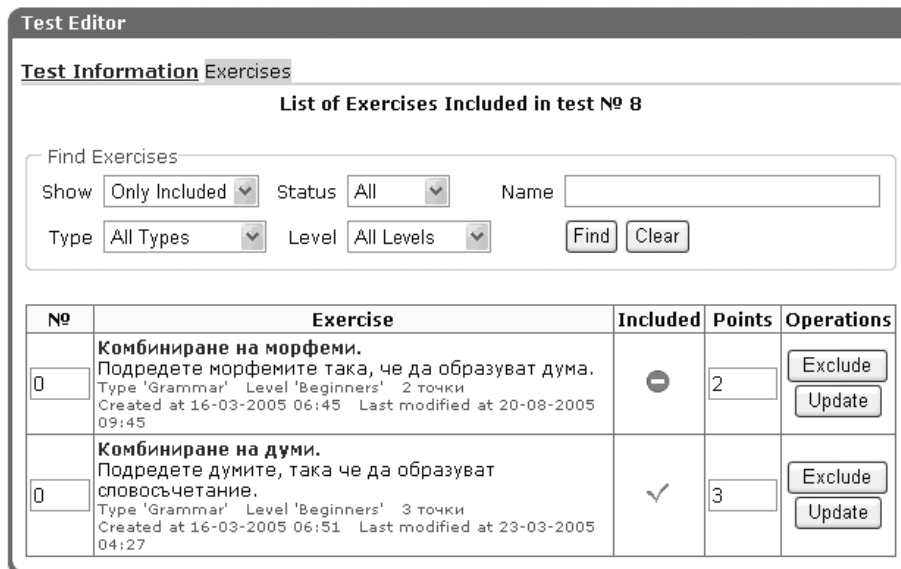


Fig. 5. The Test Editor

Test Browser

Exercise: Удвояване на предлози

Поставете правилната форма на съответния предлог

Example: Иван играе с Мария.
Иван играе със Светла.

Иван играе Мария.

Иван играе Светла.

Иван пристигна Пловдив.

Иван пристигна Видин.

<< Previous Next >> Cancel

Fig. 6. Tests as seen by a Student user

buttons take the student through the exercises of the tests in the stipulated order. At the end of the test the student's score is calculated and shown in a detailed summary (Figure 7).

Test Browser

Summary of Results for Test "Работа с предлози"

Exercise	Correct?	Points
Удвояване на предлози	✘	-
Избор на предлог	✔	2
Total Points:		2

Back to Browser

Fig. 7. Summary of the results

3 DaskaL Application

There are two logical databases: The DaskaL application works entirely through a dynamic web interface, for students as well as for educators and administrators.

From a technological point of view it is organized as a portal. A portal in this context is a web-based application that provides personalisation, single sign-on, and content aggregation from different sources and hosts the presentation layer of information systems [4]. It is organized through the portal framework Jetspeed [1]. The user interface consists of portlets giving various degrees of access according to the status of the user (guest; registered student; content creator: teacher, administrator). A portlet is defined as a Web component, usually managed by a container that processes requests and generates dynamic content. Portals use portlets as pluggable user interface components to provide a presentation layer to information systems [4]. The architecture is shown in Figure 8. Users may customize parts of the user interface. The application uses two separate databases, currently served by MySQL:

- for administrating the portal: users, status, permissions, customization, etc.
- for administrating the learning framework: exercises, dictionaries, tests, etc.

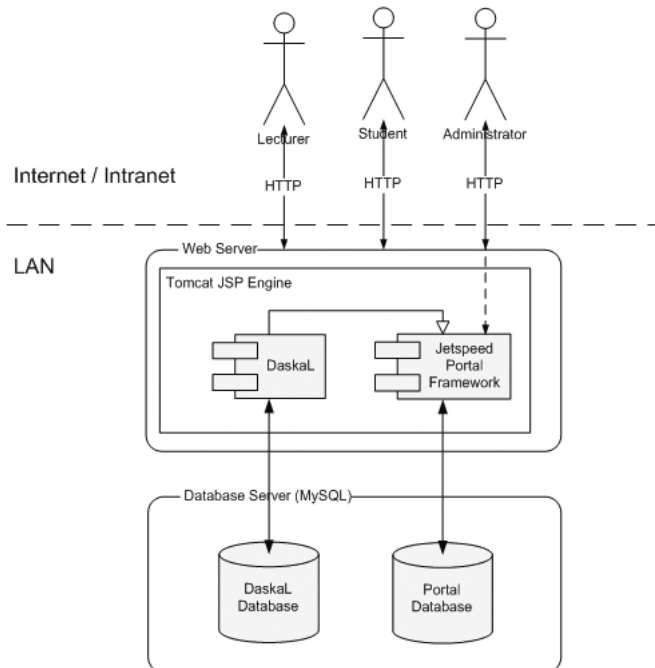


Fig. 8. The DaskaL architecture

The architecture is multilayered, with the application layer divided into several sublayers implementing different functions. The system allows three kinds of users: – Educator, Student (which for this purpose includes Guest users), and

Administrator. The administrator functions are implemented exclusively by Jet-speed, with new functions created within its framework for serving the students and lecturers. DaskaL is a pure Java application where the JSP technology is used for portlets implementation. The runtime environment is supplied by the JSP Engine – Apache Tomcat 5 [2]. For database server is used MySQL. There are two logical databases:

- the Portal Database – with the structures and data necessary for the operation of the portal – user accounts, ownership and permissions, user customizations, etc.
- the DaskaL Database – with the structures needed for DaskaL’s functionality – exercises, tests, answers, etc.

Figure 9 shows the model of the DaskaL Database. The DaskaL Database is logically independent of the Portal database. Connection between the two is served by the table of user accounts. The DaskaL Database contains two groups of tables:

- For exercises and tests, with the description of the exercises, possible correct and false answers, description of the tests and students’ test results.
- Dictionary tables supporting Unicode and thus allowing for dictionaries in almost any language.

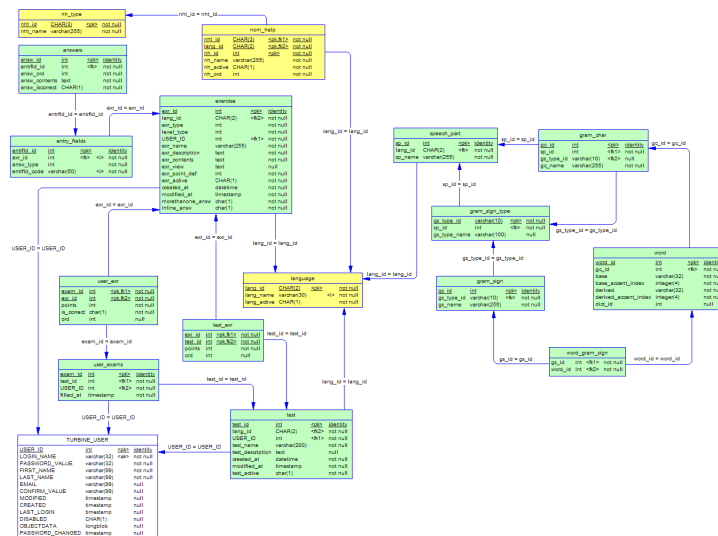


Fig. 9. The DaskaL Database

4 Conclusions and Future Directions

DaskaL provides a convenient tool for creating, editing, storing, retrieving and presenting foreign-language exercises on the World Wide Web, in almost any language supported by Unicode. The use of morphological dictionaries further speeds up development of exercises. Future development of the application includes XML import/export options for exercises, tests, and dictionaries. This will be important for exchange of exercises and tests between different organizations using DaskaL, for import into other computer-aided learning systems, for external (i.e. outside the DaskaL framework) archiving of exercises and tests in a device-independent format, and for hand-tooling exercises.

References

1. Apache Portals – Jetspeed, <http://portals.apache.org/jetspeed-1/>
2. Apache Tomcat, <http://jakarta.apache.org/tomcat>
3. Koeva S. Bulgarian Grammatical dictionary. Organization of the Language Data, Bulgarski ezik, 1998, vol. 6: 49-58.
4. Question mark Glossary, <http://www.questionmark.com/uk/glossary.htm>

Aspects of an XML-Based Phraseology Database Application

Denis Helic¹ and Peter Ďurčo²

¹ University of Technology Graz
Institute for Information Systems and Computer Media
`dhelic@iicm.edu`

² University of St. Cyril and Methodius Trnava
Department of German Language and Literature
`durco@vronk.net`

Abstract. In this paper we discuss technical aspects of a phraseology database application that is being developed in the scope of the Ephras project. Thereby we give a special attention to the data modeling perspective of such an application. We argue that the phraseology data is simply a particular kind of semi-structured data. Therefore, this data should be represented and managed by technologies that are specifically targeted at management of semi-structured data. Currently, the most prominent of such technologies is eXtensible Markup Language technology. Thus, in the remainder of the paper we discuss different implications of this technology on the application architecture and its implementation.

1 Introduction

The Ephras project is a project funded by the European Commission under Socrates/Lingua2 programme. The goal of the project is to develop a computer supported phraseology learning material for four European languages - German, Slovak, Slovenian and Hungarian language. The project aims to eliminate the lack of such phraseology learning material, as well as to meet the demands for multilingual learning material in the enlarged European Union. The Ephras learning material will be composed of a searchable database of 4x1000 phraseology data items in four languages (i.e., 1000 data items in each of the languages) accompanied with 150 interactive tests to selected phrases in four languages.

In this paper we concentrate on the first component of the Ephras learning material - the Ephras phraseology database application - by discussing its requirements and features, as well as a number of important technical issues related to that component. The most important requirements and features of this application can be summarized as following.

Firstly, the source language is German with 1000 phraseology data items, where each of these data items is a single phrase in German with one or more meanings. Additionally, each German phrase is involved in a so-called equivalence relation with data items from other three languages (target languages). The

equivalence relation expresses rather complex phraseology relationships between different data items. It can represent one of the following:

- A one-to-one relation between a single-meaning German phrase and a single-meaning phrase from any of the target languages.
- A one-to-many relation between a single-meaning German phrase and a number of different single-meaning phrases from any of the target languages.
- A one-to-one relation between a multiple-meaning German phrase and a multiple-meaning phrase from any of the target languages.
- A one-to-many relation between a multiple-meaning German phrase and a number of different single-meaning or multiple-meaning phrases from any of the target languages.
- Sometimes there is no direct phraseology equivalent for a German phrase in the target languages. The equivalents in that case are non-phraseology data items, i.e., a single word or a free collocation. Thus, the equivalence relation in such a case is a one-to-one relation between the German phrase and its corresponding free collocation.

Thus, the equivalence relation is 2-dimensional, where in the first dimension we have a one-to-one or a one-to-many relation between a German data item and the corresponding data items in the target languages. Orthogonal to that relation there is a relation between meanings of different data items, which can be either single-meaning or multiple-meaning data items, i.e., this relation is a typical many-to-many relation.

Further, in addition to the primary direction of the equivalence relation (i.e., the direction from German to other three languages) the secondary direction of this relation can be established as well (see Fig. 1). In some special cases (e.g., when only one-to-one relations are present) it is possible to infer the equivalence between data items from the target languages. For example, starting from a Hungarian data item it is possible to find its equivalent in Slovak by implicitly using the existing one-to-one relations between those two data items and their German counterpart.

Another important feature of the Ephras phraseology database application is a so-called description model. The description model for phrases in all four languages has been developed according to the latest phraseology principles and includes the following: basic form (i.e., the content of a phrase), meanings, style, grammar, collocation, pragmatics, examples, variants, keywords, synonyms, categories and multilingual comment.

From the user point of view the application has the following properties. The user can search within the database using any of four languages as the starting language. The search results are presented in a list form where the user can click on a particular search result and obtain the full description of the data item. The links to the related data items (e.g., equivalents) are included in the data item description.

The rest of the paper is organized as follows. The next section discusses the aspects of representing the phraseology data from the data modeling point of view. The subsequent section describes in details the application architecture

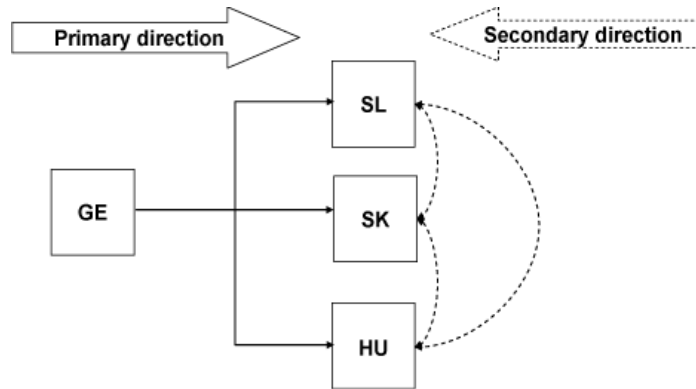


Fig. 1. Directions of Equivalence Relation

and the influence of the chosen data representation on that architecture. Finally, we give a number of conclusions and pointers for the future work.

2 Data Representation

To efficiently represent the phraseology data it was necessary to look closely on some of its features. Here is a list of some specific properties of the phraseology data:

- Data is structured only to a certain extent, i.e., there are data fields that have a different structure for different data items. For example, the meanings field contains typically textual content that possesses a varying internal structure - a single meaning or a list of meanings. Another example includes the multilingual comment field, where the content can be decorated, thus including text in bold or in italics.
- Whenever a data field has an internal structure as it is the case with the meanings or multilingual comment field, the ordering of elements within this internal structure is important and embodies semantic significance. For example, if a meanings field contains a list of meanings then the ordering of these meanings within the list possesses a certain denotation and needs to be maintained.
- Data items are interrelated by means of typed relations such as the equivalence relation discussed above. The equivalence relation is an ordered relation with varying arity and dimensions, e.g., the relation can be a one-to-one, an arbitrary one-to-many relation or even a many-to-many relation between meanings of data items. Additionally, data items can be involved in a so-called variance relation (e.g., a phrase is a variant of another phrase in the same language).

From the data modeling point of view, data with such properties is referred to as semi-structured data. Typically, semi-structured data is irregularly, partially or implicitly structured. Further, for such data there is no a-priori or rigid database schema but only a so-called a-posteriori data guide can be identified [1, 2]. Obviously, the phraseology data in question can be classified as semi-structured data.

Generally, semi-structured data is modeled as a labeled graph [1]. The nodes represent data items, have unique identifiers and can be either atomic or composite. Composite data items are related with other data items by means of labeled edges, where labels represent the relation types. A simple model representing a couple of phraseology data items can be seen in Fig. 2.

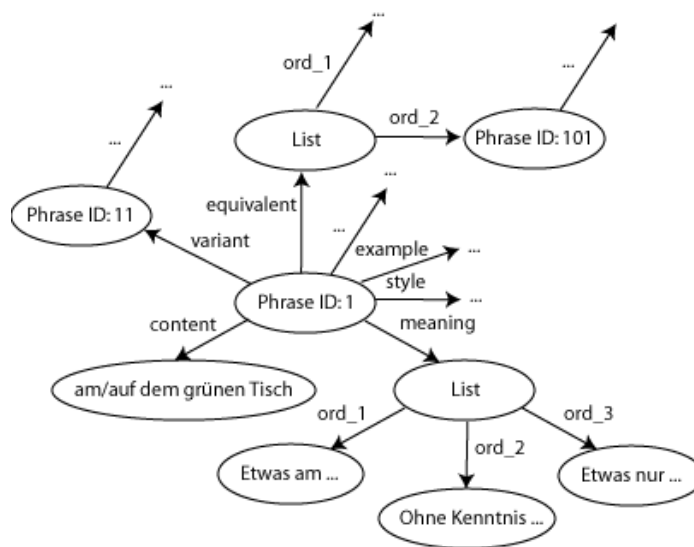


Fig. 2. Model of Semi-structured Phraseology Data

Recently, through the emergence of the Web and the related mark-up technologies, such as eXtensible Markup Language (XML), the latter evolved to a de-facto standard for managing semi-structured data [3, 4]. An XML document is a hierarchy of elements with ordered sub-elements. Each element has a name (also referred to as label or tag). The basic XML model is a labeled ordered tree where labels represent node names. Edges are always directed (to preserve the tree order) and do not have labels. Additionally, XML supports a referencing mechanism between nodes, which basically facilitates modeling of arbitrary graphs. In this way semi-structured data might be represented by means of XML documents. An excerpt from an XML document encapsulating the above depicted phraseology data items is shown in listing 1.1.

```

<phrases>
  <phrase id="1">
    <content>am/auf dem grünen Tisch</content>
    <meanings>
      <meaning>Etwas am ... </meaning>
      <meaning>Ohne Kenntnis ... </meaning>
      <meaning>Etwas nur ... </meaning>
    </meanings>
    <style> ... </style>
    <examples>
      <example> ... </example>
    </examples>
    ...
  </phrase>
  <phrase id="11" variant="1">
    ...
  </phrase>
  <phrase id="101" equivalent="1">
    ...
  </phrase>
  ...
</phrases>

```

Listing 1.1. XML Document Encapsulating Phraseology Data Items

3 Implications of XML on Application Architecture

The architecture of the Ephras phraseology database application closely follows the well-known three-tiered architecture of user-oriented database applications (see Fig. 3). The three tiers are:

- User interface module that manages the user interaction and presents the data items to the user.
- Application logic module which implements the core application functionality by representing the data items and the operations that the user can perform on these data items (e.g., get an equivalent, get a variation of a phrase, etc.). This functionality is supported in a standard object-oriented manner, i.e., as a collection of interacting objects. Additionally, this module provides a bridge to the underlying data management module.
- Data management module which abstracts the access to an external XML-based database system by means of a programmatic interface. In addition, the external XML-based database system manages the XML representation of the phraseology data items.

Using XML for data management in the Ephras phraseology database application has a number of important aspects. First of all, the application deals

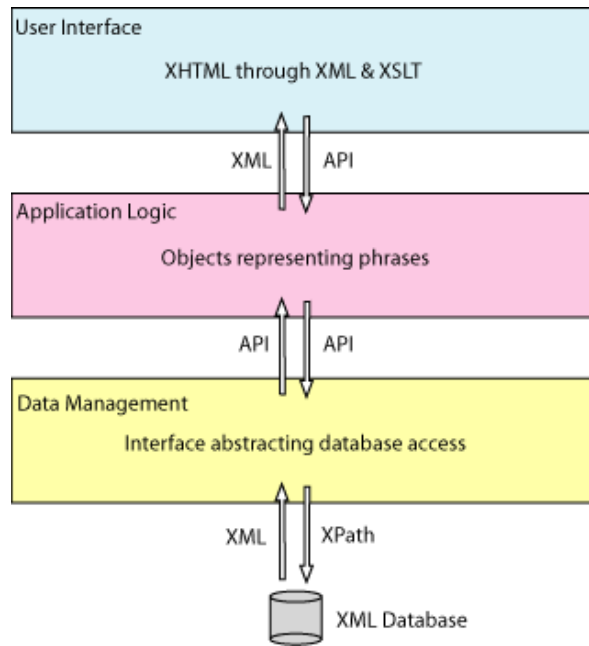


Fig. 3. Architecture of Phraseology Database

with the phraseology data from four different languages, namely German, Slovak, Slovenian and Hungarian. These four languages contain different characters, which are encoded using a particular character encoding schema. For example, German characters are encoded using ISO 8859-1 character set (Latin-1 or West European encoding). On the other hand, the characters from the remaining languages are encoded using ISO 8859-2 character set (Latin-2 or Central and East European encoding). Thus, the only possibility to combine characters from those four languages in a single XML document is to encode them using ISO 10646 Unicode character set. Technically, this does not constitute a problem, since XML documents might be encoded using Unicode character set. Additionally, all XML documents support UTF-8 and UTF-16 Unicode encodings, which define how to encode Unicode characters in a space-saving manner. In the Ephras phraseology database application we decided to use UTF-8 encoding for that purpose.

The second important aspect of using XML in the Ephras phraseology database application is related to communication between the data management module and the underlying XML-based database system. Obviously, the language for querying the database must be an XML query language. In this application the chosen language is XPath query language. XPath is a simple query language that works directly with the underlying tree-based model of an XML document supporting queries that retrieve subtrees of the whole XML tree. Thereby, different

matching criteria can be applied, such as element-based matching criteria (e.g., give me all phrase elements), attribute-based matching criteria (e.g., give me all phrase elements that have a certain attribute with a certain value) or content-based matching criteria (e.g., give me all phrase elements with a certain word in the content). For example, the first query in Listing 1.2 retrieves all phrases from the database and the second query retrieves all phrases that contain word 'Tisch'.

```
/phrases/phrase
/phrases/phrase[content[contains(., 'Tisch')]]
```

Listing 1.2. XPath Queries for Retrieving Data Items

Finally, the third aspect of XML in the Ephras phraseology database application is related to the user interface module. Originally, XML is specified as a meta-document format that can be used to define families of document formats. Definition of presentation instructions for such document families is not a part of XML specification and is defined elsewhere - namely by a number of so-called style-sheet specifications. Basically, a style-sheet is a separate document which defines how a certain XML document should be presented. Currently, Cascading Style Sheets (CSS) and eXtensible Stylesheet Language - Transformations (XSLT) are typically used for that purpose. CSS is used to specify formatting instructions for XML documents whereas XSLT provides possibilities to transform an XML document to another XML document for which presentation instructions already exist. The best known example is transformation of arbitrary XML documents into HTML or XHTML documents, which can be subsequently presented using a standard Web browser. In this application we have chosen the latter approach and thus transform XML documents into XHTML documents and present them in a Web browser to the user.

4 Conclusion and Future Work

In this paper we presented the Ephras phraseology database application, which is a database application for management of phraseology data in four different European languages. For the purpose of implementing this application it was important to examine the defining features of such phraseology data. The most important technical result of this examination is the conclusion that phraseology data should be classified as semi-structured data. Since XML is a de-facto standard for management of semi-structured data today, applying XML database technology for implementing the application was an obvious choice. The subsequent discussion of a number of aspects of XML, such as querying facilities or presentation of the data provides an insight in a number of implementation issues.

Currently, the application is still in the development phase. The XML database, the data management module as well as the application logic module are already implemented. The user interface module is still under development. The first version of a complete system will be available in the beginning of 2006.

References

1. Abiteboul, S.: Querying Semi-Structured Data. In: Proceedings of the 6th International Conference on Database Theory - ICDT '97. (1997) 1–18
2. Buneman, P.: Semistructured Data. In: PODS '97: Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, New York, NY, USA, ACM Press (1997) 117–121
3. Goldman, R., McHugh, J., Widom, J.: From Semistructured Data to XML: Migrating the Lore Data Model and Query Language. In: Proceedings of the 2nd International Workshop on the Web and Databases (WebDB '99). (1999)
4. Vianu, V.: A Web Odyssey: From Codd to XML. In: PODS '01: Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, New York, NY, USA, ACM Press (2001) 1–15

VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech

Dana Hlaváčková and Aleš Horák

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
{hlavack,hales}@fi.muni.cz

Abstract. The paper presents new lexicon of verb valencies for the Czech language named VerbaLex. VerbaLex is based on three valuable language resources for Czech, three independent electronic dictionaries of verb valency frames.

The first resource, Czech WordNet valency frames dictionary, was created during the Balkanet project and contains semantic roles and links to the Czech WordNet semantic network. The other resource, VALLEX 1.0, is a lexicon based on the formalism of the Functional Generative Description (FGD) and was developed during the Prague Dependency Treebank (PDT) project. The third source of information for VerbaLex is the syntactic lexicon of verb valencies denoted as BRIEF, which originated at FI MU Brno in 1996.

The resulting lexicon, VerbaLex, comprehends all the information found in these resources plus additional relevant information such as verb aspect, verb synonymity, types of use and semantic verb classes based on the VerbNet project.

1 Introduction

The beginnings of building the verb valency frame dictionary at the Faculty of Informatics at Masaryk University (FI MU) dates back to 1997 [1]. Since then, the dictionary, denoted as Brief, has undergone a long development and has been used in various tools from semantic classification to syntactic analysis of Czech sentence [2]. Currently, the dictionary plays a key role within an experimental high-coverage syntactic analysis using the data from the Czech WordNet. The data in this dictionary can be entered in several mutually convertible formats:

brief:

jíst (to eat) <v>hTc4,hTc4-hTc6r{na}, hTc4-hTc7

verbose:

display:

jíst

= co

= co & na čem

= co & čím

jíst něco

jíst něco na něčem

jíst něco něčím

Lemma variants:

Princeton WordNet – plan:2
 Definition: make plans for something
 VALLEX 1.0: vymyslet₂ / vymyslit₂
 VerbaLex: vymyslet:1, vymyslit:1, naplánovat:3

Word entries:

Princeton WordNet – arrive:1, get:5, come:2
 Definition: reach a destination; arrive by movement or progress
 VALLEX 1.0: dojít₁
 VerbaLex: dojít:1, dorazit:1, dostat se:1, přicestovat:1, přijet:1, přijít:1

Fig. 1. Examples of verb frame entry heads for verbs with lemma variants and for synonymic verbs.

The Brief dictionary contains about 15 000 verbs with 50 000 verb valency frames, thus making it an invaluable language resource with high coverage. However, the different verb senses are not distinguished here.

Another advance in the Czech verb valency processing came during the work on the Czech WordNet within the Balkanet project [3]. The Czech WordNet has been supplemented with a new language resource, Czech WordNet valency frames dictionary. The new acquisition of this dictionary were the semantic roles and links to the Czech WordNet semantic network.

During the work on enhancing the list and adding new entries into it, we have come to the need of comparing the quality and features of the list with the parallelly created valency lexicon of Czech verbs denoted as VALLEX 1.0 [4]. In cooperation with the VALLEX team, valency frames from Czech WordNet were transformed to an augmented VALLEX format, which was named VerbaLex.

The FI MU VerbaLex dictionary is being actively developed, checked and supplemented with new data. Currently, VerbaLex contains 3 469 verb literals which, when gathered in synonymic groups, share 1 807 verb frames. Nowadays, several linguists are working on a bulk of 15 000 more verbs being added to VerbaLex.

2 Linguistic Requirements for the VerbaLex Format

In this section, we present the substantiation of the main differences between VerbaLex and VALLEX 1.0 valency frames notation.

The lexical units in WordNet are organized into synsets (sets of synonyms) arranged in the hierarchy of word meanings (hyper-hyponymic relations). VerbaLex differs from VALLEX 1.0 in augmentation of the original format, detailed differentiation of valency frames and above all semantic roles (deep cases). For that reason, the headwords in VerbaLex are formed with lemmata in a synonymic relation (synset subsets) followed by their sense numbers (standard Princeton

WordNet notation). The standard definition of synonymy says that two synonymic words can be always substituted in the context. However, the synonymy in synsets is understood like very close sense affinity of given words, the substitution rule cannot be applied in all cases here. In VALLEX 1.0, a headword is one lemma, possibly two or more lemmata in case of lemma variants.¹ Lemma variants in VerbaLex are considered as independent lemmata and they are distinguished by their WordNet sense numbers. An example of two verb frame entries in VALLEX 1.0 and VerbaLex is displayed in the Figure 1.

In VerbaLex, each word entry includes an information about the verb aspect (perfective – **pf.**, imperfective – **impf.** or both aspects – **biasp.**). VerbaLex valency frames are enriched with aspect differentiations for examples containing the verb used with the given valency frame. This is important in case of synonymic lemmata with different aspect:

Princeton WordNet – wade:1
 Definition: walk (through relatively shallow water)
 VerbaLex: brodit se:2 **impf.**, přebrodit se:1 **pf.**
 frame: AG <person:1>_{kdo1}^{obl} VERB SUBS <substance:1>_{čím7}^{obl}
 example: přebrodit se blátem **pf.** / he wade through mud
 example: brodit se pískem **impf.** / he wade through sand

The constituent elements of frame entries are enriched with pronominal terms (e.g. *kdo* – who, *co* – what) and the morphological case number. This notation allows to differentiate an animate or inanimate agent position:

Princeton WordNet – bump:1, knock:3
 Definition: knock against with force or violence
 VerbaLex: narazit:1 **pf.** / narážet:1 **impf.**
 frame: AG <person:1>_{kdo1}^{obl} VERB OBJ <object:1>_{do}^{obl} čeho2,na co4
 PART <body part:1>_{čím7}^{obl}
 example: I bumped to the wall with my head
 frame: OBJ <vehicle:1>_{co1}^{obl} VERB OBJ <object:1>_{do}^{obl} čeho2,na co4
 example: the car bumped to the tree

2.1 Verb Usage and Verb Classes

VerbaLex captures additional information about types of verb use and semantic verb classes. Three types of verb use are displayed in the lexicon. The primary usage of a verb is marked with abbreviation *prim*, metaphorical use with *posun* and idiomatic and phraseological use with *idiom* (this follows the VALLEX 1.0 notation). The assigned semantic verb classes are adopted from the Martha Palmer's [5] VerbNet project. The verb classes list is based on Beth Levin's [6] classes with more fine-grained sets of verbs.

¹ the lemmata with small phoneme alternation that are interchangeable in any context without any change of the meaning – *bydlet/bydlit*, to live (where).

There are 395 classes in the current development version of VerbNet, which was provided by Martha Palmer's team. But this number seems to be too much for Czech verbs, therefore the list of verb classes will be adapted to the conditions of the Czech language:

Princeton WordNet – cry:2, weep:1
 Definition: shed tears because of sadness, rage, or pain
 VerbaLex: brečet:1, plakat:1, ronit:1
 class: nonverbal_expression-40.2

Princeton WordNet – take care:2, mind:3
 Definition: be in charge of or deal with
 VerbaLex: dbát:2, starat se:2, pečovat:3
 class: care-86

Princeton WordNet – be:11, live:5
 Definition: have life, be alive
 VerbaLex: žít:1, být:2, existovat:3
 class: exist-47

3 Semantic Roles

VerbaLex has introduced a different concept of semantic roles (i.e. functors in VALLEX 1.0) as compared to VALLEX 1.0. Currently, the list of semantic roles and the way of their notation establish one of the main differences between VALLEX 1.0 and VerbaLex valency frames (see also [7]). The functors used in VALLEX 1.0 valency frames seem to be too general and they do not allow distinguishing different senses of verbs. We suppose that a more specific subcategorization of the semantic role tags is necessary, therefore an inventory of two level semantic role labels was created.

The first level contains the main semantic roles proposed on the 1stOrderEntity and 2ndOrderEntity basis from EuroWordNet Top Ontology [8]. On the second level, we use specific literals (lexical units) from the set of Princeton WordNet Base Concepts with relevant sense numbers. We can thus specify groups of words (hyponyms of these literals) replenishable to valency frames. This concept allows us to specify valency frames notation with large degree of sense differentiability.

For example the literal `writing implement:1` is a hypernym for any implement that is used to write.

Princeton WordNet – draw:6
 Definition: represent by making a drawing of, as with a pencil, chalk, etc. on a surface
 VerbaLex: kreslit:1, malovat:1
 frame: AG <person:1>_{kdo1}^{obl} VERB ART<creation:2>_{co4}^{obl}
 INS<writing implement:1>_{cim7}^{obl}
 example: my sister draws a picture with coloured pencils, the famous artist was drawing his painting only with charcoal

The left-side valency position is most frequently occupied by the semantic role AG, an agent. The agent position in a valency frame is understood as a very general semantic role (functor ACT) in VALLEX 1.0. This label does not allow to distinguish various types of action cause. Two level semantic role labels in VerbaLex are able to define cause of action quite precisely. The main semantic role AG is completed by an adequate literal depending on the verb sense and valency frame. Thus, we can identify whether the agent is a person AG(person:1), an animal AG(animal:1), a group of people AG(group:1), an institution AG(institution:1) or a machine AG(machine:1). For some verbs with very specific sense, hyponyms of these literals are used. For example:

Princeton WordNet – sugar:1, saccharify:1
 Definition: sweeten with sugar
 VerbaLex: sladit:4, osladit:1, pocukrovat:1
 frame: AG <person:1>_{kdo1}^{obl} VERB SUBS <food:1>_{co4}^{obl}
 SUBS <sugar:1>_{čím7}^{obl}
 example: sugar your tea with brown sugar

In VALLEX 1.0, each valency frame starts always with functor ACT. In our opinion, it is useful to differentiate the sense of the left-side valency position (subject position) in more detail. According to our definition of agent AG (sb or sth doing sth actively) this position may be also occupied by other semantic roles. The subject position can contain objects OBJ, substances SUBS or a semantic role denoting abstract concepts – human activity ACT, knowledge KNOW, event EVEN, information INFO, state STATE. For example:

Princeton WordNet – follow:6, come after:1
 Definition: come after in time, as a result
 VerbaLex: přijít:25 / přicházet:25, následovat:4
 frame: EVEN <event:1>_{co1}^{obl} VERB EVEN <event:1>_{po čem6}^{obl}
 example: heavy rain followed flood

Princeton WordNet – fall:3
 Definition: pass suddenly and passively into a state of body or mind
 VerbaLex: zachvátit:2, zmocnit se:2
 frame: STATE <state:4>_{co1}^{obl} VERB PAT <person:1>_{koho4}^{obl}
 example: he fall into a depression

Quite a large number of semantic roles inspired by EuroWordNet Top Ontology roughly correspond with the PAT functor in VALLEX 1.0. The PAT label covers quite different senses, which can be very well identified.

In our inventory, PAT is defined as: the semantic role of an entity that is not the agent but is directly involved in or affected by the happening denoted by the verb in the clause (definition of literal patient:2 from Princeton WordNet).

Princeton WordNet – experience:1, undergo:2, see:21, go through:1
 Definition: go or live through

Table 1. List of semantic roles from VerbaLex that are used in examples.

AG	the semantic role of the animate entity that instigates or causes the happening denoted by the verb in the clause, we extended this definition for inanimate entity that does sth actively (e.g. machine)
ART	a man-made object taken as a whole
SUBS	that which has mass and occupies space
PART	a portion of a natural object, something determined in relation to something that includes it, something less than the whole of a human artifact
INS	a device that requires skill for proper use
OBJ	a tangible and visible entity; an entity that can cast a shadow
EVEN	something that happens at a given place and time
STATE	the way something is with respect to its main attributes

VALLEX 1.0: absolvovat₂
 frame: ACT₁^{obl} PAT₄^{obl}
 VerbaLex: absolvovat:2, prožít:1 / prožívat:1 /
 frame: AG <person:1>_{obl} VERB EVEN <experience:3>_{obl}
 example: he underwent difficult surgery

Some second level literals cannot be adopted from Princeton WordNet Base Concepts – especially specification of roles considered as “classic” deep cases. These literals (e.g. **agent:6**, **patient:2**, **donor:1**, **addressee:1** or **beneficiary:1**) do not have any hyponyms in Princeton WordNet and cannot be substituted by any word.

For such cases, the literal **person:1** is used (or another suitable literal with large number of hyponyms, e.g. **AG(person:1)**, **PAT(animal:1)**). This “classic” semantic roles are consistent with some functors in VALLEX 1.0 (**ACT**, **PAT**, **ADDR**, **BEN** etc.). A list of VerbaLex semantic roles that are used in the presented examples is displayed in the Table 1.

3.1 Special Semantic Roles

VerbaLex describes not only the valency and semantic frames, it also includes other relevant information about Czech verbs, such as the verb position. In a free-word order language like Czech the position of the verb within the verb frame is usually not strictly specified.

VerbaLex uses a special semantic role, **VERB**, which marks the (usual) position of the verb in its verb frame. Such default verb position is not needed only for analysis of verb valencies, it can be also directly used in the process of generation of Czech sentences, e.g. as an output of a question-answering machine.

The left side of the verb position is traditionally occupied by the sentence subject, which is also the case marked in most of the verb frames in VerbaLex. However, there are some cases, where the verb frame has to obey different rules

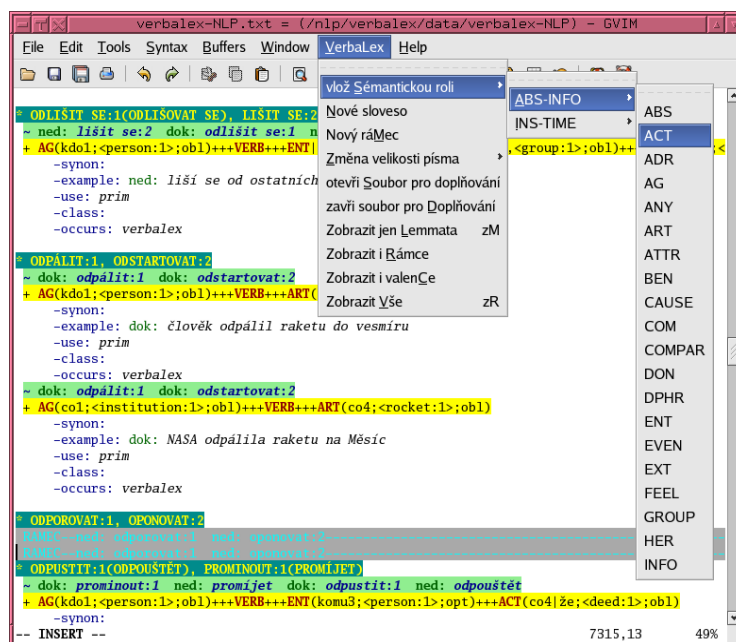


Fig. 2. The tool for editing verb valency frames dictionary in the VerbaLex format.

– e.g. sentence *Dalo se do deště* (It started to rain) cannot contain any subject. For the notation of such cases, VerbaLex uses another special semantic role **ISUB**, an inexplicit subject.

4 Implementation of Editing and Exporting Tools

For the sake of editing the newly adopted verb valency frame format VerbaLex, we have implemented a new set of editing and exporting tools.

The main interactive tool for user editing of the valency dictionary, named `verbalex.sh`, is based on a highly configurable multi-platform editor VIM (see the Figure 2). Such approach enables a linguistic expert to easily enter computer-parseable data in a fixed plain text format and still, thanks to the flexible color syntax highlighting, he or she has a full visual control of possible errors in the format.

The editing itself is not fixed to one platform, users can run the same environment under any of the current popular computer operating systems (VIM editor runs on nearly any platform).

The authoring tool `verbalex.sh` currently offers these functions to the editing user:

- free editing of the dictionary entries

```

<headword_lemmata>
  <lemma ord='1' sense='2' aspect='pf'>chopit</lemma>
  <lemma ord='2' sense='2' aspect='pf'
    aspectual_counterpart_lemma='uchopovat'>uchopit</lemma>
  <lemma ord='3' sense='2' aspect='impf'
    aspectual_counterpart_lemma='uchopit'>uchopovat</lemma>
  <lemma ord='4' sense='3' aspect='pf'
    aspectual_counterpart_lemma='brát'>vzít</lemma>
  <lemma ord='5' sense='3' aspect='impf'
    aspectual_counterpart_lemma='vzít'>brát</lemma>
  <lemma ord='6' sense='4' aspect='pf'
    aspectual_counterpart_lemma='chápat se'>chopit se</lemma>
  <lemma ord='7' sense='4' aspect='impf'
    aspectual_counterpart_lemma='chopit se'>chápat se</lemma>
</headword_lemmata>

```

Fig. 3. An example of XML structure of aspectual counterpart tuples within one dictionary entry.

- regular expression searching in the dictionary
- template-based adding of a new verb entry or a new verb frame to the current entry
- menu-based adding of new semantic role to the current frame
- multilevel folding – hiding/unhiding of valency attributes, valencies or full valency frames
- visual marking of the current frame for further inquiry
- interactive merging of definitions from two parallel sources

Moreover, the interpreted approach of the tool makes adding of new features to the editing system easy to implement.

The plain text format edited by a human expert is in further processing transformed into an XML standard format which enables conversions into different formats used for visual checking, searching and presentation of the valency dictionary.

The XML schema used in VALLEX 1.0 had to be changed to suit the augmentation of the format in VerbaLex. The changes include

- adding `class` attribute to frame `slot` tag to cover wordnet basic concept literals
- including the wordnet word sense in the lemma tags
- shifting the verb aspect to `headword_lemma`, which now enumerates all the aspectual counterpart tuples. An example of such XML substructure can be found in the Figure 3.

The resulting XML structure is then transformed into various output formats with the use of modified tools from VALLEX 1.0. The export formats are HTML with navigation among the characteristic features of the dictionary

entries, Postscript document for printing including page index of all verbs and PDF, which allows navigation through the document in the same visual form as for hardcopy printing.

5 Conclusions and Future Directions

We have displayed the details of the VerbaLex verb valency frames dictionary and described the augmentation of the PDTB VALLEX 1.0 format that was needed for encapsulation of new semantic roles and links to the Czech wordnet entries.

The nearest development of VerbaLex dictionary includes adding several thousands of verbs and implementation of sophisticated checks of the correctness of the entered data with direct linking of the editing tool to wordnet editor and to the syntactic analyzer.

6 Acknowledgements

This work has been partly supported by Czech Science Foundation under the project 201/05/2781 and by Grant Agency of the Academy of Sciences of CR under the project 1ET400300414.

References

1. Pala, K., Sevecek, P.: Valence českých sloves (Valencies of Czech Verbs). In: Proceedings of Works of Philosophical Faculty at the University of Brno, Brno, Masaryk University (1997) 41–54
2. Smrz, P., Horak, A.: Determining Type of TIL Construction with Verb Valency Analyser. In: Proceedings of SOFSEM'98, Berlin, Springer-Verlag (1998) 429–436
3. BalkaNet: BalkaNet Project Website, <http://www.ceid.upatras.gr/BalkaNet/> (2002)
4. Stranakova-Lopatkova, M., Zabokrtsky, Z.: Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In M. González Rodríguez, C.P.S.A., ed.: LREC2002, Proceedings. Volume III., ELRA (2002) 949–956
5. Hoa Trang Dang, Karin Kipper, M.P., Joseph Rosenzweig: Investigating Regular Sense Extensions Based on Intersective Levin Classes. In: Proceedings of Coling-ACL98, Montreal CA (August 11-17, 1998) <http://www.cis.upenn.edu/~mpalmer/>.
6. Beth Levin, ed.: English Verb Classes and Alternations: A Preliminary Investigation. The University of Chicago Press, Chicago (1993)
7. Pala, K.: Valency Frames and Semantic Roles (in Czech). In: Proceedings of Slovko 2005 Conference, Bratislava (2005)
8. 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)
9. EuroWordNet: EuroWordNet Project Website, <http://www.i11c.uva.nl/EuroWordNet/> (1999)

Orwell's 1984 – Playing with Czech and Slovak Versions

Jaroslava Hlaváčová

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague
Malostranské nám. 25
hlava@ufal.mff.cuni.cz

Abstract. The contribution will describe an experiment with the automatic translational tool Česílko that was designed for translation between very close languages, namely Czech and Slovak. We had at our disposal the Czech version of the Orwell's novel 1984, morphologically annotated, and the Slovak version without the annotation. We automatically translated the Czech version into Slovak and compared the result with the automatic morphological annotation of the Slovak version. We evaluated the experiment using manually annotated part of the Slovak version. During the experiment we had to deal with different inconsistent morphological tagsets used for the annotation of different input data. Conversions among them made the hardest problems of the whole work. The contribution will concentrate on overcoming inconsistent tagsets, as the problem is quite common.

1 Overview

The impulse for the experiment was a decision of our Slovak colleagues to use Orwell's novel *1984* for a manual morphological annotation. There exists the Czech counterpart — Czech translation of the same text morphologically annotated during the project MULTEXT-East that ran in 1995 – 1997 ([7]). It consisted in creating morphological tagsets for several languages of the Central and Eastern Europe and their use for annotation of the Orwell's novel.

The aim of the recent experiment was to preprocess the Slovak text by the automatic morphological annotation to speed the manual work of human annotators. Our idea was to try, if the automatic translator could help. We used our tools developed before, especially the automatic translator Česílko. With its help, we translated the Czech version automatically into Slovak. At the same time, we automatically morphologically annotated the Slovak version. Then we searched word forms from the original Slovak version in the translation, comparing their morphological tags. As the both Slovak texts (automatic and manual) differed, we had to develop a simple Slovak-Slovak aligner.

In the following sections, we will denote by *C* the Czech and by *S* the Slovak translation of the text *1984*.

2 Česílko – Translational Tool for Very Close Languages

Czech and Slovak are very close languages. So close, that an idea of a literal translation is not so crazy as it certainly is for the majority of other language pairs. The automatic translational tool Česílko takes advantage of the closeness. Its results were very promising ([4]). There exists even its extended version for Lithuanian, with additional syntactic modules overcoming the greater distance between the languages ([3], [6]).

There are 2 possible ways, how to use the tool Česílko:

1. for translation between Czech and Slovak
2. for morphological annotation of Slovak texts

2.1 Translation from Czech to Slovak

The translation is based on morphological analysis of the input text, in our case the Czech version. It assigns a set of pairs [lemma, morphological tag] to every word form. The set of pairs can contain tens of items for certain word paradigms, but the average is 3.6 pairs per word form. The basis for the morphological analysis is large monolingual morphological dictionary of Czech covering more than 800 000 lemmas (see [1]).

Next step is tagging, or disambiguation, and lemmatisation. It consists in choosing one pair [lemma, morphological tag] from the set of all possible ones. The tagging is based on statistics (see [2]) and achieves the accuracy between 92 and 93%.

For the translation itself, the bilingual Czech - Slovak dictionary is used, containing the data necessary for translation of lemmas and appropriate tags. It translates the Czech lemma assigned by the tagging and produces its concrete form in the second language (Slovak) according to the original (Czech) morphological tag selected by the tagger.

The translator is able to skip over the first two steps — morphological analysis and tagging — if the input text already contains a unique pair [lemma, morphological tag] for every word form. Thus, we could use our manually annotated data as the input.

2.2 Morphological Analysis and Tagging of Slovak

The morphological analysis of Slovak works identically as the morphological analysis for Czech described in the previous paragraph. It uses large monolingual Slovak morphological dictionary created by J. Hric covering more than 100 000 lemmas for the morphological analysis, and the same statistical methods for tagging.

3 The Data – Orwell's novel *1984*

For our experiments, we used 2 inputs:

1. the Czech version **C**
2. the Slovak version **S**

3.1 The Czech 1984 and its Pre-Processing

We had at our disposal the whole text of 1984 manually morphologically annotated, the result of the project MULTEXT-East ([8], [7]). As the morphological annotation of the Czech text was made manually, we can suppose that it was errorless. However, it was necessary to unify the tagsets. The original tagset (let us denote it **TC1**) for the manual annotation was different from the tagset **TC2** (described in [1]) used by the translator and the both sets were not possible to transfer 1:1.

Namely, the tagset **TC2** uses compound values for several morphological categories, which is not the case of the tagset **TC1**. For instance masculine gender is not always further distinguished between animate and inanimate (especially for pronouns where there is often difficult or impossible to decide the right possibility) and has a compound tag for the both. Or, some morphological categories are possible to assign the value X, meaning "there can be any appropriate value". It is used for instance for the category of case with indeclinable nouns. The tagset **TC1** does not allow these possibilities. To make the both tagsets compatible, we had to exclude the detailed tags of **TC1** that did not have their counterpart in the tagset **TC2**, and replace them with less detailed tags containing the compound values. Of course, we have lost some information.

The incompatibility between the tagsets **TC1** and **TC2** was the first main source of errors. The other was the "translation" of the Czech tagset **TC2** into the Slovak one **TS1**, because of some differences, though tiny, between the both grammars. The tagset we used for automatic annotation of the Slovak version was created by J. Hric on the base of the Czech tagset of J. Hajič.

Later, we will mention the last source of errors, which is again connected with incompatible tagsets, this time the Slovak ones.

For the translation described in the previous section we used the annotated text **C** with converted tags, skipping the first two steps of the general procedure.

Let us denote **CTS** the Slovak translation of the Czech text (Czech Translated into Slovak).

3.2 The Slovak 1984

The Slovak version of 1984 we used for our experiments was the "manual" translation of the novel, made by the human translator Juraj Vojtek [9].

The automatic morphological analysis of the Slovak text was processed, followed by the automatic statistical disambiguation (tagging). It consisted in choosing one pair [lemma, morphological tag] from all possible pairs proposed by the Slovak morphological analyzer. Let us denote **SA** the result of the automatic morphological analysis (Slovak Analyzed). Every word form now has assigned (possibly several) pairs [lemma, morphological tag], one of them automatically selected as the correct one.

There is an example of a single word form *piesku* (in English *sand* in genitive, dative or local) after this phase of procession:

```
<f>piesku<MDl>piesok<MDt>NNIS3-----A-----<MMl>piesok<MMt>NNIS2--
---A-----<MMt>NNIS3-----A-----<MMt>NNIS6-----A-----
```

It has the following attributes assigned:

- <f> original word form, taken from the input text;
- <MMl> lemma assigned on the basis of the morphological dictionary; can be multiple;
- <MMt> tag assigned by the morphological analyzer; can be multiple;
- <MDl> lemma chosen by the tagger from the set created by the morphological analysis; always unique;
- <MDt> tag chosen by the tagger from the set created by the morphological analysis; always unique.

4 Slovak-Slovak Aligner

The first experiment consisted in comparing the Slovak texts *CTS* and *SA*. As the Slovak translation *SA* from English was processed by an alive translator (a writer) and the translation *CTS* from Czech was automatic, there is no surprise that the both texts differ. We could neither align them on the basis of the same positions, nor it was possible to align sentences.

Table 1. shows the main numerical differences between the both texts:

	<i>SA</i>	<i>CTS</i>
# words	83 897	79 860
# delimiters	19 165	20 505
# sentences	6 974	6 714

However, as the both versions are in the same language, Slovak, there is no need to use complicated aligners designed for pairs of different languages. Our simple aligner consisted in trying to find the same word forms in the both texts. We did not search lemmas because in different translations they could appear in different grammatical forms which would not help us. If there appears the same word form in the both texts, there is bigger chance, that the both morphological tags could be the same, though there is quite great homonymy among the forms of one lemma. However, the homonymy usually exists only among one or two categories (see our example above, where the 3 MMt tags express the homonymy among genitive, dative and locative), the rest of the assigned tag could be right and could help a human annotator anyway.

The aligner worked in the direction *SA* → *CTS*.

We looked within certain limits around the same position of the *SA*. The limit (k) became the parameter of the aligner. As the texts have not the same number of positions, we had to assign a relative position to every word of the both texts. The relative position is a number from the interval (0, 1) expressing the relative location of the word within the whole text. It was calculated according to the following formula:

$$rel = \frac{1}{N} * abs$$

where N is number of words in the text, abs absolute position of the word expressing the order of the word — for the first word $abs = 0$, for the last one $abs = N - 1$. The relative position became the other attribute (rel) of our word forms:

```
<f>piesku<MDl>piesok<MDt>NNIS3-----A----<MMl>piesok<MMt>NNIS2--
---A----<MMt>NNIS3-----A----<MMt>NNIS6-----A----<rel>0.00059597
```

To put it together, we sought in **CTS** every word form from **SA** within $\pm k$ words beginning on the nearest relative position in **CTS**.

After preliminary experiments we found out that the most aligned word forms were prepositions and conjunctions. Because of their high frequency in any text, the alignment often made a pair of two items that did not belong to each other. That is why we excluded prepositions and conjunctions from our considerations.

If we found the same word form meeting the above conditions, we added a new attribute MPt to the word form in **SA**:

```
<f>piesku<MDl>piesok<MDt>NNIS3-----A----<MMl>piesok<MMt>NNIS2--
---A----<MMt>NNIS3-----A----<MMt>NNIS6-----A----<rel>0.00059597<MPt>
NNIS3-----A----
```

In our example we see that the attributes MDt and MPt are equal. It means that the word form *piesku* from **SA** was found within our limits in **CTS** and the morphological tag chosen by the tagger is the same as that one assigned by the automatic translator.

The aligner sometimes found more than one identical word forms in **CTS** within the given span, especially for higher values of the parameter. These items had to be ignored because it is not possible to decide automatically, which is the right one. From the rest, more than 3/4 of the aligned word forms had the same morphological tags ($MDt = MPt$).

Table 2. shows results of this experiment for the parameters 10, 20, 30, 40, 50.

Parameter k	Identical words	More MPt 's	$MDt = MPt$	% of unique alignment
10	3 829	681	2 514	79.86
20	6 268	1 827	3 496	78.72
30	7 973	2 912	3 877	76.61
40	9 297	3 894	4 125	76.35
50	10 942	4 714	4 765	76.51

It should be added, that there are 67 713 word forms that are neither prepositions, nor conjunctions. Even if the amount of equal tags is quite high, the automatic translation is not reliable enough to be used for assigning tags to a manual translation.

5 Evaluation

For the evaluation of the automatic morphological annotation we used the initial part of the novel *1984*, representing approximately one fifth of the whole text,

which had already been manually annotated with the new Slovak morphological tagset.

5.1 Differences Between Tagsets

Our Slovak colleagues decided to use again another tagset, different from the previous ones, and the incompatibility between the two systems represented a further source of errors. As the both systems of annotations are not possible to map 1:1, we had to adapt the conversion table in order that it could be used for comparing the manual and automatic results.

The biggest difference between the tagsets consists in annotating a special property of words — paradigm — by our Slovak colleagues. They distinguish up to 7 paradigms (nominal, adjectival, pronominal, numeral, adverbial, incomplete and mixed) for some parts of speech. This information is generally not possible to get from the Czech system of morphological tags. We had to ignore it and compare the results without the part of the Slovak morphological tag describing the paradigm.

Another big difference is Slovak distinguishing between adjectives and passive participia, even for already lexicalized items, which is not the case of the Czech system. In the evaluation we considered them equal.

Other incompatibilities were based again in possible compound values of the tagset *TC2*, as has been described in the section 3.1. The Slovak system does not allow these possibilities.

We solved these problems by using simple regular expressions. Some morphological tags of the Czech tagset were translated into the Slovak format with some positions "dotted", so that they could be taken as regular expressions — one regular expression then could match with several Czech morphological tags. For instance the Czech tag NNFPX-----A---- for feminine (F) nouns (NN) in plural (P) with indeterminable case (X) was converted into SUfp. , where S means noun, U incomplete paradigm, f feminine, p plural and . (dot) stands at the case position. It is not part of the Slovak tagset, but the whole tag can be used as a regular expression with an arbitrary sign at the end.

Having settled up the problems with different tagsets, we converted Czech morphological tags MDt and MPt into Slovak SDt and SPt. We also added the tag MAN from the manual annotation, so that the final word forms looked like:

```
<f>piesku<MDl>piesok<SDt>SSis3<SPt>SSis3<MAN>SSis3
```

The following example has a tag in the form of regular expression. The dot stands at the position of the paradigm, that was not possible to get from the Czech system. The rest of the manual tag is the same; we could evaluate such cases as successful.

```
<f>jeden<MDl>jeden<SDt>N.is1<SPt>N.is1<MAN>NFis1
```

5.2 Results

We introduced the last attribute AGR (agreement). It can have the following values:

- D** , if $MAN = SDt$ (agreement between the manual annotation and Slovak annotation)
- P** , if $MAN \neq SDt$ and $MAN = SPt$ (agreement between the manual annotation and the translation)
- 0** , otherwise (no agreement)

The table 3 shows the final results for word forms without delimiters:

AGR	Count	%	Explanation
D	14 226	80.35	$SDt = MAN$
P	86	0.49	$SPt = MAN \& SDt \neq MAN$
0	3 392	19.16	no agreement

The reason of not very high agreement lies in the multiple conversions among incompatible tagsets.

Let us have a look at items of no agreement. Almost one quarter of them (22.8%) are unknown words ($SDt = Q$) — words that were not present in the morphological, nor in the translational dictionary. Half of them are proper names (for instance *Winston* has the frequency 544). There is always problem with proper names because there will never exist a dictionary that would contain them all. However, it is possible to recognize them with other methods, for instance guessers ([5]).

Many errors are caused by the incompleteness of the dictionaries. One of the useful results was the list of words that should be added. However, due to the subject matter of the novel, there is quite a lot of very unusual words that are not used in the current language — *ideozločinec* (in English *thought-criminal*), *podpododdelení* (in English *subsubdivision*), some of them were even not translated into Slovak — *newspeak*, *facecrime*, or have a special Slovak ending — *goldsteinizmus*. These types of words do not belong to general dictionaries, it is necessary to recognize and determine them by different means (an automatic recognition tool, a guesser).

5.3 Conclusions

Though the automatic translational tool Česílko itself is reported to be very good, it is not possible to use it directly for annotation of the original Slovak text. However, it is possible to align the manual and automatic texts very easily on the basis of individual word forms.

Different approaches to basic inputs bring a lot of hardly surpassable barriers that are necessary to overcome at the cost of losing accuracy. For better results, it would be necessary to "translate" the dictionaries into the final tagset. Unfortunately, it cannot be done entirely automatically — it would demand a lot of manual work.

Acknowledgements

The work reported in this paper arose from the Project of Scientific and Technical Collaboration of the Czech Ministry of Education with Slovakia, number 150. It has also been supported by the grants of the Czech Ministry of Education 1ET101120503 and 1ET101120413.

References

1. Hajič, J.: Disambiguation of Rich Inflection. (Computational Morphology of Czech) Praha, Karolinum 2004
2. Hajič, J.; Hladká, B.: Tagging Inflective Languages. Prediction of Morphological Categories for a Rich, Structured Tagset. ACL-Colint'98. Montreal, Canada, August 1998. pp.483–490
3. Hajič, J.; Homola, P.; Kuboň, V.: A Simple Multilingual Machine Translation System. In Proceedings of Machine Translation Summit 2003 IX, pp. 157–164.
4. Hajič, J.; Hric, J.; Kuboň, V.: Machine Translation of Very Close Languages. Proceedings of the ANLP 2000, Seattle, USA, April 2000, pp. 7–12.
5. Hlaváčová, J.: Morphological Guesser of Czech Words. Proc. TSD 2001. Springer-Verlag Berlin Heidelberg 2001, pp. 70-75.
6. Homola, P.; Kuboň, V.: A translation model for languages of acceding countries. In Proceedings of the EAMT Workshop 2004
7. MULTEXT-East project: <http://nl.ijs.si/ME/>
8. Petkevič, V.: Czech translation of G. Orwell's '1984': Morphology and syntactic patterns in the corpus. Number 1692 in Lecture Notes in Artificial Intelligence, pages 77-82. Springer-Verlag, 1999.
9. Orwell George: 1984. Translated by Vojtek J. Bratislava, Slovart 1998. ISBN 80-7145-334-X

Czech Language Parsing Using Meta-Grammar Formalism with Contextual Constraints

Aleš Horák and Vladimír Kadlec

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
{hales, xkadlec}@fi.muni.cz

Abstract. This paper presents the latest results in the development of deep syntactical analysis of Czech as a representative of highly inflectional free-word order language. The implemented parsing system `synt` uses a fast head driven chart parser with the underlying grammar formalism based on the meta-grammar approach with effective evaluation of additional constraints needed for capturing the contextual features of analysed phrases.

1 Introduction

The syntactic analysis of running texts plays a crucial role in many areas of advanced written and spoken text processing ranging from grammar checking, machine translation or phrase identification to knowledge mining and ontology acquisition. The problem of syntactical parsing is often reduced to shallow syntactic analysis, e.g. [1], which is sufficient in many applications where the speed of the processing is more important than obtaining an exact and deep syntactic representation of sentence. On the other hand, when the final aim is a thorough meaning representation of the input sentence, a complete parsing is inevitable. This is also the case of our system, which is being implemented as a part of the Normal translation algorithm of natural language sentences to constructions of Transparent intensional logic [2].

Currently there is only one comparable syntactic analyser for Czech [3]. It is based on NCFDG (non-projective context-free dependency grammar) formalism, which is supposed to be suitable for free-word-order languages like Czech. However, it is difficult to implement an effective analyser which uses this formalism. That is why, within the `synt` design, we have chosen an approach with a fast CFG backbone parser supplemented with contextual constraints for the analysis.

2 Description of the System

We bring into play three successive grammar forms. Human experts work with the meta-grammar form (G1), which encompasses high-level generative constructs that reflect the meta-level natural language phenomena like the word

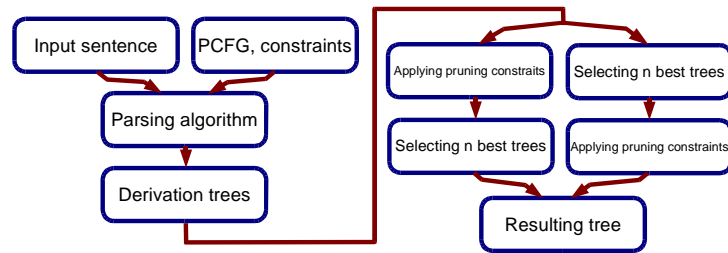


Fig. 1. Modules and data flow in `libkp`

order constraints, and enable to describe the language with a maintainable number of rules. The meta-grammar serves as a base for the second grammar form (G2) which comes into existence by expanding the constructs. This grammar consists of context-free rules equipped with feature agreement tests and other contextual actions. The last phase of grammar induction (G3) lies in the transformation of the tests into standard rules of the expanded grammar with the actions remaining to guarantee the contextual requirements.

The number of rules naturally grows in the direction $G1 < G2 < G3$. The current numbers of rules in the three grammar forms are 253 in G1, 3091 in G2 and 11530 in G3, but the grammar is still being developed and enhanced.

In the current stage of the meta-grammar development, we have achieved an average of 92.08% coverage¹ with 83.7% cases where the correct syntactic tree was present in the result. However, the process of determining the correct tree is still premature.

3 Parser

We restrict our work to lexicalized grammars, where terminals can only appear in lexical rules in the form of $A \rightarrow w_i$. This restriction allows us to simplify the implementation and it also enables to separate a lexicon from the grammar². The parsing module of `synt`, the `libkp` library provides an efficient implementation of standard parser tasks:

- syntactic analysis of sentences in natural language based on context-free grammars that can be large and highly ambiguous;
- efficient representation of derivation trees;
- pruning of the trees based on the application of contextual constraints;
- selecting n most probable trees based on computed probabilities of edge values (e.g. the frequency characteristics obtained from tree-banks);

¹ measured on 10.000 Czech corpus sentences with an average time of 276 mili-seconds per sentence.

² Actually we do not use Czech lexicon in our system because the lexical rules are created with the morphological analyser `ajka` [4].

- visualization and printing of the parsing trees in a graphical form.

All these functions are implemented as plug-ins that can be modified as needed or even substituted with other implementations. For example, we have compared four different parsing algorithms which use identical internal data structures (Earley’s top-down and bottom-up chart parser [5], head-driven chart parser [6] and Tomita’s GLR [7]). All these implementations produce the same structures, thus applying contextual constraints or selecting n best trees can be shared among them. The data flow in `libkp` library is shown in the Figure 1.

Princeton WordNet – curl:4, wave:4 Definition: twist or roll into coils or ringlets VerbaLex: natočit:3 frame: AG <person:1> ^{obl} _{whoNom} VERB PART <hair:6> ^{obl} _{whatAccus} PAT <person:1> ^{obl} _{whomDat} example: Mary curls her friend’s hair
--

Fig. 2. An example of VerbaLex verb frame.

3.1 Evaluation of Contextual Constraints

The contextual constraints (or actions) defined in the meta-grammar G1 can be divided into four groups:

1. rule-tied actions
2. agreement fulfillment constraints
3. post-processing actions
4. actions based on derivation tree

The rule-based probability estimations are solved on the first level by the rule-tied actions, which also serve as rule parameterization modifiers.

Agreement fulfillment constraints serve as chart pruning actions and they are used in generating the expanded grammar G3. The agreement fulfillment constraints represent the functional constraints, whose processing can be interleaved with that of phrasal constraints.

The post-processing actions are not triggered until the chart is already completed. Actions on this level are used mainly for computation of analysis probabilities for a particular input sentence and particular analysis. Some such computations (e.g. verb valency probability, see Section 3.3) demand exponential resources for computation over the whole chart structure. This problem is solved by splitting the calculation process into the pruning part (run on the level of post-processing actions) and the reordering part, that is postponed until the actions based on derivation tree.

The actions that do not need to work with the whole chart structure are run after the best or n most probable derivation trees are selected. These actions are used, for example, for determination of possible verb valencies within the input sentence, which can produce a new ordering of the selected trees, or for the logical analysis of the sentence.

3.2 Implementation

In the `libkp` library every grammar rule has zero, one or more semantic actions. The actions are computed bottom-up (like in `bison` GNU tool). These actions serve the purpose of:

- computing a value used by another action on the higher level;
- throwing out incorrect derivation trees.

For example, the following grammar rule for genitive constructions in Czech has three semantic actions:

```
npn1 -> np np +0.0784671532846715
    test_gen ( $$ $2 )
    prop_all ( $$ $1 )
    depends:1 ( $$ $1 $2 )
```

First line contains a grammar rule with its frequency obtained from a tree-bank. The contextual constraints are listed on the lines below it. The number 1 after the colon represents an internal classification of the action. We can turn an evaluation of actions with specified type on and off. The `$$` parameter represents the return value. The `$n` parameter is a variable where we store a value of n -th nonterminal of the rule. Notice that the presented notation is not entered directly by users. It is generated automatically from the meta-grammar G1.

The representation of the values It was shown that parsing is in general NP-complete if grammars are allowed to have agreement features [8].

The pruning constraints in `libkp` are weaker than general feature structures. It allows us to have an efficient implementation with the following properties. A node in the derivation tree has only limited number of values, e.g. the number of values for noun groups in our system is at most 56 [9]).

During the run of the chart based parsing algorithm the results of the parsing process are stored in a packed shared forest of Earley's items [5]. To compute the values, we build a new forest of values instead of pruning original packed shared forest. The worst-case time complexity for one node in the forest of values is therefore 56^δ , where δ is the length of the longest right-hand side grammar rule. Notice that this complexity is independent on the number of words in input sentence.

The values in the forest of values are linked with Earley's items. An item contains a single linked list of its values. The value holds a list of its children – one dimensional arrays of values. This array represents one combination of

values that leads to the parent value (there can be more combinations of values leading to the same value). The i -th cell of the array contains a reference to a value from i -th symbol on the RHS of the corresponding grammar rule.

3.3 Verb Valencies

In case of a really free word order language, we need to exploit the language specific features for obtaining the correct ordering of the resulting syntactical analyses. So far the most advantageous approach is the one based upon valencies of the verb phrase — a crucial concept in traditional linguistics.

We are currently preparing a comprehensive list of verb frames (VerbaLex), see [10], featuring syntactic dependencies of sentence constituents, their semantic roles and links to the corresponding Czech WordNet classes. An example of such verb frame is presented in the Figure 2. The list currently contains more than 3000 verbs which, when gathered in synonymic groups, share about 1700 verb frames.

The part of the system dedicated to exploitation of information obtained from a list of verb frames is necessary for solving the prepositional attachment problem in particular. During the analysis of noun groups and prepositional noun groups in the role of verb valencies in a given input sentence one needs to be able to distinguish free adjuncts or modifiers from obligatory valencies. We are testing a set of heuristic rules that determine whether a found noun group typically serves as a free adjunct. The heuristics are based on the lexico-semantic constraints supplemented with the information obtained from VerbaLex.

Certainly, when checking the valencies with VerbaLex, we discharge the dependence on the surface order. Before the system confronts the actual verb valencies from the input sentence with the list of valency frames found in the lexicon, all the valency expressions are reordered. By using the standard ordering of participants, the valency frames can be handled as pure sets independent on the current position of verb arguments. However, since VerbaLex contains an information about the *usual* verb position within the frame, we promote the standard ordering with increasing or decreasing the respective chart edge probability.

4 Conclusions

The presented parsing system `synt` has already proven its abilities in analysis of running texts in both speed and coverage of various sentence types. We believe, that with continuous development of the grammar, we obtain a quality general purpose system for deep syntactic analysis of natural language texts even for language with such extent of non-analytical features as the Czech language is.

The current development of the system lies mainly in probabilistic ordering of the obtained analyses with the usage of language specific features such as augmented valency frames in VerbaLex.

Acknowledgements

This work has been partly supported by Czech Science Foundation under the project 201/05/2781 and by Grant Agency of the Academy of Sciences of CR under the project 1ET400300414.

References

1. Srihari, R., Li, W.: A Question Answering System Supported by Information Extraction. In: Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, U.S.A (2000) 166–172
2. Horák, A.: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. PhD thesis, Faculty of Informatics, Masaryk University, Brno (2002)
3. Holan, T., Kuboň, V., Plátek, M.: An Implementation of Syntactic Analysis of Czech. In: Proceedings of the 5th IWPT, Charles University, Prague, Czech republic (1995) 126–135
4. Sedláček, R., Smrž, P.: A New Czech Morphological Analyser *ajka*. In: Text, Speech and Dialogue, 4th International Conference, TSD 2001, Czech Republic, Springer-Verlag, Berlin (2001) 60–67
5. Earley, J.: An Efficient Context-Free Parsing Algorithm. In: Communications of the ACM. Volume 13. (1970) 94–102
6. Kay, M.: Algorithm Schemata and Data Structures in Syntactic Processing. In: Report CSL-80-12, Palo Alto, California, Xerox PARC (1989)
7. Tomita, M.: Efficient Parsing for Natural Languages: A Fast Algorithm for Practical Systems. Kluwer Academic Publishers, Boston, MA (1986)
8. Barton, G.E., Berwick, R.C., Ristad, E.S.: Computational Complexity and Natural Language. MIT Press, Cambridge, Massachusetts (1987)
9. Smrž, P., Horák, A.: Large scale parsing of Czech. In: Proceedings of Efficiency in Large-Scale Parsing Systems Workshop, COLING'2000, Saarbrücken: Universitaet des Saarlandes (2000) 43–50
10. Horák, A., Hlaváčková, D.: Transformation of WordNet Czech Valency Frames into Augmented vallex-1.0 Format. In: Proceedings of LTC 2005, Poznan, Poland (2005) accepted for publication.

Analysis of Rule Based Phonetic Transcription Technique Applied to Slovak Language

Jozef Ivanecký

Technical University of Košice
Faculty of Electrical Engineering and Informatics
Department of Cybernetics and Artificial Intelligence
ivanecky@gmail.com

Abstract. Correct phonetic transcription is a key requirement for any automatic speech recognition or text to speech system. In this paper we describe our effort toward automatic phonetic transcription for Slovak language. We give an overview on possible techniques suited for the phonetic transcription and we explore the ability to create a rule based transcription system. We focus also on the syllabical and morphological segmentation as necessary part of rule based transcription for Slovak language as well as the possibility to use it in the real application.

1 Introduction

Among mostly used techniques for computer speech recognition belongs today Hidden Markov Models (HMM) and Neural Networks (NS). In both cases a bigger amount of training data is required. Training input for such systems is recorded speech as well as transcription of recorded utterances. The quality of final system is very dependent exactly on the quality of the transcription.

Transcription itself is possible to create manually. In case of huge training set it is power consuming job and can result in bigger amount of errors. If such data are used for the training, a drop in the final accuracy can appear. On the other side not using such data in the recognition process can have also unexpected influence. The manual transcription can be done by two different ways:

- The input data are transcribed without listening what was really said (pronounced). In such case is necessary to generate all possible pronunciations. Advantage of such approach is, that we can obtain bigger amount of data useful in the future for a recognition process. Disadvantage is that transcriber does not have to cover also the pronunciation of the speaker.
- The input data are listened and transcription contains what was really said. Such approach is invaluable especially if we have data from different dialects regions. This approach is of course more time consuming but sometimes necessary.

The second approach is automatic transcription. In contrast with the first one it is possible to eliminate many human mistakes. On the other side the creation

of such system can be very complicated. In some languages so complicated, that it is not used at all. Automatic transcription of Slovak but for speech synthesis is for example in [2]. Possibility of automatic transcription in Czech are discussed in [11], [1].

Rule based approach The easiest way to create phonetic transcription and start from the scratch are rule based systems. Such approach is also used when other techniques – required a training data set – is not possible to use. The requirements for creation of such system is existence of the rules – resp. existence of possibility to create such rules – for “correct” pronunciation in case, that an orthographic representation of the text is available. For the Slovak language very good source of such rules is [8].

One of the assumption for creation of rule based system is existence of relation between orthographic and orthoepic symbol sequence. It may be relation $1 : 1$, $1 : N$ or $N : 1$. Relations $N : N$ has to be possible to split just to the rules containing $1 : 1$, $1 : N$ or $N : 1$ only. If it is not possible, it is not possible to use rules to generate pronunciation for particular language. As example for Slovak language can be:

- ‘ $1 : 1$ ’ relation: ‘ $a \rightarrow a$ ’; grapheme a always represents phone a .
- ‘ $1 : N$ ’ relation: ‘ $i \rightarrow \{ I, I_{-}^{\wedge} \}$ ’; grapheme i represents either phone I or phone I_{-}^{\wedge} . It is context dependent¹.
- ‘ $N : 1$ ’ relation: ‘ $\{ \acute{i}, \acute{y} \} \rightarrow I$ ’; graphemes \acute{i} , \acute{y} always represent phone I .²

If it is possible for particular language to create the set of such rules then it is possible to create the system which will generate from input sequence of graphemes G the appropriate sequence of phones F . Quality of output sequence depends on the rules selection. For example if only orthoepic or also dialect pronunciation was considered etc. [4].

Statistical methods The methods discussed in this section need certain amount of training data. The training data are necessary for creation of certain statistics which are later used for generating the sequence of output phones from input orthographical representation. In the most of the cases decision trees are used. The basic idea is simple.

- The bigger training set is created. It consist of the W, Φ pairs where W is word in orthographic form and Φ is appropriate phonetic transcription. For each such pair the graphemes to phonemes assignment is created. This is the key task. First the assignment from Φ to W is created and then backward by simple rules assignment $W \rightarrow \Phi$. This is outcome from assumption, that number of graphemes is never smaller then number of phonemes. It is true

¹ In this case if i is followed by short vowel, it can be either group of vowels or diphthong.

² Such relations is possible to simplify to set of ‘ $1 : 1$ ’ relations.

for English. For Slovak just in case when we do not use stand-alone phone for the glottal stop. In Czech language – where glottal stop is more common – mentioned assumption is not correct. To generate assignment $\Phi \rightarrow W$ is possible to use for example HMM [7].

- From the pairs W, Φ created in previous step is for each symbol w created a decision tree based on the phone sequence ϕ generated for previous graphemes w and information about orthographical context of symbol w . So create tree is used later for conversion from orthographical to phonetical representation.

Advantage of such approach is fact that in case of big training set the final system will be able to deal with the foreign word better than rules based system. Such solutions are part TTS system where are very often a words, not included in the pronunciation dictionary. The hybrid system consisting of both discussed methods are also popular.

2 Rules based system for Slovak

The decision whether to create rules based transcriptions system for Slovak or use some statistical approach was determined by lack of sufficient amount of the training data required for the statistical approach. The rule based approach required the analyze of all phonetical phenomena in Slovak language. The analyzed problems can be divided in to following sections:

- **Vowels** – There is distinct relation between graphemes and phones in Slovak. This is true if vowel is between two consonants or between consonant and word boundary. The problem can be the transcription of vowel “ä” but it is discussed later.
- **Diphthongs** – Slovak has four diphthong: *ia, ie, iu, ô*. Not all combination of vowel *i* and *a, e, u* are diphthongs. Here was necessary to distinguish if sequence of vowels is a real diphthong or vowels sequence.
- **Vowels sequence** – The vowels sequence in Slovak are only in prefixed words, compounded words and foreign words. The pronunciation of vowels sequences in foreign words was adapted to pronunciation of Slovak words. The definition of the pronunciation rules was not problematic.
- **Hard vocal begin and glottal stop** – In Slovak language in contrast with for example Czech it appears very rarely. Because there are no exact rules for pronunciation of this phenomena, we did not consider it in the rules definition.
- **Vowel *ä*** – The pronunciation of *ä* is considered as advanced pronunciation. In the standard pronunciation there are big regional differences. From this reason we used for transcription of *ä* phone *E*.
- **Vowels *ö, ő, ü, ů*** – The vowels *ö, ő, ü, ů* are in Slovak language only in loan-words. In Slovak pronunciation there are very often replaced by the closest vowel. To stay within the Slovak phone set, we used such approach for the rules definition.

- **Voice assimilation** – In the case of consonants pronunciation the main problem is voice assimilation. The assimilation appear on the morphematic borders. While morphematic boundary on the words borders is clear, inside the word it seems to be as bottle neck of the proposed system. From the [9] is clear that to creation the morphematic boundary detector the morphematic dictionary is required. We solved this problem by simple morphematic dictionary.
- **Doubled consonants** – The doubled consonants in Slovak are pronounced on morphematic boundaries. In case of 3 or more consonants the doubled consonants appear very reary. Other simplification are applied here. Here we had to deal again with all problems related to morphematic boundaries detection.
- **Soft and hard consonants** – The graphemes *t*, *d*, *n*, *l* have two possible pronunciation in Slovak language. Either soft or hard one. To be able to deal with all pronunciation phenomena, the syllabic boundaries are required. We used statistical approach described in [5]. In case of pronunciation of graphemes *t*, *d*, *n*, *l* many exceptions exist. They were added in to the exception dictionary.
- **Consonants [m, F, n, N, ʃ]** – In case of the pronunciation of [m, F, n, N, ʃ] there are two main problems:
 - There is no exact definition of pronunciation of grapheme *n* in Slovak language.
 - Neither IPA or SAMPA has phonetic repertoire required for the resolution used in Slovak phonetic [6].
 From this reasons we did not fully follow the rules defined in [8]. With the simplification the rules definition was without special needs.
- **Other consonants sequences** – for the correct transcription of the consonant sequences is very often necessary to know morphematic boundaries. From this reason we have to deal here with the all morphematic problems described above. The rules without morphematic boundary needs were defined and exception were added in to the exception dictionary.
- **Other rules** – Because the pronunciation of graphemes *r* and *l* was not included in [8], to defined the rules we used [10]. The definition of the pronunciation rules was not problematic was not problematic in this case. The next set of rules here are the graphemes were pronunciation is unambiguous and therefore they were not mentioned in [8]. For the computer implementation they had to be exactly defined.

As we can see from above mentioned, there are several problematic domains in the rule based transcription of Slovak language:

- In the first case it is transcription of foreign and loan-words. We expected this problem and therefor we focused on domestic words. The rules for foreign words were defined only if it was not too complicated. Many of them are in the exception dictionary.
- The second problem is need of morphematic boundaries for some rules from the assimilation section. To create reliable morphematic analyze is necessary

to use morphematic dictionary. This is the issue we were not currently able to deal with.

- The syllabic boundaries problem is necessary to solve in case of pronunciation of grapheme *j* and some others. The solution of syllabic boundaries is described in [5] and is possible to use it for transcription purposes. In this case the statistical approach was used. No external dictionaries are here required.
- The problem of non orthoepic pronunciation is specific problem which was solved just with the information from [8] and own experiences. From this reason many local pronunciation specifics were not included in the rules.

If it is possible to solve all the mentioned problems, it will be possible to create reliable rule based system for Slovak language.

3 Experiments

All the described rules were implemented in Perl language. From 255 pronunciation rules obtained from [8] and [10], 257 transcription rules were defined. Achieved results is possible to resume as follows:

- The vowels rules (vowels, diphthongs, vowels groups) were implemented in full range and testing showed the high reliability during the transcription generation.
- The consonants rules were also implemented in full range but quality of the transcription depend on the quality of the morphematic and syllabic segmentation. The next problem influencing the quality is amount of loan-words and exceptions. In the system is included also exceptions dictionary but language coverage is unknown.

During the implementation was very important to achieve the right order of the rules. The calling the rules in order as they are for example in [8] would lead to the incorrect results.

Before the rules are applied the input word is tested against pronunciation dictionary. In case the word does not belong to the exceptions, the next step is syllabic and morphematic segmentation. After that the transcription rules are applied. The rules are applied word by word considering context within the sentence. In case of transcription of large text the first step is splitting the text in to the single sentences. The sentences are then processed sequentially. Words within the sentence are processed as described above.

For the testing 100 randomly selected words were used. The words were selected from the test set for the syllabic segmentation and words with incorrect syllabic segmentation were removed. The reason for such solution was attempt to eliminate syllabic segmentation errors and get better picture about the quality of the transcription itself. For each word from the test the correct morphematic segmentation was added in to the morphematic dictionary. Testing itself was done first without the morphematic dictionary and then with the morphematic

dictionary. We wanted to know also influence of morphematic segmentation to overall quality. The generated transcription was compared with the manually created transcription.

The reason for transcription of entire sentences instead of isolated words was fact that the morphematic boundaries are also between two words and for correct transcription of assimilation phenomena the surrounding words have to be considered. The problem is detection of morphematic boundaries inside of words. The same we can say about syllabic boundaries and about the words belonging to exception dictionary.

In the following table is the transcription accuracy for the test set with and without morphematic dictionary:

	T. with morphematic dictionary	T. without morphematic dictionary
1 transcription	81 %	75 %
2 transcriptions	9 %	8 %
Overall	90 %	83 %

The analyze of the output showed that:

- Difference between the results achieved with and without morphematic dictionary is generated mostly by the words where voice assimilation rules were not applied. The voice assimilation is part of the transcription where morphematic boundaries are required. In one case it was pronunciation of doubled consonants. The problem were following 7 words: nepredpokladal (J E p r E t p O k l a d a l / J E p r E d p O k l a d a l), trikmi (t r I g m I / t r I k m I), podpíšu (p O t p I: S U / p O d p I: S U), odpálená (O t p a: l E n a:, O t p a: L E n a: / O d p a: l E n a:, O d p a: L E n a:), krúžkoch (k r U: S k O x / k r U: Z k O x), vládcom (v l a: t s t s O m / v l a: d t s O m), jesenná (j E s E n a: / j E s E n n a:).
- The second set of errors — 10 % — were mostly problems with the transcription of $t, d, n, l - t, d', \tilde{n}, l'$. Such kind of errors have to be fixed by exception dictionary because Slovak language has in case of pronunciation of $t, d, n, l - t, d', \tilde{n}, l'$ too many exceptions. In this case it was group of following words: minifutbal (m I J I f U d b a l), kandidátska (k a n J I d a: t s k a), benevolenciu (b E J E v O l E n t s I _ ^ U, b E J E v O L E n t s I _ ^ U), kabinetu (k a b I J E t U), minimálnym (m I J I m a: l n I m), veterinárny (v E c E r I n a: r n I), Tibete (c I b E c E), teroristických (c E r o r I s c I t s k I: x), jednej (j E d J E I _ ^), pevného (p E U _ ^ J E: h Ö). From above mentioned word is clear that in the most cases problem are foreign words and loan-words. If we wanted to transcribe the pure Slovak words only, the accuracy would be above 95 % if we use also morphematic dictionary.

As an example of the system we show here the transcription of the sentence *Egyptská správa nehnuteľností zamestnáva tanečnú majstra s pedálikom*. In the following table is for each word from the input sentence its phonetic transcription in the SAMPA coding.

Graphemes	Phones
Egyptská	E g I p s k a:
správa	s p r a: v a
nehnuteľnosť	J E h \ n U c E L n O s c I:
zamestnáva	z a m E s t n a: v a
tancmajstra	t a n d z m a j s t r a
s	s
pedálikom	p E d a: l I k O m p E d a: L I k O m

4 Summary

In this paper we described the analyze of possibilities to generate phonetical transcription for Slovak language. The proposed method was also implemented and tested on the real data. The achieved result showed that the rules based transcription problem is not possible to solve without external data.

References

1. Černý, M. – Matoušek, V. – Mautner, P.: An automatic creation of the pronunciation dictionary. In: *Proceedings of 3rd Slovenian-German and 2nd SDRV Workshop*. Ljubljana, 1996.
2. Daržágin, S. – Franeková, Ľ. – Rusko, M.: *Konverzia a rečová syntéza slovenčiny*. Jazykovedný časopis, 45, 1994, 1, s. 31–43.
3. Ivanecký, J.: Automatická transkripcia slovenčiny v počítačovom rozpoznávaní reči. In: *Slovenčina a čeština v počítačovom spracovaní*. Bratislava, Veda 2001, s. 109–116.
4. Ivanecký, J.: SAMPA v slovenčine a jej význam z pohľadu viacjazyčných systémov na rozpoznávanie reči. In: *Slovenčina a čeština v počítačovom spracovaní*. Bratislava, Veda 2001, s. 98–108.
5. Ivanecký, J.: Štatistický prístup pri určovaní slabičných hraníc. In: *Slovko 2003*. Bratislava, Veda 2003.
6. Ivanecký, J. – Nábělková, M.: *Fonetická transkripcia SAMPA a slovenčina*. Jazykovedný časopis, 53, 2002, s. 81–95.
7. Jelinek, F.: *Statistical Methods for Speech Recognition*. Cambridge – Massachusetts – London, The MIT Press 1998. 283 s.
8. Kráľ, Á.: *Pravidlá slovenskej výslovnosti*. Bratislava, Slovenské pedagogické nakladateľstvo 1983. 632 s.
9. Páleš, E.: *Sapfo – Parafrázovač slovenčiny*, Bratislava, VEDA 1994. 305 s.
10. Pauliny, E.: Slabičné [r][l] v slovenčine. *Slovo a slovesnosť*, 38, 1977, s. 307–310.
11. Psutka, J.: *Komunikace s počítačem mluvenou řečí*. Praha, Nakladatelství Akademie věd české republiky 1995.

Construction of Spoken Corpus Based on the Material from the Language Area of Bohemia

Marie Kopřivová and Martina Waclawičová

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

1 Introduction

Presently the Institute of The Czech National Corpus has two different corpora of the Czech language in the form of transcripts of voice recordings. They are the Prague Spoken Corpus (in Czech: PMK), recorded in Prague in the period 1988–1996 and the Brno Spoken Corpus (BMK), recorded in Brno in the period 1994–1999.

Since 2001 we conduct more recordings, especially in Prague. However, the goal is to obtain recordings from the whole area of Bohemia, not only Prague. We are helped in this by Czech language students who conduct recordings in the places of their residence, in various parts of Bohemia. Starting this year, universities in České Budějovice, Hradec Králové, Plzeň and Ústí nad Labem are also going to join the project.

These recordings will become a part of the spoken corpora in the framework of Czech National Corpus. The following article describes the process of recording — conditions, used procedures and techniques and further processing and transcription as well as our future plans and goals.

2 Places of recording

The primary target of this project is to obtain recordings of the prototypical spoken language; to capture the commonly used spoken language as a collection of language means used in every day's casual situations. Therefore, we do not concentrate on capturing particular dialects or the Common Czech.

Most of the recording took place in Prague, mainly for the practical reasons. In Prague we have a specific language situation — a mixture of people from different areas and the language is mostly levelled out. There is also a not insignificant presence of local varieties of the common speech from the border areas, Central Bohemia, North-East Bohemia, South-East Bohemia and the transitional Bohemian-Moravian area.

We determine the degree of language levelling in Prague and the border areas in combination with the influence of the area where the speaker grew up and how strong the local language variations of the area are. The influence of neighbouring areas is taken into account as well.

There are also noticeable generation differences in the language levelling. Provided we manage to obtain a sufficient number of recordings from other cities and countryside, the corpus data may enable us to follow the specifics of the speech in cities and possible differences from the speech of countryside.

3 The aim of recording, socio-linguistic categories

The main objective of the recordings is to capture the language in casual situations — therefore, we try to capture conversations of two or more subjects, who know each other well. These conversations usually take place in private, during unofficial and informal events; the topics aren't given in advance. The speakers involved are characterized — similarly to their characterization in the already existing corpora — by three socio-linguistic categories with binary values. They are the following: gender (male / female), education (university / high school or elementary school), age (less / more than 35 years).

We do not record speakers under 18 years of age because their speech might be too specific (children's or students' slang) and varying. Given the recent size of the corpus we concentrate only on adult speakers. In addition to this information we note the type of the situation (formal — at an office, at a physician etc.; informal — conversation between friends, no limitations — these types of conversation are the most important to us). These four characteristics are essential for cataloging in the corpus. Besides these, we record additional information in a separate database.

Entries in the database can be sorted by the variables describing the speaker or the situation or by the characteristics of the recording.

For the speakers we record the following information:

1. Age at the time of the recording — enables us to create a more detailed classification by age groups.
2. Education — since the recordings are mostly done by students, there is a majority of people with university education in our database.
3. Place and region of birth — we use Bělič division of regions — Central Bohemia, North-East Bohemia, South-West Bohemia, Central Moravia, East Moravia and Silesia and the border areas.
4. Place of residence during childhood — according to the same region division.
5. Whether or not the speaker lived in the place, where the recording was done.

Furthermore, we try to record other details concerning the language situation:

1. Type of the situation (e.g. a visit, conversation during a meal at home or in a restaurant, a party, a party game). Our recent recordings are mostly from visits or conversations during meals — occasions for longer conversations. Now we try to focus on the recording of shorter communication situations — e.g. meeting in a corridor, on a street, talking to a shop assistant etc. These situations are very frequent in every day's life but are demanding for the person conducting the recording — he or she needs to find the basic

information about all the participants. Therefore, there are not many of these recordings.

2. Topic — if a major topic is present.
3. Physical presence of the speakers.
4. Readiness of the speaker — in case of a lecture, moderated discussion etc.
5. Dialogue / Monologue — we record the number of speakers and the degree of their partaking in the conversation. For example, we distinguish situations in which there is one person asking questions and more people reply.
6. Environment — private or public (e.g. a conversation with a physician takes place in a private environment but still remains formal).
7. Relationship between the speakers — we distinguish three stages — they don't know each other, they know each other, they are friends.

The recordings can be also searched based upon technical information:

1. Length of the recording — varying from 2 minutes (meeting on a street) to 2 hours (a visit, a party game).
2. Month and year of recording — recordings take place since 2001.
3. Place and area — areas according to Bělič.
4. Number of speakers — usually two or three, the maximum number was 12 during a party game.

Additional information can list the main topics, if there were any.

4 Transcription of the recordings

The recordings are transcribed using a modified version of folkloristic transcription, which has been adjusted for the needs of computer processing according to the custom of the Czech National Corpus. Intonation or other phonetic phenomena — different pronunciations of one phoneme or assimilation — are not recorded in the transcript.

The speakers are assigned numeric codes; should the person conducting the recording take part in the conversation, he or she gets a code ending with zero. This is because it might happen that the speaker, knowing he or she is being recorded, does not behave naturally — asks questions to keep up to the subject etc. Such recordings have to be excluded from the corpus or marked as formal. However, it appears these cases are very rare among our recordings.

The transcribing itself is very similar to common writing; the few differences are as follows: there are no capitals at the beginnings of sentences (because of computer processing), capital letters are used for names and some abbreviations only. Unfinished or interrupted sentences are marked explicitly — both of these are very common. The literary form of a word is kept in such cases, where the written form normally differs from the spoken form (e.g. *i-y*; *dě*, *tě*, *ně*; *bě*, *pě*, *mě*; voiced / unvoiced sounds). On the other hand, we try to capture the specific features of common speech, including regional features (e.g. *dóle*, *vzádu*, *kamen*, *zrouna*). In the cases where the spoken words are commonly pronounced differently from the proper pronunciation, we capture this difference — e.g. *sem*

(= jsem), *pudu* (= půjdu), *von výjde* (= vyjde), *já si to vezmu* (= vezmu), *dyť, kanička, řeben, kerej, práznej, muskej*. The transcripts can therefore contain doublets.

The segmentation of the written transcript is up to the person writing it. Intonation, semantic and grammatical units are taken into account. That means the transcript is not segmented by pauses in the speech but it closely resembles a normal written text.

In practise it appeared that most people conducting transcripts naturally transcribe every sentence three times — the first time they write the sentence in a form it would have in a normal written text (slightly different word order, without hesitation sounds and filling words); the second time they change the word order according to the speech and add the filling words; after the third listening they add all the hesitation sounds as well.

5 Conclusion

Presently, we have transcripts in the length of about 500 000 words. Among the speakers the most numerous are people with university education and younger than 35 years. Both genders are represented equally. For the future recordings we want to concentrate on balancing the number of speakers from different sociolinguistic categories and on recording others types of situations — especially short talks. We experiment with recording the communication taking place during the whole day, which might help us to identify the types of situations we are missing. Acoustic issues also need to be addressed — improving the quality of recordings in some cases and solving the problem of linking recordings with their transcripts.

References

1. Bělič, J.: *Nástin české dialektologie*. Praha, 1972.
2. Čermák, F.: *Pražský mluvený korpus*. <http://ucnk.ff.cuni.cz>.
3. Hladká, Z.: *Brněnský mluvený korpus*. <http://ucnk.ff.cuni.cz>.
4. Čermák, F. — Sgall, P. : Výzkum mluvené češtiny: jeho situace a problémy. *SaS* 58. 1997.
5. *Český národní korpus — BMK*. Ústav Českého národního korpusu FF UK, Praha, 2001. Available from WWW: <http://ucnk.ff.cuni.cz>.
6. *Český národní korpus — PMK*. Ústav Českého národního korpusu FF UK, Praha, 2001. Available from WWW: <http://ucnk.ff.cuni.cz>.

Multimedia Reading Book – Utilization of an XML Document Format and Audio Signal Processing^{*}

Marek Nagy

Department of Applied Informatics,
Faculty of Mathematics, Physics and Informatics, Comenius University
Mlynska dolina, 842 48 Bratislava, Slovak Republic
mnagy@ii.fmph.uniba.sk,
<http://www.ii.fmph.uniba.sk/~mnagy>

Abstract. A goal of the project Multimedia Reading Book (MRB) was to collect audio patterns and their subsistent texts - transcriptions by a playful and competitive form. The project was designated for children of elementary schools (6 - 11 y.o.). Kids took part in the project by writing a text - story, recording a story sound fluently and separately word by word. In a phase of story processing, a transcription of fluently sound was created and, thanks to a designed segmentation algorithm, a transcription of separately spoken words was made too. An XML representation of text allows to use processed stories on a www page of the project and allows a future utilization in other interactive applications (for example reading tutor). The collected and preprocessed patterns from Slovak schools make next development of speech recognition and its application on children reading learning activities possible.

1 Introduction

Using a computer in an education process is not unlikely at Slovak elementary schools. But computers are mostly used to develop basic computer skills. Teachers put off everything to a special class where children are sitting behind the computers and developing skills like a mouse clicking, a keyboard typing and so on. Older children are learning to make web pages, programming, ... It is unusual to see computers on other classes like a history, a biology, ... and for younger children on reading and writing classes. As I could find out [2] an appropriate activities with computers helps teachers to make classes more interesting and effective. Teachers can leave mechanical and repeated educational things to computers and saved time devote to children.

For example at reading classes, one child is reading an article and other schoolmates are listening to him. A teacher is also following and fixing him. If it is a child with good reading skills then weaker readers are running late. And if a weak schoolmate is reading better readers are bored. A solution to this

^{*} This work was partially supported by national grant UK/379/2005.

problem can be in using a special computer application - a reading tutor. See [7]. A click-able story is a simplified version of it. Children are reading for yourself and if they do not know how to read a word they just click on the word and listen to its sound [8]. To make such a reading book I suggest the Multimedia Reading Book (MRB) project. It can be later extended by a speech recognition mechanism which will correct children reading mistakes. The project is designed so that children help to develop the MRB for other children. The result of the project is the MRB with many click-able stories.

2 Description of project contributions

Contributions to the MBR project were sent by the on-line form at the web page. A correct story contribution had to contain: a text of the story, a sound of the fluently spoken story and a sound of the separately spoken story word by word.

The sound of separated words allows better cutting because the MRB is intended so that children can click on one word and listen to a sound of this word. I tried to use a fluently spoken story but to detect boundaries of words was very difficult (sometimes impossible) either automatically or manually. Of course coarticulation effects at the start and at the end of words were hearable. Therefore I decided to record a sound of the story in two variants: fluently and separately.

Some personal data had to be filled too: a name, a surname, an age, a sex, a grade and a school. A part of the submitted contribution could be a picture for a variegation. The contribution (the story) is intended to be like a package which can be used on the MRB project web page but later it can be plugged into other applications. (For example reading tutor [7].)

All the form data were uploaded on a project server where an appropriate XML document was created (named *story.xml*) and sounds files were stored (named *fluently.wav*, *separately.wav*). The picture was possibly stored too (named *picture.jpg*). The contributions were hierarchically saved in directories which were named by a unique number (*ID*). Now, the XML document is ready to immediately provide itself for client browsers. The story is correctly shown thanks to an XSL transformation file. The both sound files were send and processed at the computer cluster CPR [9].

An example of the *story.xml* file:

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/xsl" href="../work_xsl.xml"?>
<work id="312">
  <navigation prev="0" next="451"/>
  <accompanied-by sound="yes" wordsounds="yes" picture="yes"/>
  <person name="Lukas" surname="Krofta" age="8" grade="2.a"
    pseudo="pseudonym" gender="boy"
    school="ZS" address="Street, locality">
```

```

        teacher="name of teacher"/>
<headline>Vylet na Zamkovsku chatu.</headline>
<story>
  <p> My sme ziaci malicki,</p>
  <p> no mame sikovne nozicky.</p>
</story>
</work>

```

The element *navigation* serves as links to a previous and next story¹. If the story is accompanied by a sound, sounds of words or a picture then an appropriate element attribute (*accompanied-by*) is switch on. The element *person* contains all personal data. Then the *headline* and text of the *story* follow. The story can be formatted by a paragraph element *p*.

I have obtained 125 stories but 5 are without a sound of separated words.² Some contributions do not have completely spoken stories with the separate manner.³ A sample rate (16kHz) and a bits precision (16b) of the sounds have not been fulfilled often. And it causes downgrade of a word boundary detection and downgrade a sound articulation.

Totally it has been collected 12170 words in fluently spoken sounds. It includes 4114 different words. And it has been collected 10998 words in separately spoken sounds. It includes 3826 different words.⁴

3 Word boundary detection

Word boundary detection is used to detect separately spoken words in the sound *separately.wav*. It is a bit easier because words could be separated from each other by a short silence. But the files contain different quality sounds with different parameters. I also suppose, that majority of children do not understand what is an intention of this sound kind and they have read words too quickly.

At the first, a frequency vector *hist()* of frame energies is computed along the whole story sound. The frame energy E_i is converted by a logarithmic scale and is rounded to an integer e_i .

$$e_i = \lfloor \log(E_i) \rfloor, \quad i = 1, \dots, T$$

$$\xi_h(e) = \begin{cases} 0 & e \neq h \\ 1 & e = h \end{cases}$$

$$hist(e) = \sum_{i=1}^T \xi_e(e_i)$$

¹ Stories are a double linked list what make simulating a book possible

² The web form can not check if a child sends two identical sounds.

³ Teachers of these children were invited to repair and complete the contribution but without a response.

⁴ A neutral label #sil is not considered. See section 4 about labeling.

Energy threshold θ is computed as local minimum between two outer frequency maxima e_L, e_R .

$$\theta = \min_{e_L < e < e_R} (hist(e)), \quad \text{where}$$

$$e_L = \min_e (hist(e-1) < hist(e) > hist(e+1))$$

$$e_R = \max_e (hist(e-1) < hist(e) > hist(e+1))$$

The possibility to set the threshold manually is also leaved. The size of the detection context frame N is experimentally set to value 34. For every frame positions energies above the threshold $ecount(t, \theta)$ are counted.

$$\zeta_h(e) = \begin{cases} 0 & e \leq h \\ 1 & e > h \end{cases}$$

$$ecount(t, \theta) = \sum_{i=t}^{t+N-1} \zeta_\theta(e_i)$$

A word starts at position t when $ecount(t, \theta) > 0.43N$. The end of the word $t + N - 1$ is now testing by $ecount(t, \theta) < (1 - 0.66)N$ what is test on number of energies below the threshold θ is greater than 66% of N . The process of finding the start and the end of words is mutually alternating. Before the word boundary detection waves are scaled to satisfy precondition $RMS = 0.18$ (root mean square) by the sox [11] utility.

4 Labeling

The algorithm mentioned earlier serves to detect word boundary and to create a label file *separately.lab*. Names of labels are initially taken from generic names: word001, word002, . . . By the ideal case every generic label (a detected segment) can be substituted by one word from *story.xml* without needs to insert or to delete one. To label fluently spoken sounds is simpler. Whole story is extracted from *story.xml* and every word is placed at a separate line.⁵ Therefore *story.xml* must be checked for grammatical and formating errors. Words in the story must be separated by at least one blank character (space, newline, . . .). The punctuation must be appended to the nearest previous word. It is needed make the story click-able by automatic manner.⁶

And at this point the automatic has finished. Now, the labels files must be corrected manually. The manual process starts by listening to the fluently spoken sound and correcting the *fluently.lab*. Every punctuation is deleted. Numbers may appear in the story and they must be rewrite by words.

⁵ Boundary of words are not computed in this case.

⁶ A correspondence between a word of the story and an appropriate sound is offered by a strict word order.

The next step is to correct *separately.lab*. Before it, the generated segments are relabeled according to words from *fluently.lab* in same order. This manner helps to correct word boundary in editing tool HSLab [6].

A problem can appear if the story bundle is not consistent. What variants may come? We can assume that the *story.xml* is consistent with the *fluently.wav*. If not, it will be put in harmony manually.

- A word of the story is missing in the *separately.wav*. At the appropriate position is inserted a short empty segment named #sil.
- An extra word is in the *separately.wav*. It is omit and it is not labeled.

At the end, according to *separately.lab*, words sounds will be cut out from the *separately.wav*. And then to make the click-able web page by an XSL transformation is easy.

5 Conclusion

The project MRB allowed to demonstrate basic computer skills by children 6-11 years old. At this age children learn reading, writing and arithmetic. They tried to fill out the web form with personal data and to write story by a computer keyboard. These skills are widely used and educated on Slovak schools. But a new extension consists in a story bundle creation. Children had to use an audio application and tried to record their own voice. After it they had to manage consistency of three things (two sound files and one story file). Children did it excellently though with a teacher help. 25 schools and 125 children take part in the project.

The final product - MRB [3, 10] has a bit different goals as to develop basic computer skills. The main goal is to develop reading skills of children. The MRB can be used in a classroom where children learn to read. They are reading a story and if they are not sure how to read a word they can click on the word and they will listen to a sound of the word (Click & Listen). This is just basic functionality which can be extent. A speech recognition can be utilized and children can train a pronunciation of words [7].

When you listen to stories of the MRB it is evident that children hurry to read whole the story text and therefore do not read it carefully. A teacher cannot notice all such mistakes but a computer can be consistent and follow children reading with patience [7].

References

1. Nagy, M.: PenguinQuart - a Digit Speech Pattern Collector, Bratislava, Slovak republic, Slovko 2003 (2003)
2. Nagy, M.: PenguinQuart - hlasom ovládaná edukačná hra, Trenčín, Slovak republic, INFOVEK 2004 (2005) (in Slovak).
3. Nagy, M., Ondrišková J.: Multimediálna Čítanka. Active Advice Ltd., Bratislava, Slovak republic (2005) (in Slovak).

4. Nagy, M.: Utilizing an education game PenguinQuart to develop a speech recognition of slovak digits, Bratislava, Slovak republic, Informatics 2005 (2005)
5. Psutka, J.: Komunikace s počítačem mluvenou řečí. ACADEMIA, Praha, Czech republic (1995) (in Czech).
6. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book 3.2. (2002)
7. Project LISTEN - A Reading Tutor that Listens, Carnegie Mellon University, USA:
<http://www.cs.cmu.edu/~listen/>
8. Starfall – Where Children Have Fun Learning to Read!,
<http://www.starfall.com/>
9. Computer Cluster CPR at Comenius University,
<http://www.ii.fmph.uniba.sk/~mnagy/index.php?file=Cluster.xml>
10. Multimediálna Čítanka, on-line,
http://www.infovek.sk/predmety/1stupen/citanie/citanka/rok2004_2005/index.html
11. SoX - Sound eXchange, <http://sox.sourceforge.net/>

Morphological Idiosyncrasy in Hungarian Multiword Expressions

Csaba Oravecz, Viktor Nagy, and Károly Varasdi

Research Institute for Linguistics, Hungarian Academy of Sciences,
Budapest
{oravecz,varasdi,nagyv}@nytud.hu

Abstract. The paper presents an approach for the extraction of Multiword Expressions from large corpora that is based on the morphological idiosyncrasy of certain word combinations rather than on some statistical association measure operating on co-occurrence counts. We investigate the usability of information extracted from suffix distributions and examine whether the inflectional idiosyncrasies on the members of multiword units could be a good indication of collocativity or idiomaticity.

1 Introduction

The extraction or identification of Multiword Expressions (MWEs) from large corpora has long been considered as a serious but somewhat under-appreciated problem of natural language processing [1]. In recent years, however, a wide scale of different methods have been developed and investigated for the computational treatment of MWEs. A majority of them is commonly based on some statistical association measure operating on candidate data in the form of positional or relational n-gram lists [2], generated from a corpus with varying level of linguistic annotation. For languages which offer a richer information source than English as far as individual word forms are concerned, research has just begun to focus on properties other than simple co-occurrence [3]. With Hungarian being a morphologically rich language, one such property naturally offers itself: the morphological idiosyncrasy of certain word combinations.

In our paper we will investigate the usability of information extracted from suffix distributions in MWE detection/extraction for Hungarian, and try to examine whether, at least for certain types of MWEs, the inflectional idiosyncrasies on the members of multiword units could be a good indication of collocativity or idiomaticity.

The remainder of the paper is structured as follows. In Section 2 we will give a brief description of the method based on the idiosyncratic behavior of suffix distributions in MWEs, which could be utilized (at least in theory) to demarcate MWEs from productive word combinations or to identify different MWE classes. Section 3 will discuss the data used and the analysis we experimented with, while in Section 4 we will present a case study how the applied technique can perform on selected MWE candidates. Some remarks on problems that the method has

to face follow in Section 5 and conclusions and suggestions for further work will end the paper in Section 6.

2 Extraction Method

We call a word combination morphologically or morphosyntactically idiosyncratic if the suffix distribution of its individual members as attested in the co-occurrence pattern significantly deviates from their distribution calculated from all of their occurrences. This is in some contrast to the method presented in [3], where the values for specific inflectional features are compared only in within a multiword combination and the proportion of different values (eg. singular versus plural with respect to number) is considered as an indication of morphosyntactic preference for the particular MWE. We attempt to take a more general approach and try to measure the suffix distribution in and outside the multiword unit, and use the information as an indication whether the unit can be taken as a MWE. Thus, we hope to utilize this approach in place of a co-occurrence based association measure as an independent classifier not necessarily as a post-processing step only.

The working hypothesis is the following. The candidate list consists of two-word combinations where there is some syntactic relationship between the members. We refer to the morphosyntactic features which do not depend on this specific relationship as free features. These are then either inherent features of the members or originate from outside the given syntactic relation. (For example, in a verb/object relation in Hungarian, it is only the accusative case that is enforced by this relation, therefore it is not a free feature in this combination, whereas all other morphosyntactic features, such as number, possessive, etc. are considered as free.)

The assumption is that there is a measurable difference between the suffix distribution in terms of the free features of the members of MWEs when they appear in a MWE, and the distribution that is calculated when their occurrence outside the MWE is also taken into account, and this difference can be used to identify potential multiword expressions¹. It is very much possible, however, that the syntactic relationship will already cause a significant deviation in the inflectional pattern so an alternative hypothesis should also be considered: the presence of a MWE can only be predicted when the distribution of the free features within the multiword unit differ from that measured also outside the unit but still in the same syntactic relationship.

Since both members of the multiword unit can be subject to the above analysis, the result could in theory be used to pinpoint which member of the unit has inherited more idiosyncratic properties due to its appearance in the MWE.

¹ For instance, in the expression *gyenge láb(ak)on* “weak+on leg(s)=on shaky ground(s)” the second member of the unit is almost exclusively used in the suppressive case in this specific construction. However, outside the construction it can take any other case as well as other inflectional features, like the possessive. (See the case study in Section 4 for details.)

3 Experiments

3.1 Data

The data to generate the required statistics over the suffix distributions is drawn from the 150m word Hungarian National Corpus [4], which is a POS disambiguated corpus of contemporary Hungarian. Each token in the corpus is provided with a morphosyntactic description originating from the annotation of the HUMOR morphological analyzer [5] in the following way: The direct output notation of the morphological analyzer is not suitable to be applied directly as a MSD set for two reasons: a) it is not designed to return a POS tag and a lemma for each analysis of a given word form and b) it returns several analyses at varying levels of specificity. For illustration purposes an example is given in Figure 1, which shows the analysis of *lehetőségekben* 'within possibilities'. As regards point a) note that the leftmost item in each line is tagged with a POS label but this POS may change as derivational suffixes are added to the stem. In the first line we find that the noun stem *lehetőség* features in the lexicon as a unit and in this particular case the two inflectional suffixes PL and INE obviously did not modify the POS status of the resulting word form. However, in the following line the derivation suffix COL does turn the adjective stem into a noun but this fact remains implicit in the analysis. Point b) is illustrated by lines 2-4, which unfold a derivational tree at successively finer levels. The multitude of analyses in themselves do not create any ambiguity as in this particular example they all amount to the same reading as a noun. They are mentioned here merely to illustrate the need to interpret the analyzer's output to make the data tractable.

1. <i>lehetőség</i> [FN]+ <i>ek</i> [PL]+ <i>ben</i> [INE]
2. <i>lehető</i> [MN]+ <i>ség</i> [COL]+ <i>ek</i> [PL]+ <i>ben</i> [INE]
3. <i>lehet</i> [IGE]+ <i>ő</i> [MIF]+ <i>ség</i> [COL]+ <i>ek</i> [PL]+ <i>ben</i> [INE]
4. <i>lesz</i> [IGE]= <i>le</i> + <i>het</i> [HAT]+ <i>ő</i> [MIF]+ <i>ség</i> [COL]+ <i>ek</i> [PL]+ <i>ben</i> [INE]

legend:

FN = N	MIF = Present participle
MN = Adj	COL = Adj →N deriv. suffix
IGE = Verb	INE= inessive case
HAT = modal	PL = plural

Fig. 1. A sample output of the morphological analyzer

For the final MSD notation all derivational details about the internal structure of the rightmost POS category are eliminated. Only the lemma, the POS category and the inflectional structure is preserved, which is sufficient, however, as a basis for investigations about the suffix structure and suffix distributions of word forms in the corpus. So the above example is transformed into the simplified form in Figure 2. This format represents roughly the same information as

lehetőségekben=>lehetőség\ [N] [PL] [INE]

Fig. 2. The MSD notation in the corpus

(and can in principle be mapped into) the EAGLES compliant encoding scheme developed in Multext–East [6].

A general advantage of this process for higher level of language processing in an agglutinative language is that individual suffixes that carry important linguistic information are identified as separate elements in a possible suffix sequence attached to a word stem, and this information is encoded in a concise and easily accessible way.

3.2 Analysis

The inflectional features are analyzed along multiple dimensions which are presented in Table 3.2. For each candidate (C), the inflectional features of the

Type	Position				
	1	2	3	4	5
Nominals	number	possessive	anaphoric possessive	case	degree (only for Adj)
Verbs	mode	definiteness	number/person	-	-

Table 1. Inflectional features according to word types

member word forms (w) of the multiword unit are taken along these dimensions and converted into parameters. A parameter represents a feature (F) value (v) pair in the word form’s inflection. For all parameters a relative frequency is calculated:

$$P(F_i = v_j | w_k, C) = \frac{c(F_i = v_j \text{ in } w_k \text{ within } C)}{c(C)} \quad (1)$$

For all features in a word form, the following is obviously true: $\sum_j \frac{c(F_i=v_j)}{c(C)} = 1$. The same distribution is also calculated for the member units in and outside the multiword:

$$P(F_i = v_j | w_k) = \frac{c(F_i = v_j \text{ in } w_k)}{c(w_k)} \quad (2)$$

This is the point where the two possibilities mentioned in Section 2 can each be considered, which is reflected in the change of the denominator in equation (2): wordcounts are either calculated from all occurrences or only from occurrences in the same syntactic relationship the word form has in within the multiword unit.

4 Case Study

As an initial experiment we examined the inflectional patterns of several candidates from a candidate list of lemmatized adjective+noun combinations compiled from the corpus. In Figures 3–8, some illustrative results from the analysis are presented. The lemmatized selected candidates are *gyenge láb* (a true MWE when in the superessive case, meaning “on shaky grounds”) versus *hátsó láb* (fully productive compositional combination, meaning “rear leg”). In the figures below, the black bar represents the distribution of features within the multiword whereas the striped bar stands for the general distribution.

Figure 3 shows a clear preference for the plural in the *gyenge láb* MWE, which is in nice contrast with the more even distribution for the same feature in the non-MWE *hátsó láb* (Figure 4). There is also some difference in the distribution of the possessive feature with respect to the two candidates (Figure 5 vs. Figure 6), but the most significant deviation is manifest in the distribution of the case features: the MWE is clearly biased towards a specific value (Figure 7) while the non-MWE has a balanced distribution (Figure 8).

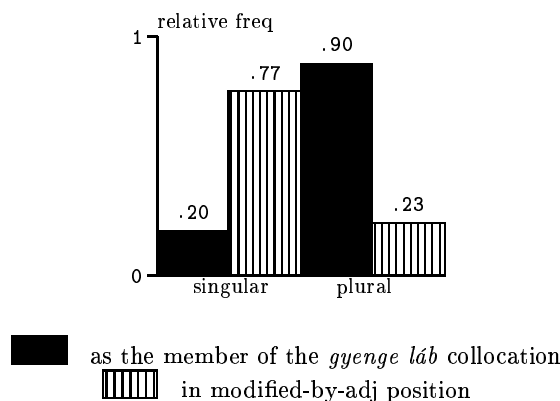


Fig. 3. Distribution of the number feature of *láb* in the collocation *gyenge láb* (in similar syntactic position)

5 General Problems with the Method

Although the analysis presented above does indicate that distributional differences in the inflectional features of multiword units could be a useful information source in search of MWEs, how the method can be generalized into an efficient classifier is currently under research. In particular, at this stage of the work we are experimenting which of the several similarity measures could be used to compare the different distributions [7] and how it can be converted into an indicator value of MWE status. However, several other problems also arise and should be considered.

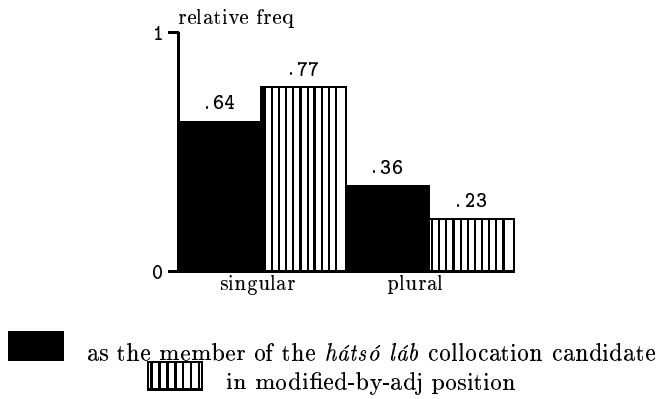


Fig. 4. Distribution of the number feature of *láb* in the collocation candidate *hátsó láb* (in similar syntactic position)

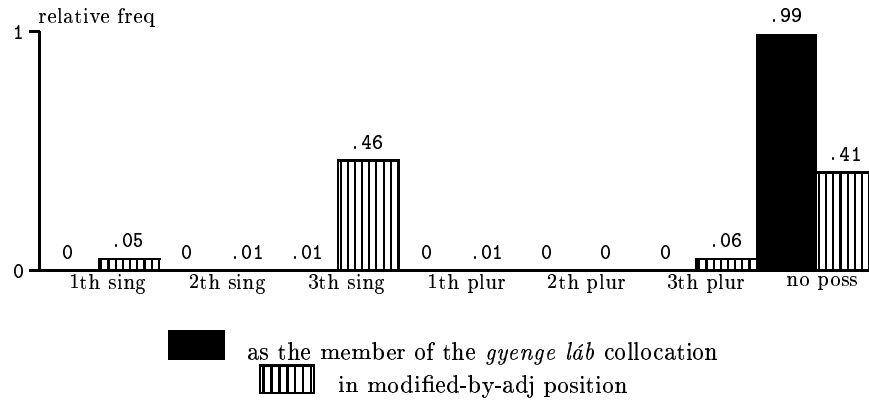


Fig. 5. Distribution of the possessive feature of *láb* in the collocation *gyenge láb* (in similar syntactic position)

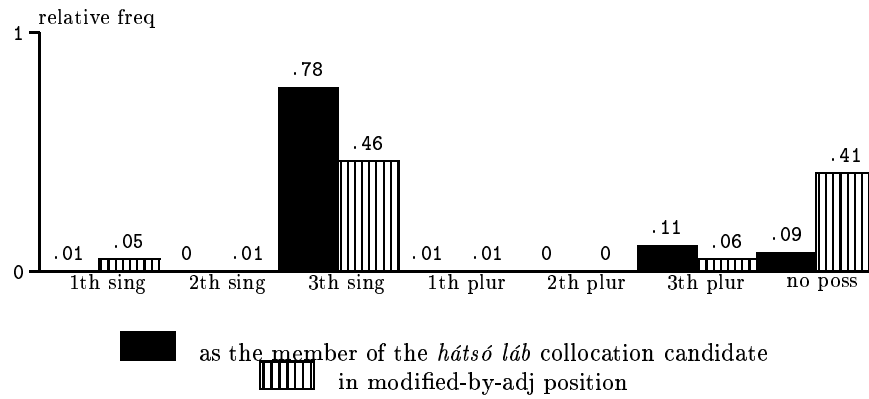


Fig. 6. Distribution of the possessive feature of *láb* in the collocation candidate *hátsó láb* (in similar syntactic position)

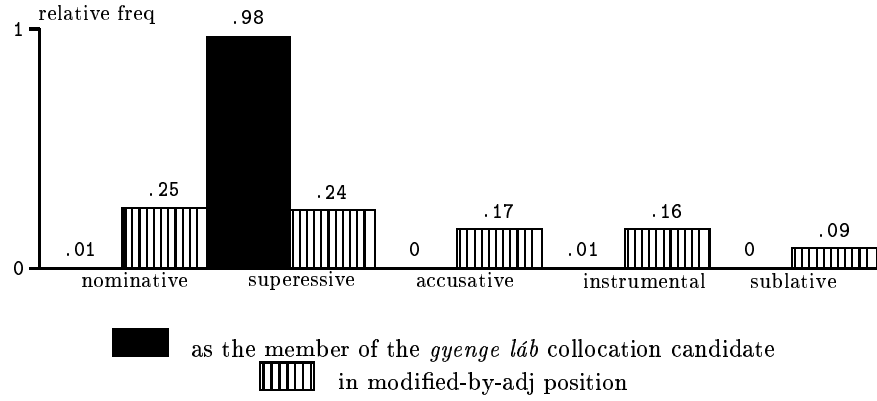


Fig. 7. Distribution of the case feature of *láb* in the collocation candidate *gyenge láb* (in similar syntactic position; only frequent cases)

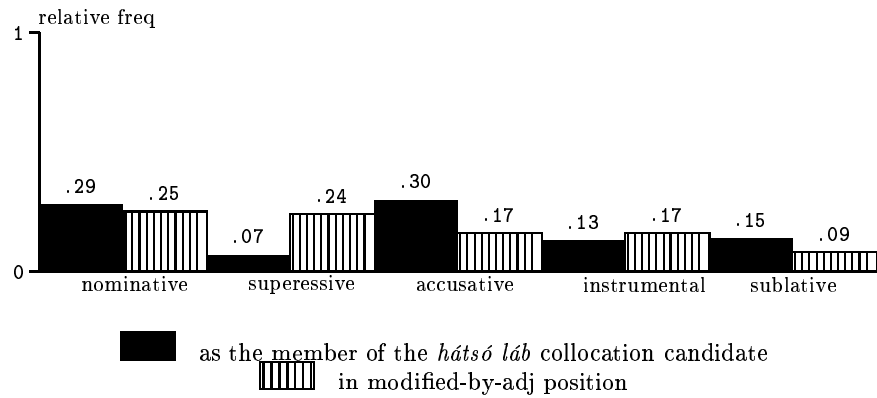


Fig. 8. Distribution of the case feature of *láb* in the collocation candidate *hátzó láb* (in similar syntactic position; only the frequent cases)

Beside the ubiquitous data sparseness, one of the problems that can cause significant difference in the distributions is polysemy. The inflectional distribution of a word form is the sum of the distributions of its different meanings. However, the particular word as a member of a collocation candidate is rarely polysemous, since the other member usually disambiguates its meaning. The distribution for the different meanings can differ, and it follows that the distribution of the collocation candidate also significantly differs from the joint distribution even when the multiword is not an MWE.

Furthermore, in a head-argument relation the distribution of some of the free features of the argument can behave idiosyncratically if they correlate with the features of the head. For example, "I lost his head" is semantically ill formed, because the person feature of the verb and the object have to be in agreement. So the distribution of the person feature of the head comes not from the opacity or morphological rigidity of the phrase, but the distribution of the person feature of the verb "lose". So for the time being it still remains an open question how the relevant features, and only those, can be selected for the comparison of distributions.

6 Conclusion and Future Work

We have examined whether we can use yet another information source for MWE extraction, which is based on the inflectional characteristics of word forms appearing in multiword units and tries to utilize their morphological idiosyncrasy. A clear advantage is that only minimal linguistic annotation is required. This approach might be suited in languages where word combinations carry rich inflections with long suffix sequences following the stem.

There is ample room for further work with respect to testing the general usability of the method. A first step is to examine how the procedure applied in our case studies can be extended into a useful measure of collocational status of multiword units, and it will also be worth exploring next how much the results are language specific, applying the method to other languages of similar characteristics.

References

1. Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for nlp. In: Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), Mexico City, Mexico (2002) 1–15
2. Evert, S., Krenn, B.: Computational approaches to collocations. Introductory course at the European Summer School on Logic, Language, and Information (ESSLLI 2003), Vienna. (2003)
3. Evert, S., Heid, U., Spranger, K.: Identifying morphosyntactic preferences in collocations. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004) 907–910

4. Váradi, T.: The hungarian national corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas (2002) 385–389
5. Prószéky, G., Tihanyi, L.: Humor – a morphological system for corpus analysis. In: Proceedings of the First TELRI Seminar in Tihany, Budapest (1996) 149–158
6. Erjavec, T., Monachini, M.: Specifications and notation for lexicon encoding, COP Project 106 Multext-East, Deliverable D1.1 F (Final Report) (1997)
7. Lee, L.: Measures of distributional similarity. In: 37th Annual Meeting of the Association for Computational Linguistics. (1999) 25–32

Valency Frames and Semantic Roles (in Czech)

Karel Pala

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno
pala@fi.muni.cz

Abstract. In this paper we pay attention to valency frames of Czech verbs and two-level notation for semantic roles. This work is now continuing in NLP Laboratory at FI MU where the VerbaLex valency database is being built (see also Hlaváčková, Horák in this volume). First, we pay attention to the inventories of semantic roles (deep cases) as they exist in various projects. We discuss their aspects, particularly, their low compatibility with the real lexical data existing in corpora (as an example, some of the semantic roles used in Vallex and Verbalex are discussed). We deal with a two-level inventory of the semantic roles designed for VerbaLex, which exploits the selected items from the EuroWordNet Top Ontology and from the Set of Base Concepts introduced in EuroWordNet. It enables us to get closer to the lexical units in a corpus text and allows us to handle the verbs whose semantic roles are inevitably too general (e.g., *vidět* (*see*), *slyšet* (*hear*), *držet* (*hold*), *dostat* (*get*), ...). The complex valency frames exploiting Word Sketches are introduced. We ask a question whether the complex valency frames can reasonably work also for the verbs in languages other than Czech, e.g. for Bulgarian. The experience of the Balkanet project shows that the answer to the question is positive.

1 Introduction

From the semantic point of view the verbs represent lexical (also logical) units denoting relations, processes, events, action, states etc. This is immediately reflected on the syntactic level by the fact that the noun and adverbial constituents in natural language sentences are organized around the verbs as their arguments. The meaning of the noun groups occurring in sentences together with a verb depends on the meaning of the particular verb. Thus we can say that the verbs or verbal expressions definitely determine the meaning of the noun groups in sentences. The deep cases or semantic roles have been introduced for the description of the meaning of the noun groups [6].

The verbs together with their semantic roles are now in the centre of the attention of many research projects, for example, the following ones can be mentioned as interesting: FrameNet [1], Context Pattern Analysis [7], SALSA for German [18], for Czech – Vallex [23] and VerbaLex [8,9].

As we said, the meaning of a verb determines meaning of the whole sentences, i.e. what it is about, what it refers to. Thus, inevitably, in lexical resources for NLP (lexical databases) verbs have to be described as completely as possible – the different kinds of the verb frames have been developed for this purpose depending on quite often different theoretical approaches.

We will be working with **valency frames** that can be characterized as data structures (tree graphs) describing predicate-argument structure of a verb which contains a verb itself and the arguments determined by the verb meaning (their number usually varies from 1-5). The argument structure also displays the semantic preferences on the arguments. On the syntactic (surface) level the arguments are most frequently expressed as noun or pronominal groups in one of the cases (seven in Czech and similar numbers in other highly inflectional languages). The semantics of the arguments is typically described as belonging to a given semantic role (or deep case), which represents the subcategorization features (or selectional restrictions). Thus valency frames typically consist of:

1. the syntactic (surface) information about the syntactic valencies of a verb, i.e. what **morphological cases** (direct and prepositional ones in highly inflected languages such as Czech) are associated with (required by) a particular verb, and also **adverbials**,
2. **semantic roles** (deep cases) that represent the subcategorization features (or selectional restrictions) required by the meaning of the verb. The number of the roles (deep cases) in various theories and resources ranges from 10-60. We will refer to them as to **inventories** (or collections) of the semantic roles.

This sort of information about verbs is typically given in the individual valency lists, there are, however, noticeable differences between the particular inventories of the semantic roles. The collections of the roles typically depend on the theoretical framework the particular research group is following and it can hardly be said that one is better than the other. What can be observed, however, is the fact, that (according to my knowledge) the inventories are usually being devised without really thorough testing against empirical (corpus) data. Hanks and Pustejovsky [7] are an exception, their context patterns are built from a corpus data.

Surface valencies can be usually obtained from the standard (both paper and electronic) dictionaries, e.g. for Czech we exploited the representative Dictionary of Written Czech (SSJČ [3]) or Dictionary of Literary Czech (SSČ [4]) where the syntactic (surface) valencies can be given either explicitly or implicitly through the examples. What, however, is missing in these dictionaries is the information about the possible semantic roles. The explanation is rather simple: when SSJČ was written and printed (1960) the theory of the deep cases had not existed. But even in the contemporary dictionaries of English like NODE [13] we would not find systematic information of this sort in an appropriate formal notation.

2 What Do We Already Have

2.1 Vallex 1.0

With regard to the Czech resources containing valency frames we should first mention Valency Dictionary of Czech (Vallex 1.0,) that has been developed in UFAL at the Charles University [23], according to [23] Vallex 1.0 presently contains about 5000

Czech verbs. Theoretically, it is based on the functional generative approach (FGD, [17, 20]) and uses the inventory of the semantic roles (functors or actants) that was developed earlier by Sgall, Hajičová, et al [20] and originally containing about 50 items. The inventory has been modified and now it contains 33 functors (semantic roles) (see [23]). It is to be remarked that it is not well-balanced, while functors with adverbial meanings are quite detailed others are too general or even missing. The list is now used:

- 1) in building the Prague Dependency Tree Bank (PDT) – for its annotation on the tectogrammatical level [23]. The verb frames in Vallex 1.0 contain information about Czech morphological cases and the semantic roles (functors) taken from the mentioned inventory. They are also associated with the particular senses of the verbs and contain information about idiomatic constructions belonging to the respective verbs.
- 2) It can be seen that the inventory of the semantic roles used in Vallex is quite closely associated with FGD theory, with the consequence that the individual functors are rather general and therefore not allowing to discriminate sufficiently the more subtle semantic (lexical) distinctions typical of the verb arguments. It can be remarked that with regard to the functors (semantic roles) FGD tries to reflect the distinction [20] between linguistic meaning and logical meaning (probably expressed by a logical form). A question can be asked how well this theoretical distinction could be justified if tested against empirical (corpus) data.

2.2 Verbalex 1.0

The second Czech resource being now developed is a list of Czech valency frames – the work is going on within the Verbalex project at FI MU ([8] and [9]). Verbalex now contains approx. 3469 verb literals with 1807 valency frames gathered in the synsets. The goal is to have 15 000 verbs from BRIEF [15] in Verbalex soon. It differs from Vallex 1.0 in several points:

- 1) verb entries are linked to the Czech [15] and Princeton WordNet 2.0 [5], i.e. they are organized around the respective lemma in synsets with numbered senses,
- 2) the inventory of the semantic roles is inspired by the Top Ontology and Base concepts as they have been defined within EuroWordNet project [21]. Thus we work with roles like AG(ENT), ART(IFACT), SUB(STANCE), PART, CAUSE, OBJ(ECT) (natural object), INFO(RMATION), FOOD, GARMENT, VEHICLE and others (see [8, 9]), that do not appear in Vallex and other inventories.
- 3) we use **two-level notation** that consists of the general labels, as the just mentioned ones, and subcategorization features (selectional restrictions) which are represented by the literals taken from PWN 2.0, e.g. AG(person:1)

animal:1), or PAT(garment:1), SUBS(beverage:1), etc. This solution allows us to specify large groups of words through the hypero/hyponymy (H/H) relation and obtain a higher degree of the sense discrimination. As we show below, even more detailed sense specification (subcategorization) is necessary, therefore, we enhance the two level notation to obtain what will be called a **complex** notation (complex valency frames – CVFs).

The valency frames in Verbalex include both the surface and deep valencies in the way shown below. The example for one of the senses of *vidět:4* (*see:1* in PWN 2.0) shows how valency frames in Verbalex are constructed:

```
* SPATŘIT:2, UVIDĚT:1, VIDĚT:4
~ dok: spatřit:2 dok: uvidět:1 ned: vidět:4
AG(kd01;<person:1,animal:1>;obl)+++VERB+++ANY((koho4|co4|
S?);<anything:1>;obl)
- synon: dok: dohlédnout:3
- example: dok: spatřil dívku
- example: ned: turista viděl les, vlk viděl zajíce
- use: prim
```

For the verb *vidět* (*see*) the following frame(s) can be found in Vallex 1.0:

1. ACT(1;obl) PAT(4,že,zda,jak;obl) [f-10122-4] [D] %modified
 -freq: 26
 -example: vidí chlapce / že chlapec přichází; vidí
 perspektivu; vidí se
 -ewn: 1,2
 -class: vnímání
 -use: prim
 -reciprocity: ACT-PAT
2. ACT(1;obl) DIR3(;qua) MANN(;typ) [f-10122-5] [E]
 -example: vidí (na protější budovu) dobře
 -ewn: 1
 -class: vnímání (schopnost)
 -use: prim
3. ACT(1;obl) PAT(4,že,jak,zda;obl) LOC(;obl)
 [f-101226] [F]%modified
 -example: vidí na Petrovi únavu / že je unavený; vidí
 na trhu zeleninu,
 -class: vnímání, -use: posun ...(+ other 5 senses)

3 What Is Missing

As an example consider verbs, *vidět* (*see*) and *slyšet* (*hear*) (other verbs like *řít* (*say*), *vědět* (*know*), *zapomenout* (*forget*), *dostat* (*get*), *držet* (*hold*) can be examined in a similar way). Their common feature is that while their left argument can be most typically labeled as AG(person:1|animal:1|organization:1), their right argument can refer, in fact, to any entity. We can *see people, animals, buildings, trees, see, stars, things, ideas* ..., obviously the list can be quite large. In most of the inventories the role

(functor) PAT will be used to label the right argument for the mentioned verbs and researchers will be quite happy about it (as explicitly states e.g. Z. Žabokrtský in his dissertation on p. 70 [23]). Let us have a look in corpus (SYN2000 [24]) how the verbs of perception *vidět* (*see*) and *slyšet* (*hear*) behave in Czech corpus texts: then a question comes up immediately: is not such labeling **too general**, can it really describe the obvious semantic distinctions as the corpus evidence shows? Let us go to corpus data and see what nouns (entities denoted by them) appear as right arguments of the verbs like *slyšet* (*hear*) or *vidět* (*see*) (or any other). To explore the corpus data we use the Word Sketch Engine (see [10]) which allows us to obtain automatically the Word Sketch table for Czech verb *vidět* (*see*) from the corpus SYN2000 (the similar data can be obtained from BNC, the comparison is quite interesting).

has_obj4	8215	has_subj	6555
důvod	212	divák	206
příčina	97	Norman	23
budoucnost	82	člověk	327
věc	138	návštěvník	54
karta (žlutá, červ.)	55	Gott	10
situace	106	Smiley	12
problém	109	Leto (r. Duna)	11
obrys	15		
perspektiva	22	není	6
běльмо	6	odborník	40
spousta (lidí)	34	máma	11
šance	43	trenér	50
možnost	83	svědek	21
světlo	43	analytik	16
svět	118	ekonom	19
rozdíl	50	Cipro	6
východisko	35	Kalibán	5
film	64	Brandon	7
silueta	9	pes	25
nebezpečí	28	oko	45
kus (světa)	29	Ježíš	14
tvář	39	šance	28
přízrak	7	fanoušek	15
smysl	33	Brenda	6
chyba	29	pozorovatel	13

Table 1. Word Sketch of the verb *vidět* (*see*), freq. = 61979

It is obvious that in SYN2000 the verb *vidět* behaves in the following way: its subject or left argument can be labeled with the general tag AG(ENT) (freq. 6555) which can be subcategorized with tags denoting **persons** or **animals** (*pes (dog)*). The noun *oko (eye)* in the list is obviously an error caused by incorrect tagging. The verb's object or right argument can be generally labeled as PAT(IENT) which is further subclassified with the tags containing the **lists of** the nouns one can see in the Table 1 (freq. 8215).

Obviously, even the two level notation mentioned above cannot be regarded as satisfactory because it cannot capture the co-occurring nouns as we can see them in the Word Sketch Table 1 for *vidět (see)* above. To overcome this problem we introduce an enhanced notation – we are going to speak about **complex valency frames (CVFs)** in which the labels of the semantic roles will consist of two or more levels. CVFs allow us to take into account the corpus data (for Czech obtained from SYN2000, for English from BNC), and therefore we get a more adequate picture that is not dependent mainly on our introspection. It can be, however, objected that the data obtained from Word Sketches may contain errors in tagging and may not be quite complete (size of the particular corpus may not be sufficient) but in any case we can be sure that they offer a reasonable approximation of the verb collocability. In this way we are going to develop CVFs also for other Czech verbs. For English the slightly different corpus based approach is applied by Hanks and Pustejovsky in their Context Pattern Analysis [7]. If we examine the Word Sketch for *vidět (see)* as well as the individual lines in the respective concordance and try to classify them using some ontological categories we can obtain the following complex valency frame (or structure) for *vidět (see)*:

(CVF1) AG(osoba | zvíře | organizace) – *vidět* – PAT(SITUAT{*situace, problém, věc, svět, rozdíl*})
 CAUSE{*důvod, příčina, smysl, chyba, nebezpečí*}
 STARTPOINT{*východisko, perspektiva, budoucnost, možnost*}
 OBJECT{*film, karta, tvář, silueta, obrys, světlo, spousta, svět*}

We can see that the syntagmatic relations in the complex valency frame can be described by the functors (semantic roles) like AG or PAT (and others) but the subtle semantic relations require the more detailed labels as we show in (CVF1). The system of the semantic labels is being developed for Verbalex database. The reason why FGD and also other systems like, e.g. VerbNet, use only general functors probably follows from the fact that theoretically they have not been designed to consider appropriately the lexical data found in corpora.

3.1 The Verb *slyšet (hear)*

If we take verb *slyšet (hear)* and have a look at its Word Sketch Table 2 we can observe the following picture:

has_obj4	2187	12.9	post_od	118	9.9	has_subj	1594	5.6
hlas	159	38.04	pan	9	17.6	hlas	53	25.4
křik	38	36.52	člověk	5	7.39	slovo	57	22.81
zvuk	68	33.52	has_obj2	77	7.0	ucho	21	22.78
střelba	43	32.08	slovo	14	21.56	Smiley	10	21.75
hučení	13	27.98	hlas	6	14.78	volání	14	21.29
rachot	15	26.88				Leto	9	21.09
nářek	16	26.02				Jack	10	19.46
výkřik	22	25.93				člověk	75	18.53
výstřel	19	25.09				střelba	12	18.19
slovo	81	24.98				křik	8	17.96
smích	25	24.69				zvuk	15	17.17
volání	20	24.11				smích	11	17.02
řev	13	23.74				posluchač	10	16.27
ozvěna	15	23.71				Norman	6	16.05
skřípění	9	23.17				řev	5	14.46
hluk	19	23.17				rána	9	12.81
tráva	21	21.89				trenér	16	12.77
praskot	7	21.83				hluk	6	12.7
hukot	9	21.74				soused	7	12.07
bzukot	6	21.16				výkřik	5	11.8
šplouchání	6	21.1				volič	8	11.25
zpěv	17	20.8				kouč	6	11.19
pleskání	5	20.66				maminka	6	11.01
pláč	11	20.53				divák	9	10.79
klapot	6	20.29				návštěvník	7	10.18

Table 2. Word Sketch of the verb slyšet (hear), freq = 19526

(CVF2) **AG** (osoba | zvíře | organizace | ucho) – *slyšet* – **PAT**(SOUND{*hluk, rachot, hukot hučení, klapot*}|

SHOOT{*výstřel, střelba, rána*}|

VOICE{*křik, výkřik, řev, zpěv, pláč, nářek, smích, volání*}|

WORD{*slovo*}|

NOISE{*bzukot*<hmyz>, *šplouchání*<voda>, *pleskání*<voda>}|

IDIOM1{*trávu růst*<neodůvodněné podezření>}|

The verb *slyšet* (*hear*) offers a picture that is not so different from what we could observe for *vidět* (*see*). We have tried to use ontological categories that seem to fit to *slyšet* (*hear*). It can be seen that in both (CVF2) the PWN 2.0 sense numbers are not used because we consider CVFs experimental so far. However, there is an important reason why the PWN sense numbers cause some problems here – PWN 2.0 is not based on the corpus data and the sense discrimination in it is too fine grained (and arbitrary). The Word Sketches we are working with are inevitably leading to the other discrimination of the senses than the one occurring in PWN 2.0. The compromise has to be found – the sense numbers used in Czech WordNet have as their translation equivalents only those PWN 2.0 synsets that can be regarded as acceptable if confronted with corpus data. In general, PWN 2.0 and 2.1 obviously call for a serious reconstruction which unfortunately is not on the horizon.

We could continue with the other verbs mentioned above, having their Word Sketches at hand. A reasonable procedure will be to begin with highly frequent verbs belonging to the semantic classes that are easy to distinguish, as e.g. verbs of perception above or verbs of eating, drinking, verbs expressing emotional states, verbs of weather, etc. The main task now is to prepare the full system of the subcategorizing semantic labels fitting to all the processed Czech verbs (15 000).

The more detailed description concerning the two-level notation used in Verbalex and the way how the valency frames are represented using XML can be found in [9] (in this volume).

4 Can CVFs Be Universal?

The existing valency lists are usually independent databases that may be parts of other lexical resources as e.g. WordNets. In our case, Czech valency list, i.e. VerbaLex, is incorporated into Czech WordNet and through ILI also to PWN 2.0 and other WordNets. Some other resources like VerbNet are linked to Princeton WordNet 2.0 via sense numbers as well.

In the Balkanet project (see [11]) we have included the list of Czech Valency Frames (1500 verbs) in the Czech WordNet to obtain the more detailed description of the predicate-argument structure of verbs (now being developed as VerbaLex, see below [8,9]). The valency frames developed for Czech have also been successfully used for obtaining valency frames in Bulgarian and Romanian WordNet [11]. As it was expected the necessary changes have been related mostly to the surface valencies.

5 Relation to Ontologies

In the recent developments in the field of the NLP we observe a growing interest in the data structures known as **ontologies** through which researchers are trying to classify semantically lexical resources covering various domains. Above we have paid an attention to the one of such structures, the Top Ontology in particular, developed within EuroWordNet (see [21]). It yields a general semantic classification of the English and

other languages' word stock and can be used as a base for a semantic classification and subclassification of the verbs, and particularly, it can be related to the inventories of the semantic roles for verb frames. The ontologies represent theoretical constructs designed from the „top“ and as such they are not directly based on the empirical evidence, i.e. corpus data. Thus there is a need to confront the ontologies and the inventories of the semantic roles that can be derived from them with the corpus data and see how well they can correspond to them. However, Hanks and Pustejovsky [7] try to build the inventory of the roles from the „bottom“. It is in agreement with the need to develop the subcategorizing tags allowing to capture and classify the lists on nouns yielded by the Word Sketches.

6 Conclusions

We have proposed the notation for complex valency frames for Czech verbs that are based on the data extracted from the corpus and obtained by means of the Word Sketches. The proposed notation for complex verb frames has an experimental nature and calls for further testing and validation.

The next step is to write the CV frames for at least 5000 most frequent Czech verbs (and their English equivalents) selected from the list of verbs that are already included in Czech WordNet. This has to be done mostly manually though the various software tools such as Verbalex tool (see [8, 9]), VisDic [22] and dictionary writing system DEBII [2] presently developed at NLP Lab. FI MU will be used.

Together with this Czech WordNet has been linked with AJKA, morphological analyzer for Czech [19], which presently contains approx. 450 000 Czech word stems and is able to capture some selected word-formation relations.

It is obvious that the presented CVFs with the more detailed subcategorization features are far from being complete – it should be enlarged appropriately, in our estimation up to 100 tags may be necessary if we want to get reasonably close to the real lexical data as they occur in corpus texts. In other words, what we are trying to develop should, in fact, lead to a broader but more specific ontology than TO.

Validation of the proposed frames through the real texts will be done within a small but well defined domain (law).

It also has to be tested how the proposed CVFs will work for the larger semantic fields where the semantic classes of verbs will come significantly into the play. The important assumption here is that semantic classes of the verbs should be helpful in checking the consistency of the inventory of semantic roles since in one class we can expect roles specific only for that class. For example, with verbs of clothing the role like GAR(ment) and its respective subclassification can be reliably predicted, similarly it should work for other verb classes, such as verbs of eating, drinking, wearing, emotional states, weather verbs and many others.

Acknowledgements

This work has been partly supported by the Czech Grant Agency under the project 201/05/2871 and by Czech Ministry of Education under the project LC536.

References

1. Baker, C. F., Fillmore, Ch. J., Lowe, J. B.: *FrameNet Project*, in: Proceedings of the Coling-ACL, Montreal, Canada, 1998.
2. DEBII, <http://nlp.fi.muni.cz/projekty/deb2/clients/>, 2005.
3. Dictionary of Written Czech (SSJČ), Academia, Praha, 1960.
4. Dictionary of Literary Czech, Academia, Praha, 1986, 1991.
5. Fellbaum, Ch.: (Ed.) *WordNet: An Electronic Lexical Database*, MIT Press (1998).
6. Fillmore, Ch. J.: The Case for Case, in: *Universals in Linguistic Theory*, eds. E. Bach and R. Harms, Holt, Rinehart and Winston Inc., 1968, p.1-88.
7. Hanks, P., Pustejovsky, J.: *A Pattern Dictionary for Natural Language Processing*, 2004, forthcoming.
8. Hlaváčková, D., Horák, A.: Transformation of WordNet Czech Valency Frames into augmented VALLEX 1.0 Format, Proceedings of LTC Conference, Poznaň 2005.
9. Hlaváčková, D., Horák, A.: Verbalex – New Comprehensive Lexicon of Verb Valencies for Czech, Proceedings of the Slovko Conference, Bratislava, 2005 (in this volume), <http://nlp.fi.muni.cz/verbalex/>.
10. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine, in: Proceedings of the 11th Euralex Congress, Université de Bretagne – Sud, p.105-116, 2004.
11. Deliverable D 8.1, Koeva, S.: Bulgarian VerbNet, EU project Balkanet, 2004.
12. Levin, B.: *English Verb Classes and Alternations: a preliminary investigation*, The University of Chicago Press, 1993.
13. *New Oxford Dictionary of English*, (ed. by P. Hanks), Oxford University Press, 1998.
14. Pala, K., Ševeček, P.: Valence českých sloves (Valencies of Czech Verbs), *Studia Minora Facultatis Philosophicae Universitatis Brunensis*, vol. A45 (1997), Brno, p. 41-54.
15. Pala, K., Smrž, P.: Building Czech WordNet, *Romanian Journal of Information Science and Technology*, vol. 1-2, p. 89-97, Bucarest, 2004.
16. Palmer, M., Rosenzweig, J., H. Trang Dang, Kipper, K.: Investigating regular sense extensions based on intersective Levin classes, in: Proceedings of Coling-ACL 98, Montreal, August 11-17, 1998 (www.cis.upenn.edu/~mpalmer/).
17. Panevová, J.: On Verbal Frames in Functional Generative Description, Part I, *The Prague Bulletin of Math. Linguistics* 22, Prague 1974, pp.3-39; Part II, Prague 1975, pp.17-71.

18. Pinkal, M., Erk, K., Kowalski, A., Padó, S.: Building a Resource for Lexical Semantics (SALSA Project), Proceedings of the 17th International Congress of Linguists, Prague, 2003.
19. Sedláček, R., Smrž, P.: A New Czech Morphological Analyser *ajka*, *Proceedings of TSD 2001*, Springer-Verlag, LNAI 2001, p.100-107.
20. Sgall, P., Hajičová, E., Panevová, J.: The meaning of the sentence in its semantic and pragmatic aspects, ed. by J. Mey, Reidel, Dordrecht – Academia, Praha, 1986.
21. Vossen, P.: (Ed.), EuroWordNet: a multilingual database with lexical semantic networks for European languages, Kluwer Academic Publishers, Dordrecht, 1999.
22. VisDic, <http://nlp.fi.muni.cz/projekty/visdic>, 2004.
23. Žabokrtský, Z.: Verb Valency, Ph. D. Dissertation, MFF UK, Prague, 2005.
24. SYN2000, http://ucnk.ff.cuni.cz/syn2000_bonito.html, 2003.

Question Answering in Polish Using Shallow Parsing

Dariusz Piechociński¹ and Agnieszka Mykowiecka^{1,2}

¹ Polish-Japanese Institute of Information Technology,
Warsaw, Koszykowa 86, Poland
dpiechocinski@gmail.com

² Institute of Computer Science, Polish Academy of Sciences,
Warsaw, Ordonia 21, Poland
Agnieszka.Mykowiecka@ipipan.waw.pl

Abstract. Question Answering applications have capability to provide information through answering users' questions. In this paper, we present an application that answers users' questions formulated in Polish using the data from the Wikipedia. Our application is one of the first attempts at implementing such a system for Polish. The answer consists of one to five sentences selected on a basis of a calculated relevance measure.

1 Introduction

Currently one of the most common tasks performed using a computer is searching through the Internet. However, this seemingly easy task usually turns out to be much more complex than previously suspected. Data stored throughout the Internet will turn into information only if it is correctly grouped and answers a specific question. To get such information, a user must formulate a question in a specific way depending on the search engine used. Furthermore, a user must manually determine which documents are useful and more importantly, which parts answer the question at hand. The query set by a user is usually formulated as a list of keywords, which may occur in a large number of documents. These keywords may have many different meanings, completely unrelated to the ones the user has in mind. This means, that in order to find the answer to our question, we would need to repeatedly add new keywords to our list as well as specify the context in which each of them is used. Using natural language questions in standard search engines frequently produces results that have no relevance whatsoever to the researched topic.

On the other hand, deep methods giving precise analysis of natural language queries, are still not general enough to be used as a means of specifying unrestricted types of information. A reasonable solution to answer questions posed by the user, could be applying a shallow parsing analysis that helps in better formulating a query for a search engine. Systems of this kind are created for many natural languages and have many practical applications. This paper describes a computer application that is one of the first attempts at implementing such a system for Polish. The program answers questions relating on information contained

in the Wikipedia Internet encyclopedia (<http://pl.wikipedia.org>). The subject of a question has to be listed within the Wikipedia's keyword database for the program to work properly. User questions are transformed into the main subject that is used to select an encyclopedia entry and a list of other keywords that are searched within the encyclopedia article. Moreover, there is an attempt to use a question structure in order to help in finding the more adequate answer.

The searching procedure is as follows. For every sentence that is part of a definition of a keyword, a coefficient is calculated that describes its use as a potential answer to the question. The sentence receives points if it contains those elements of the question which were obtained as a result of the simple analysis done at the beginning. If a question type is specified, then elements that are adequate for this specific type also receive positive points (eg. if the question is of "when" type, some sort of date expression is expected in the potential answer). In the general case, the answer to a question is a sentence from the encyclopedic definition of our keywords. The program cites a maximum of five sentences whose coefficients are higher than a system defined threshold.

The rest of the article goes as follows. In Section 2 we briefly present some related work. Section 3 describes main aspects of our application. The results of the evaluation procedure are given in Section 4. We conclude with some remarks on future work in Section 5.

2 Related Research

Works similar to our approach have been undertaken for the English version of the Wikipedia. The QUARTZ [1] application has been presented during TREC-13 [6] competition, where questions were divided into small sets and each set was devoted to one subject. This application uses WordNet to determine an answer type (standard type like: *person*, *date* or more complex like: *actor*, *capital*, *year*) and then gets a subject definition from the encyclopedia. A name of the subject was given together with the question so there was no need to resolve it. The definition is searched for phrases of the same type as the question type. Phrases are scored according to their location in the text — those at the beginning are considered to be more important so they get a higher rank. Phrases get also scored if they contain a word from the question (or synonym).

Brill [2] presented a question answering application that uses shallow parsing methods. This approach in many cases gives satisfactory results, as question answering mostly relies on pattern matching between the user question and the candidate answer. The question was reformulated into few queries that were used in the Google search engine. Seven types of questions were defined and each had some special rewriting rules that were used for constructing the queries. Candidate answers were selected from summaries of the first 100 pages returned by Google. The final answer was selected using the n-gram model.

The Question Answering problem is not very popular in Poland and there exists no complete question answering application for Polish. However, some information is available on the WebStorm [3] program. This application uses

deep analysis methods; it resolves speech parts, phrases, proper names, etc. It uses its own dictionary of words and their cases, grammars, question and answer patterns. Grammars are used for resolving some patterns in a question and for applying a corresponding answer pattern. No test results for this application are publicly available at the moment.

3 System Overview

Our application answers users questions using information included in Polish Wikipedia. Firstly, user's question is analyzed in order to identify an encyclopedia entry in which an answer should be searched for. In the general case, the answer to a user's question is a sentence from this entry definition. For two selected types of questions special procedures of formulating a precise answer were defined. Because we gather information from the Wikipedia, questions have to address directly an encyclopedic entry (in other words, any question has to contain such an entry word). Processing a user question is then divided into three stages (illustrated more precisely in the Fig.1) :

- question analysis and an encyclopedic entry resolving,
- selecting an entry definition from the encyclopedia,
- definition analysis and answer generating.

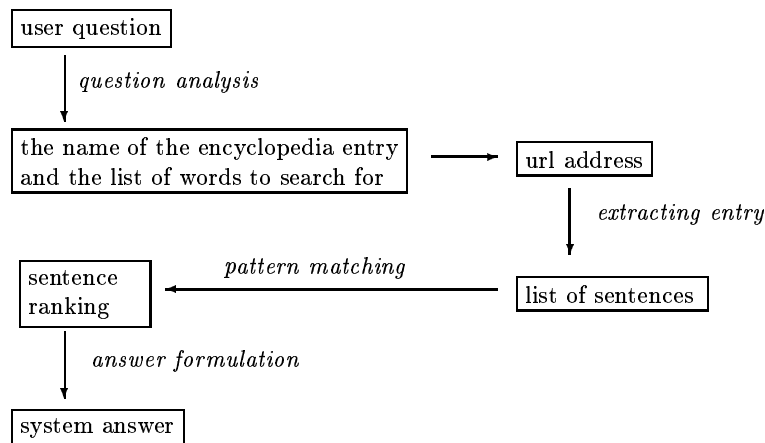


Fig. 1. A diagram of the application work flow.

3.1 Question Analysis

At this stage, we resolve the subject of a question. This subject is then compared to the elements of the list of encyclopedic entries. Moreover, there is an attempt to identify within a question a 'verb' and a 'noun phrase'. All this is done with two heuristics that use only pattern matching.

The first question analysis method relies on resolving proper names in the question.³ We are looking for proper names because in most cases the occurrence of a proper name in the question means that this is a subject of the question, thus it is a good candidate for an encyclopedic entry title. If there are several words starting with a capital letter separated by non-capitalized words, then only first sequence of the capital words is regarded as a name for an encyclopedic entry. For example, in question (1) only “Albert Einstein” string will be taken.

- (1) Kiedy Albert Einstein otrzymał nagrodę Nobla?
 when Albert Einstein received prize Nobel
When did Albert Einstein receive the Nobel Prize?

Next, there is an attempt to point out a verb and noun phrase. This is also done in two ways. In the first case, if there is any word between an encyclopedic entry and interrogative pronoun, such a word is regarded as a verb. For instance, in question (2) word “położona” will be regarded as a verb (word “jest” although it is also a verb is treated as a stopword and is passed over). In the second case (if the first one doesn’t occur), the first word after the encyclopedic entry name is regarded as a verb and a sequence following this ‘verb’ is regarded as a noun phrase. For example, in the question (1) about Albert Einstein word “otrzymał” will be a verb and sequence “nagrodę Nobla” will be a noun phrase.

- (2) Gdzie jest położona Wyżyna Anatolijska?
 where is located Upland Anatolian
Where is the Anatolian Upland located?

The second question analysis method is applied when there is no proper name in the question. This method use an assumption that question is constructed according to the following scheme: emphinterrogative pronoun + verb + encyclopedic entry. This assumption works very well with simply questions like (3), but it works worse with more complicated questions like (4).

- (3) Kto był wynalazcą dynamitu?
 who was inventor_{instr} dynamite_{gen}
Who was the inventor of the dynamite?
- (4) Jakie pierwiastki wchodzą w skład stopu o nazwie alpaka?
 what elements included in composition alloy_{gen} of name alpaka
What elements are included in the composition of an alloy named alpaka?

In question (4) an encyclopedic entry will be resolved as “wchodzą w skład stopu o nazwie alpaka”. There is no such entry in encyclopedia so our suggestion will be shortened continuously and finally it will contain only “alpaka” which will be a name of the existing encyclopedia entry.

It is clear that patterns used by those two heuristics are not sufficient for every question but the test results showed that they are good enough for most questions asked by users.

³ We assume that a proper name is just a word/words starting with a capital letter.

At the stage of question analysis takes also place the process of resolving the question type (thus expected answer type). There are only two question types identified at the moment: date questions (*When...*) and questions about date or place of the birth (or death). This second question type triggers special procedure which returns a precise answer (a full sentence).

3.2 Gathering Entry Definition from the Encyclopedia

After identifying an encyclopedic entry the application connects with the Wikipedia and downloads the source of the definition page. Then, every not essential element is removed (in particular html tags) so at the end there is only plain text of the definition. This process is described in detail in [5].

3.3 Answer Generation

At this stage every sentence from entry definition is analyzed to decide if it contains the answer and a ranking of all these sentences is made. As there exists no WordNet like database for Polish, for the purpose of sentence scoring a special dictionary file was made. This dictionary includes list of verbs and their synonyms. The dictionary was constructed in order to improve efficiency of the software in the case when the words in a question differ from words in a definition text. This is not dictionary of synonyms in the literally meaning of this word. It is rather a dictionary of expression that are used in the same context. For example word *napisal* ('he wrote/he has written') has following synonyms: *autor*, *stworzył*, *dzieła* ('author', 'he created', 'the works'). Every sentence from a definition gets scores for any of the following parts of user question: encyclopedic entry, verb (or synonym), noun phrase. Actual ranking is done according to the following rules:

- appearance of the verb (or synonym): +1
- appearance of the encyclopedic entry: +0,45
- appearance of the noun phrase: +1
- appearance of the date: +0,8.

The last rule is taken into account only if a question type is date. As a final answer there are returned sentences with rank higher than a system defined threshold. As it was mentioned earlier there are two question types that provide precise answer. Those question types are presented in (5) and (6).

- (5) Kiedy zmarł (urodził się) X?
When was X born? / When did X died?
- (6) Gdzie zmarł (urodził się) X?
Where was X born? / Where did X died?

When a question of this type is asked, then a special parsing is done. For example, if a question is about place of birth then we search sentences for 'urodził' word synonym. When we finds any then we look for prepositions: "w, we, na, pod, koło" (in, near) and for a word beginning with a capital letter. For example, if the question was as in (7) then answer could be as in (8).

- (7) Gdzie zmarł Adam Mickiewicz?
Where did Adam Mickiewicz died?
- (8) Adam Mickiewicz zmarł w Paryżu we Francji.
Adam Mickiewicz died in Paris, France.

4 Evaluation

The application was tested by various users over the Internet. During the test a total of one hundred questions were asked. These were factoid questions like in TREC competition. All answers have been manually analyzed to check if they were correct or not. For the purpose of this evaluation, the number of sentences used as an answer was limited to 5. It means that at maximum total number of five answers (five sentences from a definition) could be returned by the program. Each answer has been qualified as:

- correct: if at least a part of it was the right answer,
- incorrect: if no part answered the question, however the answer exists in Wikipedia,
- unknown: if the encyclopedia does not contain the answer.

Moreover, for each question a precision measure was counted. It was calculated as the ratio of a number of correct answers returned by the program and the total number of answers returned. The overall results were as follows: 13 incorrect, 13 unknown and 74 correct answers for 100 questions. Both precision and recall is equal 0.85. If we take into account each of the five answers independently then our total precision measure is equal to 0.503. This significant reduction in precision is due to the fact that the questions for which one answer was generated has a smaller affect than the ones that all five answers were correct. Equal number of incorrect and unknown answers is accidental. However efficiency of the application equal 74% of the correct answers is quite surprising. It means that using shallow parsing analysis could be effective. From among answers that were incorrect two weren't well-chosen because an encyclopedic entry in the question was resolved wrong; other two answers were located in the html table (part of the data which at the moment is removed automatically by the program); and the rest nine needed deeper language analysis.

In the second test, a comparison has been made between our program and the Start application [4]. Two hundred questions originated at the TREC-8 conference [7] have been asked. Start returned correct answers for 41 of them. For the purpose of our test, we translated all questions to Polish and asked to our application. This time, the correct answers were returned for 29 questions. This quite surprising outcome shows that shallow analysis is not much worse (and a lot easier to implement) as any its deep counterpart used by Start. Our software which use only shallow parsing methods was 30% worse than Start which use advanced methods of language analysis but some relatively easy ways of improvement can be pointed out.

5 Conclusion and Further Work

In the paper, we presented a question answering application for Polish. It answers users' questions using the Wikipedia encyclopedia as the knowledge database. As an answer, the program gives sentences taken from an encyclopedia article. There is no deep analysis, only shallow parsing and pattern matching methods are used. The software is available to the users over the Internet as the Java applet. A test evaluation showed a relatively high efficiency in question answering.

The application presented may be regarded as the first publicly available program that is able to answer questions stated in Polish. It uses simple methods and produces interesting results being a good starting point for further work in this field. In many cases, an improvement could be achieved by adding an external morphological analysis (e.g. Morfeusz [8]) and a synonym lexicon. The lack of Polish dictionaries of this latter kind, especially a version of WordNet, is very unlucky, as this kind of resources will be of a great value. Now, we cannot answer questions, if the definitions stored in the database do not contain the same words as queries (to improve a program a little bit, a small dictionary of synonyms was constructed manually). What is more, it would be desirable to develop more rules for resolving question types as well as to develop a method that will allow combining information from several sentences.

References

1. Ahn, D., V. Jijkoun, G. Mishne, K. Muller, M. Rijke, S. Schlobach: Using Wikipedia at the TREC QA Track. In Proceedings of the TREC13 (2004)
2. Brill, E., J. Lin, M. Banko, S. Dumais, A. Ng: Data-Intensive Question Answering. Proceedings of the TREC10 (2001)
3. Duclaye, F., P. Filoche, J. Sitko, O. Collin: A Polish question-answering system for business information. Proceedings of the BIS Conference, Poznań (2002)
4. Katz, B.: Annotating the World Wide Web using Natural Language. Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (1997)
5. Piechociński, D.: Automatyczne udzielanie odpowiedzi na pytania zadawane w języku naturalnym, Master's thesis, PJWSTK (2005)
6. Voorhees, E. M.: Overview of the TREC 2004 Question Answering Track. Proceedings of the Thirteenth Text REtrieval Conference, NIST Special Publication 500-261 (2005)
7. Voorhees, E. M., D. M. Tice: The TREC-8 Question Answering Track Evaluation. Proceedings of the Eighth Text REtrieval Conference, NIST Special Publication 500-246 (1999)
8. Woliński, M.: System znaczników morfosyntaktycznych w korpusie IPI PAN. Polonica XXII (2003)

In Search of the Best Method for Sentence Alignment in Parallel Texts^{*}

Alexandr Rosen

Institute of Theoretical and Computational Linguistics,
Faculty of Philosophy and Arts, Charles University, Prague
`alexandr.rosen@ff.cuni.cz`

Abstract. After a brief account of a parallel corpus project involving many diverse languages and a summary of two previous evaluations of sentential alignment tools, results are presented from tests of three automatic aligners on English-Czech and French-Czech literary and legal texts, clean and noisy. The results confirm that an alignment tool may perform well on one type of texts and fail on another type, and indicate that near-to-perfect alignment is possible when tools providing high precision are combined with manual checking, where the proofreader can focus only on those parts of the text that were either not aligned at all, or that were aligned less reliably. Further gains in precision are shown to be feasible when alignments proposed by multiple aligners are intersected.

1 Introduction

Once we have a text and its translation, is there a way to match corresponding sentences reliably and without too much human intervention? This question has been asked before, e.g, by [Langlais et al.(1998)], [Véronis & Langlais(2000)] and, most recently, by [Singh & Husain(2005)]. The answers do not point to a single all-purpose method. Different contexts may require different solutions and their choice should be based on a careful consideration of properties of the text pair and ways of using the result. The factors include structural distance between the two texts (how free or literal the translation is), typological distance between the two languages, size of the texts (a critical issue for statistical methods), acceptable error rate in terms of precision and recall, and acceptable amount of manual checking. Given the task to provide sentence alignment tools for a number of diverse language pairs and text genres with the obvious desideratum to reach a near-to-perfect result, an opportunistic mix is inevitable.

In Sect. 2 we provide background information on our parallel corpus project including over 20 languages with Czech as the pivot. Given a wide range of languages, distributed setting is required as linguists knowledgeable of specific language pairs are necessarily involved in the whole process of text acquisition, pre-processing, alignment and checking of the alignment results. At the same

^{*} This work was supported by Czech Ministry of Education, grant no. MSM 0021620823.

time, common shared procedures, tools, text formats and other resources are needed for the results to be integrated into a single corpus, maintained and queried by a parallel corpus manager. The solution to this challenge aims at maximising synergy effects of the large team of linguists as experts on the individual languages, and the main coordinator, providing project management and software infrastructure.

It is alignment that largely determines the usefulness of a parallel corpus, Sect. 3 deals with this issue, listing some candidate automatic alignment methods and providing data from previous evaluations. Sect. 4 presents results of testing three aligners on available texts, comparing them with previous evaluations. Based on the results, we argue for a strategy for integrating highly reliable automatic alignment with a minimum amount of human intervention. Finally, in Sect. 5 we explore options for combining results of different aligners to obtain maximum precision as a suitable step preceding manual checking of alignment results. This seems to be the least painful way to achieve minimum error rate for all sentences, and thus a corpus with highly reliable alignment. Sect. 6 summarises conclusions and suggests what should be done next.

2 The Project *InterCorp*

This parallel corpus project¹ is not unique in involving a larger number of languages: a portion of the Uppsala and Oslo's universities' OPUS project² [Tiedemann & Nygaard(2004)] includes 60 languages, and the Acquis Communautaire parallel corpus,³ compiled at the European Commission's Joint Research Centre at Ispra (Italy), includes 20 languages [Erjavec et al.(2005)]. Still, there are at least three aspects that make it different: distributed setup, preference for a balanced choice of text types, and a fair amount of texts with manually checked alignment.

The project is based upon an idea of integrating expertise and efforts of a number of project participants into a common shared resource, providing them with the necessary infrastructure and complying with their preferences: although for some languages it may not be easy to acquire enough texts, preference is given to balance rather than quantity, with literary texts and – at least in the initial stages – Czech originals the priority.

Due to the substantial involvement of a large number of participants, a distributed mode of pre-processing is inevitable: the current institutional participants consist of twelve departments and institutes, two of them outside Charles University, each responsible for at least one language pair. There are at least 20 such pairs, all of them including Czech as the pivot language, the other languages being as diverse as Arabic and Chinese. Guidance, coordination and support are provided by the main coordinator, the Institute of the Czech National Corpus.

¹ See <https://trnka.ff.cuni.cz/ucnk/intercorp/>, only Czech version is available at the time of writing.

² <http://logos.uio.no/opus/>

³ <http://www.fi.muni.cz/~zizka/Langtech/>

Some participants have already built parallel corpora of various sizes and fashions, using *ParaConc* as the segmentation, alignment and search tool [Barlow(1999),Barlow(2002)],⁴ and they continue to do so within the project. The challenge is to reconcile distributed pre-processing with the need to store, maintain and access the corpus at one place at the final stage. Thus, a battery of tools take care of the smooth transition between the ‘local’ format required by *ParaConc*, *MS Word* and other PC-based software and the canonical format adopted for the common shared corpus, making sure that – in the worst case when an electronic source is not available – a paper document goes through OCR, proofreading, conversion to tagged text, segmentation into paragraphs and sentences, sentential alignment and alignment checking, ending up in the XML format with stand-off alignment annotation.

3 Alignment

In most cases, reliable alignment of sentences is a necessary condition for a useful parallel corpus. Indeed, a parallel corpus is only as good as its alignment. In order to minimise the amount of manual checking, it is worthwhile to search for the best methods of automatic alignment.

The default alignment tool is an implementation of Church and Gale’s algorithm [Gale & Church(1991a)], integrated with *Paraconc*. The obvious question is whether there is a better alternative.

There are some published reports on comparative evaluation of sentential alignment. In *ARCADE*, a major project [Langlais et al.(1998),Véronis & Langlais(2000)], a number of important issues are brought up, but today the choice of evaluated tools would probably be different. Six systems were tested on French-English texts of various types (over 1M words per language), including an abridged translation of Jules Verne’s novel *From the Earth to the Moon*. Interestingly, this was a pitfall for all systems except one, which was based on a combination of techniques including sentence length, recognition of cognates (identical or similar strings) and bilingual lexicon look-up.

More recently, results of another detailed evaluation were reported by [Singh & Husain(2005)] (henceforth S&H). S&H aimed for systematic evaluation of four aligners on different text types. They used a mix of 21 samples from three different English-Hindi corpora, systematically varied in terms of size and noise (sentences added at random from other corpora). Due to practical constraints, only 1:1 links were considered. Three of the four systems have also been used in our evaluation, so results presented by S&H are examined more closely below.

Two of the four systems are based on methods matching most likely sentences by comparing their lengths, either in words [Brown et al.(1991)] – henceforth

⁴ <http://www.athel.com/para.html>

Brn – or characters [Gale & Church(1991b)] – **GC**.⁵ Both systems are quite fast and language-independent, but they assume some fixed points: **Brn** expects at least some sentences to be previously aligned, while **GC** requires identification and alignment of paragraphs (“hard regions”) across the texts.⁶ The other two systems use word correspondences: [Melamed(1997)] – **Mmd** – gives better results with a bilingual dictionary, although cognates such as punctuation, numbers and similar words may suffice,⁷ while [Moore(2002)] – **Mre** – generates word correspondences from input texts by combining length-based pre-alignment of sentences with a stochastic method (IBM Translation Model 1), the correspondences are used subsequently to improve the initial pre-alignment. In the available implementation **Mre** proposes 1:1 links only.⁸

The results are measured in recall, precision and F-measure, computed for the purpose of alignment evaluation in the usual way as in Fig. 1.⁹ Overall, the best results are achieved by **Mre** in precision (92.9) and **GC** in recall (84.3). On noisy texts, **Mre** compares with **GC** even better in precision (92.2 and 91.5, compared to 84.1 and 84.9). For ‘clean’ texts, precision of **GC** is better (98.7 vs. 95.1). **Mmd** scores worst, possibly due to inadequate tuning to the language pair, while **Brn** is marginally worse than **GC**.¹⁰ On the other hand, **Mre** shows marked improvements the more input it gets. With 10,000 sentences it wins on both clean and noisy texts in precision (100 and 98.4) and on noisy texts in recall (89.2). Rather surprisingly, it fails on an easy corpus sample with short sentences (precision 66.8), as opposed to more difficult samples (100 and 99.5).¹¹ The lessons learnt from the previous evaluations can be summarised as follows:

1. Quality of alignment depends to a large extent on properties of the input: on its formatting complexity – the presence of elements other than running text (graphics, tables, notes), on “structural distance” between the original and its translation (a scale from literal to free translation), on the amount of “noise” (such as omissions or segmentation differences/errors due to pre-processing), on typological distance between the two languages (important

⁵ Probably the most popular alignment tool, dubbed *vanilla aligner*. For an implementation see <http://nl.ijs.si/telri/Vanilla/>.

⁶ In fact, a “hard region” can be larger than one paragraph. With some loss in speed, it could be a chapter or even a book. Similarly, a “soft region” can be larger than a sentence – this way paragraphs may be aligned instead of sentences.

⁷ <http://nlp.cs.nyu.edu/GMA/>

⁸ <http://research.microsoft.com/research/downloads/default.aspx>

⁹ *correct links* = number of correct links among those proposed by the aligner, *reference links* = number of links in correctly aligned texts (the gold standard), *test links* = number of all links proposed by the aligner. *F-measure* combines recall and precision into a single measure. For a discussion of these measures in the context of alignment see, e.g., [Véronis & Langlais(2000)] and [Melamed et al.(2003)].

¹⁰ **Mmd** with appropriate tuning and a Czech-English lexicon was successfully used before on a large set of English-Czech data, see <http://ufal.mff.cuni.cz/pdt/Corpora/Czech-English/>.

¹¹ In the readme file that comes with **Mre** code a minimum of 10,000 sentence pairs is recommended for reliable estimation of a statistical word-translation model.

$$\begin{aligned}
 \textit{recall} &= \frac{\textit{correct links}}{\textit{reference links}} \\
 \textit{precision} &= \frac{\textit{correct links}}{\textit{test links}} \\
 \textit{F-measure} &= 2 \frac{\textit{recall} \times \textit{precision}}{\textit{recall} + \textit{precision}}
 \end{aligned}$$

Fig. 1. Measures for evaluating alignment

- especially for methods based on searching for *cognates* in the two texts), and – at least for some alignment methods – on the input size.
2. Alignment methods differ in their sensitivity to such properties.¹² Some methods can be trained or supplied with additional resources to handle difficult texts in a specific language pair, but it requires additional effort and/or availability of such resources. It seems that there is no single best all-purpose way to sentence alignment.
 3. As can be expected, word-correspondence methods fare better on noisy texts, but even standard sentence-length-based methods turn out to yield satisfactory results.
 4. Counting the number of correctly aligned sentence pairs as the evaluation result is not always a fair measure: sentence boundaries may not have been detected correctly (often there is no unanimous way to segment a text into sentences anyway), and a sentence pair where one sentence is a partial translation of the other should not be treated on par with a totally unrelated pair. Thus, alignments of sentences in *ARCADE* were measured also in terms of words and characters. However, for the practical purpose of building a parallel corpus, the “strict” measure in terms of alignment links seems to be sufficient, or even preferable.
 5. When correct alignment (gold standard) is available, both precision and recall can be obtained: selecting a method maximising precision may be the right move for some tasks, while the opposite may be needed for other tasks.

To answer our original question concerning an optimal choice of (a mix of) tools and procedures that would be best suited to a specific text type and language pair, with minimum manual checking and the goal of a near-to-perfect result, the inevitable conclusion would be that with various text types and diverse languages there is probably no universal solution. Instead, a new choice must be made each time a significantly new input occurs, based on experience and experimentation.

4 Comparison

Although we could not compete with the previous evaluation projects on the level of methodology and systematic exploration of text versions, we decided to

¹² S&H make this a key point of their report.

conduct a smaller scale evaluation of our own. We were interested in trying out candidate tools on our data, including Czech and at least two other languages. We used three aligners (**GC**, **Mmd** and **Mre**) from the set of four introduced in the previous section, some with additional resources or in a slightly modified version:

- Mmd**⁺ – Same as **Mmd**, with a 106K-entries English-Czech lexicon.¹³
- Mre**^{*} – Same as **Mre**, with some words in the input truncated by a character or two.¹⁴
- Mre**⁺ – Same as **Mre**, with more input data (the previously mentioned English-Czech lexicon and an English-Czech pre-aligned corpus of 830K/731K words¹⁵).

The systems were tested on a rather opportunistic set of text samples for which hand-corrected alignment was available.¹⁶ Nevertheless, the set at least partially reflects the needs of the project: the samples consist mostly of fiction, two language pairs are represented, and one of the sample includes substantial noise.

AC – This is the sample with the highest noise. It consists of 46 documents (in each language) from the English-Czech part of *Acquis Communautaire*¹⁷ (roughly 1% of the total number, eliminating those that did not contain usable data). All omissions and mismatches in segmentation were retained. As in the full corpus, the segments aligned are paragraphs rather than sentences, which, however, does not make too much difference as most paragraphs in these legal texts consist of a single sentence.

1984 – George Orwell’s novel in English and Czech. This is the most orderly sample, with just a few omissions in the Czech part.¹⁸

FR7 – Seven French fiction/essay books with Czech translations.¹⁹ The sample does not include any information about paragraph boundaries.

Quantitative data on the samples, including hand-corrected alignment counts, are given in Table 4. The percentage of 1:1 links provides a rough measure of the difficulty of the sample – the more such links, the easier the sample.

Table 4 gives counts of all types of links (n:n) for all samples and aligners.²⁰ The counts are compared in terms of recall, precision and F-measure.

As expected, the two aligners using lexical anchors perform significantly better on noisy texts (AC) than the length-based aligner **GC**, the difference reaching 10 and more percentage points in all measures. Interestingly, on AC, **Mre**⁺ is better than **GC** even in recall, although it outputs 1:1 links only. On the other hand, **GC** has better recall on the more orderly texts 1984 and F7, but it still lags behind **Mre**⁺ in precision. Actually, the relatively good performance of **GC** on F7 is surprising, given that the system expects “hard regions” to be paragraphs,

¹³ The lexicon we used is a GNU/FDL project, available from <http://slovník.zcu.cz/>.

¹⁴ This was actually due to the fact that the **Mre** perl scripts as downloaded from the Microsoft pages ignored the Czech locale setting. We are grateful to Bob Moore, the author of the program, and Pavel Pecina for their kind assistance in solving this

Text	Cz words	L2 words	Cz segments	L2 segments	All links	1:1 links
AC	62,010	74,986	3,025	2,699	2,685	89%
1984	99,099	121,661	6,756	6,741	6,657	97%
FR7	289,003	337,226	21,936	21,746	21,207	95%

Table 1. Size of the samples

	Reference	Test	Correct	Recall	Precision	F-measure
AC						
GC	2700	2683	2225	82.4	82.9	82.7
Mmd ⁺	2700	2686	2492	92.3	92.8	92.5
Mre	2700	2313	2218	82.1	95.9	88.5
Mre ⁺	2700	2375	2308	85.5	97.2	91.0
1984						
GC	6657	6633	6446	96.8	97.2	97.0
Mmd ⁺	6657	6606	6287	94.4	95.2	94.8
Mre	6657	6167	6110	91.8	99.1	95.3
Mre*	6657	6370	6320	94.9	99.2	97.0
Mre ⁺	6657	6441	6402	96.2	99.4	97.8
F7						
GC	21207	20868	19427	91.6	93.1	92.3
Mre	21207	19512	18801	88.7	96.4	92.3
Mmd	21207	21057	16161	76.2	76.7	76.4

Table 2. All links

rather than whole books, as was the case here. On F7, **Mmd** clearly suffers from the lack of resources and tuning.²¹ The aggregate F-measure distributes its favour rather fairly among all aligners, still pointing twice to **Mre/Mre⁺**. Tables 4, 4, and 4 rank the aligners by recall, precision, and F-measure and precision, respectively.

issue. The reason the faulty version is still mentioned is that with less input data it actually produced better results than the corrected version.

¹⁵ <http://ufal.mff.cuni.cz/pdt/Corpora/Czech-English/>

¹⁶ Except for one sample (AC) that was checked and corrected by the author.

¹⁷ See Sect. 2 for details.

¹⁸ This sample was produced and hand-corrected within the project *Multext-East*, see <http://nl.ijs.si/ME/>.

¹⁹ For this hand-corrected sample I owe thanks to Martin Svášek.

²⁰ Originally, a part of F7 (one of the novels, about one seventh of the total F7 size) was used for testing **Mmd** only. Surprisingly, the results were comparable to those obtained for **Mmd** on English-Czech samples, where additional resources were available. The unconfirmed explanation may be that this specific novel was very easy to align.

²¹ Although it did surprisingly well on F1, an easy subset of F7: with 96.7/97.0/96.8 for recall/precision/F-measure it is the winner in the French-Czech category.

Rank	AC	1984	F7
1.	92.3 Mmd ⁺	96.8 GC	91.6 GC
2.	85.5 Mre ⁺	96.2 Mre ⁺	88.7 Mre
3.	82.4 GC	94.9 Mre*	76.2 Mmd
4.	82.1 Mre	94.4 Mmd ⁺	
5.		91.8 Mre	

Table 3. Ranking for recall (all links)

Rank	AC	1984	F7
1.	97.2 Mre ⁺	99.4 Mre ⁺	96.4 Mre
2.	95.9 Mre	99.2 Mre*	93.1 GC
3.	92.8 Mmd ⁺	99.1 Mre	76.7 Mmd
4.	82.9 GC	97.2 GC	
5.		95.2 Mmd ⁺	

Table 4. Ranking for precision (all links)

Rank	AC	1984	F7
1.	92.5 Mmd ⁺	97.8 Mre ⁺	92.3 GC
2.	91.0 Mre ⁺	97.0 GC	92.3 Mre
3.	88.5 Mre	97.0 Mre ⁺	76.4 Mmd
4.	82.7 GC	95.3 Mmd ⁺	
5.		95.3 Mre	

Table 5. Ranking for F-measure (all links)

To enable fair comparison with **Mre** and the data in S&H, Table 4 gives corresponding results on 1:1 links. As can be expected, the results are better than for n:n links in all cells, except for **Mre**'s precision, where they are necessarily identical (the system outputs 1:1 links only). Again, there is no outright winner: **Mre** scores best in recall everywhere and **Mmd** in precision wherever additional resources were available (AC and 1984), while **GC** is marginally better in precision on F7. Taking into account variations in the amount of noise, structural differences, different language pairs and availability of additional resources, the results fall within the range of those reported by S&H.

Considering the overall results, conclusions of the previous evaluations seem to be largely confirmed. On noisy texts, **Mmd** and **Mre** fare better than **GC**, while on clean texts, **Mre** and **Mmd** tend to show higher precision than **GC**. Surprisingly, **GC** performs well on F7 without paragraph boundaries (with book as the hard region) and **Mmd** on an easy subset of F7 without bilingual lexicon. Further improvements might be achieved with the two lexically-based methods: **Mre** can be expected to gain further points with more input data and – possibly – lemmatisation, while **Mmd** may profit from creating more cognates by more tuning and better additional resources.

	Reference	Test	Correct	Recall	Precision	F-measure
AC						
GC	2391	2248	2156	90.2	95.9	93.0
Mmd ⁺	2391	2354	2304	96.4	97.9	97.1
Mre	2391	2313	2218	92.8	95.9	94.3
Mre ⁺	2391	2375	2308	96.5	97.2	96.9
1984						
GC	6440	6438	6274	97.4	97.5	97.4
Mmd ⁺	6404	6301	6287	97.6	99.8	98.7
Mre	6440	6167	6110	94.9	99.1	96.9
Mre*	6440	6370	6320	98.1	99.2	98.7
Mre ⁺	6440	6441	6402	99.4	99.4	99.4
F7						
GC	20116	19220	19427	92.6	96.9	94.7
Mre	20116	19512	18801	93.5	96.4	94.9
Mmd	20116	19714	15539	77.2	78.8	78.0

Table 6. Links 1:1 only

Overall, the results also confirm the conclusion that there is no single best alignment tool for all purposes, and that the success is to a large extent determined by choosing the right tool for a given text. Additionally, the choice might depend on how the automatically aligned texts will be used, and here the tradeoff between recall and precision comes into play.

For some applications, such as machine learning, maximising precision is probably the best strategy if manual checking is not an option. On the other hand, S&H claim that if the result is going to be manually checked before use, it is desirable to maximise recall: some decrease in precision is not going to make manual checking much more difficult.

This reasoning assumes that all links are going to be checked. On the other hand, if safe links can be identified in the result and only the rest is presented for manual checking, the amount of human effort could be substantially reduced. In this scenario, 100% precision is needed to obtain error-free alignment, but we might be satisfied even with a figure close to it. Recall is of secondary interest.

With precision close to 100%, the “unsafe” links are simply those that the aligner does not propose, they do not even exist as links yet. An alternative, less reliable method of automatic alignment can then be used to suggest links in this more difficult portion of the input.

In the following section, we explore an option to raise precision to make a scenario combining automatic alignment with manual checking more attractive.

5 Joining Forces

In order to push precision closer to 100%, a single text pair can be processed by more than one aligner and a correct link defined as one on which all (or most)

aligners agree. The set of proposed links would be smaller, but they would be safer: a decrease in recall, an increase in precision.

The results of the three aligners as solo performers, presented in the previous section, were intersected pairwise and all together. For convenience, the top lines of the two tables (5 and 5) give the counts already presented for solo aligners. Only two samples were used (1984 and F7), and **Mmd** – due to its poor performance – was excluded from the test on F7.

	Ref.	Test	Correct	Recall	Precision	F-measure
GC	6657	6633	6446	96.83	97.18	97.01
Mmd ⁺	6657	6606	6287	94.44	95.17	94.81
Mre ⁺	6657	6441	6402	96.17	99.39	97.76
GC/Mmd ⁺	6657	6279	6254	93.95	99.60	96.69
GC/Mre ⁺	6657	6354	6348	95.36	99.91	97.58
Mmd ⁺ /Mre ⁺	6657	6130	6114	91.84	99.74	95.63
GC/Mmd ⁺ /Mre ⁺	6657	6095	6089	91.47	99.90	95.50

Table 7. Merging results on 1984

	Reference	Test	Correct	Recall	Precision	F-measure
GC	21207	20868	19427	91.61	93.09	92.34
Mre	21207	19512	18801	88.65	96.36	92.35
Mmd	21207	21057	16161	76.21	76.68	76.44
GC/Mre	21207	17728	17661	83.28	99.62	90.72

Table 8. Merging results on F7

Both samples show the same pattern: F-measure is always better for an aligner in solo mode (**Mre⁺** and **Mre**), but a tandem of aligners always wins in precision, reaching 99.91 for **GC/Mre⁺** on 1984, with recall still at 95.36. This is an improvement of about 2.7/0.5 percentage points over their solo performance in precision. The gain is even more marked for F7: 3.6 points.

6 Conclusions and Future Work

1. Several conclusions of previous evaluations have been confirmed: quality of alignment depends to a large extent on properties of the input and alignment methods differ in their sensitivity to such properties. Thus, word-correspondence methods fare better on noisy texts, where sentence-length-based methods give mixed results.

2. Although none of the evaluated aligners was the overall winner, it was **Mre**, especially when supplied with additional resources, that often performed better than its contestants. Again, this is in accordance with a previous evaluation [Singh & Husain(2005)]. Still, the success is to a large extent determined by choosing the right tool for a given text.
3. Manual checking of alignment results can be done more efficiently with an automatic alignment method preferring higher precision to better recall. With precision close to 100, manual checking can focus only on links where good results are less likely. Such links are not even proposed by the aligner, although a different, less reliable aligner can be used in a step preceding manual checking of the difficult parts of the input.
4. In order to raise precision, sets of links proposed by different aligners can be intersected. Our results show that such a move improves precision by 0.5–3.6 percentage points.

The tests should be extended to more languages, text types and tools,²² and they would profit from a more rigorous methodology. But the present results already suggest that a near-to-perfect sentential alignment with a small amount of manual checking is a realistic perspective.

References

- [Barlow(1999)] Barlow, M. (1999). MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics*, 4(1), 319–327.
- [Barlow(2002)] Barlow, M. (2002). ParaConc: Concordance software for multilingual parallel corpora. In *Language Resources for Translation Work and Research, LREC 2002*, pages 20–24.
- [Brown et al.(1991)] Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Meeting of the Association for Computational Linguistics*, pages 169–176.
- [Erjavec et al.(2005)] Erjavec, T., Ignat, C., Pouliquen, B., & Steinberger, R. (2005). Massive multilingual corpus compilation; Acquis Communautaire and totale. In *2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (L&T'05)*, Poznań, Poland. Available at <http://www.jrc.cec.eu.int/langtech/>.
- [Gale & Church(1991a)] Gale, W. & Church, K. (1991a). Identifying word correspondence in parallel text. In *Proceedings of the DARPA NLP Workshop*.
- [Gale & Church(1991b)] Gale, W. A. & Church, K. W. (1991b). A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184.
- [Langlais et al.(1998)] Langlais, P., Simard, M., & Véronis, J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 711–717. Association for Computational Linguistics.

²² *HunAlign*, a tool developed within the *Hunghish* English-Hungarian parallel corpus project, is a hot candidate, see <http://mokk.bme.hu/resources/hunalign>.

- [Melamed(1997)] Melamed, I. D. (1997). A portable algorithm for mapping bitext correspondence. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 305–312, Somerset, New Jersey. Association for Computational Linguistics.
- [Melamed et al.(2003)] Melamed, I. D., Green, R., & Turian, J. P. (2003). Precision and Recall of Machine Translation. In *HLT-NAACL*.
- [Moore(2002)] Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144, London, UK. Springer-Verlag.
- [Singh & Husain(2005)] Singh, A. K. & Husain, S. (2005). Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 99–106, Ann Arbor, Michigan. Association for Computational Linguistics.
- [Tiedemann & Nygaard(2004)] Tiedemann, J. & Nygaard, L. (2004). The OPUS corpus – parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.
- [Véronis & Langlais(2000)] Véronis, J. & Langlais, P. (2000). Evaluation of parallel text alignment systems: the arcade project. In J. Véronis, editor, *Parallel text processing: Alignment and use of translation corpora*, pages 369–388. Kluwer Academic Publishers, Dordrecht.

Word Tests for Speech Understandability Evaluation in Slovak

Milan Rusko and Marián Trnka

Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9,
845 07 Bratislava
Milan.Rusko@savba.sk

Abstract. The paper gives a short review on availability of word lists for speech understandability evaluation in Slovak. Problems of the Diagnostic Rhyme test word list design for particular case of Slovak language is discussed. Changes in the definition of the tests are proposed that make it possible to create word list for this test in Slovak. The authors plan to use the designed tests in subjective synthesized speech quality evaluation.

1 Introduction

The subjective tests of the speech understandability are based on reproduction of pre-recorded spoken test words or sentences to listeners. The participants of the test write down the words that they heard or fill in some items in a test-sheet. The results are evaluated and a subjective measure of speech comprehensibility is obtained.

However, as far as we know, the specialized test word lists for these tests in Slovak have not been developed so far; thus, we have decided to make initial research steps in this direction. Our primary intention is to make a survey of the tests used for similar purpose (evaluation, etc.) and to define missing word and phrase sets for subjective tests in Slovak. The resulting test files in Slovak may be used for diagnostic purposes, for testing of hall acoustics with respect to speech comprehensibility, telecommunication channel measurements or for evaluation of Slovak speech synthesizers such as the one described in [1].

2 Word set for Speech Audiometry

The only Slovak test word list published and accepted in practice is the word set for speech audiometry [2]. It is described by its authors Bargár and Kollár as follows:

“The principle of speech audiometry is in reproduction of selected speech material recorded on audio tape into the tested person's headphone or bone vibrator, or by a loudspeaker into the free field. During the test, speech comprehensibility (i.e. the percentage of correctly heard and understood speech units at a certain intensity) is specified. The words for speech audiometry must be selected with respect to several factors. All words must be common, well-known even to people with lower education. Each group of words must represent the language as much as possible, and must have the same characteristics both from linguistic and phonetic aspects (the occurrence of

high and low formants, mono- and multisyllables, parts of speech, and, primarily, the same number of phonemes which should be represented here with respect to frequency in the language). Thus, each set of words will have the same value for the test”.

This definition makes this set similar to standard **Phonetically balanced word lists** – PB [3] The Slovak Word set for speech audiometry was successfully used for Slovak synthetic speech quality testing [4].

Decade 1	Decade 2	Decade 3	Decade 4	Decade 5
jazero	široký	dolina	adresa	opona
šíp	vek	tón	tyč	kmeň
takto	málo	žltý	tmavý	sto
nos	ľad	zrak	boj	dážď
vyrába	čaj	spieva	deň	heslo
lep	ozvena	lom	pozri	guľatý
živý	stan	nefajči	vlas	var
daň	zober	háj	uteká	zem
humno	pluh	múr	mnoho	búrka
česť	sedí	Iste	cieľ	žije
Decade 6	Decade 7	Decade 8	Decade 9	Decade 10
pokojne	farebný	povala	počúvaj	odvážny
stôl	hra	rok	vrch	loď
žiada	mäso	nedaj	cena	osem
dom	sud	meč	osobný	poháňa
vezmi	nevädza	nové	žiak	rak
deravý	pleť	plot	hrot	banka
tlak	soľ	zametá	dnes	svet
noc	lak	chýr	štýl	dym
ucho	domov	buk	múka	dub
sieť	šije	šiesti	ide	líce

Table 1. Selection of words for Slovak speech audiometry according to [2]

3 Diagnostic Rhyme Test – DRT

In the Diagnostic Rhyme Test, introduced for English by Fairbanks in 1958, a file of autonomous words is used to test comprehensibility of initial consonants [3,5,6]. It uses monosyllabic words that are constructed from a consonant-vowel-consonant sound sequence. In the DRT words are arranged in ninety-six rhyming pairs which differ only in their initial consonants differing by only one distinctive feature. Listeners are shown a word pair, then asked to identify which word is presented by the talker (reproduced). Carrier sentences are not used.

The word pairs have been chosen so that six phonological features could be tested. One of the words in a pair starts with a consonant characterised by certain feature, and

the second one by a consonant with a contrastive feature. The phonologically distinctive features tested in the DRT are summarised in Table 2.

4 Diagnostic Rhyme Test for Slovak Language

During the development of the DRT for Slovak language, several problems had to be resolved.

- a) Consonant pairs with distinctive phonological features had to be defined for Slovak language. We accepted the categorization given by Pauliny [7] and Kráľ and Sabol [8].

Feature	Distinctions (Phonological contrasts)	List of consonant pairs with contrast features
Voicing – Vc	voiced Vc – voiceless Vc ^o	b-p, v-f, d-t, dz-c, z-s, d'-f, dž-č, ž-š, g-k, h-ch.
Nasality – N	nasal N – non-nasal (oral) N ^o	m-v, n-l (Lt), ň-ŕ (Lt)*, n-r (Lt ^o), ň-j (Lt ^o)*. m-b (V ^o)* n-d (V ^o), ň-d' (V ^o)*
Sustentation – O	occlusive O – non-occlusive O ^o	p-f, b-v, c-s, dz-z, č-š, dž-ž, k-ch, g-h.
Sibilantion – S	sibilant S – non-sibilant S ^o	c-t, dz-d, č-f, dž-d'. A/ A ^o : c-p, dz-b, s-f, z-v** A ^o /A: č-k, dž-g, š-ch, ž-h**
Graveness – A	acute A – grave A ^o	m-n p-t, b-d, f-c, v-dz v-l (Lt), v-r(Lt ^o).
Compactness – D	diffuse (non-compact) D – non-diffuse (compact) D ^o	n-ň, l-l', r-j p-k, b-g, f-ch, v-h, t-ť, d-d', c-č, dz-dž, s-š, z-ž,

* With nasals, the absence of the lateral (Lt) feature is phonetically irrelevant ([2], page 284), thus /n/ makes a phonologically contrasted pair both with lateral /l/, and with non-lateral /r/. Similarly, /ň/ makes a phonologically contrasted pair both with lateral /l', and with non-lateral /j/

Due to lack of words with contrast sounds for N, in accordance with the procedure used for English language simultaneous inequality of the sonority distinction can be admitted. Pairs given in the fourth and fifth lines of the listed pairs for nasality will be obtained.

** Due to lack of words with contrast sounds for S in accordance with the procedure used for English language simultaneous inequality of the acute distinction can be admitted. Pairs given in the second and third lines of the listed pairs for sibilancy will be obtained. Complete list of distinctions for consonants in Slovak language can be found in [8], pp. 274 – 289, and in [7], pp. 107 - 144.

Table 2. Characteristics of consonant pairs with phonological distinctions for the Slovak Diagnostic Rhyme Test

- b) A substantial problem is in finding a sufficient number of source words for selection. The Slovak Paronymic Dictionary [9] was used in the preliminary stage of the search. Later we decided to develop a software tool that enables browsing large text corpora and searching for words satisfying user's definition. Two corpora were used in the development – The Slovak National Corpus and our own collection of texts. All the three text sources appeared to be helpful, as every of them contained some words, that were not found in the others.
- c) In Slovak, syllables without a vowel are admissible, i.e. with syllabic-forming /r/, /l/, /t/, /ŕ/ (such as mrk, tŕk ...). We decided not to include such syllables in the test and we tried to make the DRT for monosyllabic words with /á/, /é/, /í/, /ó/, /ú/, /a/, /e/, /i/, /o/, or /u/ only. The vowel /ä/ changes to /e/ in neutral pronunciation, and thus it is not included in the test as well. Analysis of words with diphthongs (/ia/, /ie/, /iu/, /ô/) has revealed that such words are rather rare and their occurrence is not sufficient to create a DRT for words with diphthongs. Thus, our DRT has 20 lines only (two lines for each of ten Slovak vowels).
- d) The lack of words is a general problem for all vowels. This is caused by the fact that Slovak is an inflexion language and the number of monosyllabic nouns and verbs is rather low. We had been trying to select preferably nouns in the basic form (nominative case) and verbs in infinitive.
- For the sake of phonetic enrichment, in some cases a different form was used instead of infinitive at a later stage.
 - It came out soon that different grammatical cases of nouns and other parts of speech must be admitted. The use of proper nouns had to be considered as well.
 - There are critically few monosyllabic words with /é/. Thus, we have also included the forms used in Slovak spelling (bé, cé, dé, etc.). Similar violation of the definition (CV syllable structure) can be found in the English version too.
- e) Development of several different DRT sets would be very useful (in the Danish version, three sets are applied). It came out, however, that there are not enough monosyllabic words in Slovak to make even one set. This is why we had to introduce words where the first vowel differs by two distinctions (this violation of the definition can be found in the English version too).
- f) To obtain even more test words we decided to allow CCVC syllable structure where it was inevitable.

- g) To obtain considerably bigger amount of test material we had to modify the definition of the DRT and to allow bisyllabic words. Thus we defined and created a new test set – a bisyllabic DRT in Slovak (BDRT).

So the summary of the steps of our work on DRT is as follows:

- a) choose of consonant pairs with contrastive features for Slovak
- b) designing a software text-searching tool for automatic retrieval of words with defined features
- c) after some discussion excluding monosyllabic words with /r/, /l/, /ř/, /ĺ/, /ä/, /ia/, /ie/, /iu/, /ô/ in syllable nucleus from the DRT word list
- d) introduction of nouns in different grammatical cases, verbs in different forms (not only infinitive, proper nouns, and spelling alphabet words to enrich the list
- e) introduction of words with consonant pairs differing in two features
- f) allowing CCVC monosyllabic words where inevitable
- g) creating a new test set - a bisyllabic DRT in Slovak (BDRT)

Other tests, such as MRT, DMCT, SAM, etc. are beyond the scope of this paper and therefore we shall only mention them in brief. Their design for Slovak will be described in future publications.

5 Modified Rhyme Test (MRT) and Diagnostic Alliteration Test (DALT)

The modified Rhyme Test uses 50 six-word lists of rhyming or similar-sounding monosyllabic English words. Each word is constructed from a consonant-vowel-consonant sound sequence, and the six words in each list differ only in the initial or final consonant sound [5]. Listeners are shown a six-word list and then asked to identify which of the six is spoken by the talker. A carrier sentence is usually used.

The mistakes are evaluated independently for initial and final consonant position.

The DALT is a final consonant rhyming word test that is structured the same as the DRT.

Generally speaking there are the same problems in MRT and DALT design for Slovak as those with monosyllable DRT.

6 Diagnostic Medial Consonant Test (DMCT)

The DMCT employs a list of ninety-six two-syllable word pairs that differ in only the middle consonant (for example, bobble-bottle). These differences are organized in six categories, and scores in each category can be used to identify specific problems. Averaged together, the six scores provide a single measure of intelligibility. Listeners are shown a word pair, then asked to identify which word is reproduced. Carrier sentences are not used.

7 Conclusion

The design of the Diagnostic Rhyme Test in Slovak brings several problems. The biggest of them is the lack of appropriate monosyllabic words in Slovak which makes it impossible to create a complete DRT word list for Slovak. Some tricks – slight changes in the definition of the DRT which we hope will not affect its functionality – allowed us to substantially enrich the test material. Even much bigger number of test words was obtained when bisyllabic words were introduced in the test.

Our research is still in its initial phase and the effect of the bisyllable structure on the test results words should be studied before the expanded version of the DRT comes to a wider use in practice. Still we hope that the definition of the Slovak DRT is an important step towards the design of an integrated set of tests for subjective speech quality evaluation in Slovak.

Acknowledgements

The Slovak National Corpus [10] was used for research and development of the Slovak DRT.

This work was partly supported by VEGA grant No. 2/5124/25 and by the state task of research and development 2003 SP 20 028 01 03.

References

1. Petriska M., Sachia Daržágín S.: Slovenská difónová databáza pre Text-To-Speech systém Festival. In: Proceedings of the 6th International Acoustic Conference. Noise and vibration in practice, Slovenská technická univerzita, Strojnícka fakulta, Bratislava, 2001. - ISBN 80-227-1544-1, pp. 105-108.
2. Bargár Z., Kollár A.: Praktická audiometria, Osveta 1986, str. 159-160
3. Goldstein M.: Classification of Methods Used for Assessment of Text-to-Speech Systems According to the Demands Placed on the Listener. Speech Communication vol. 16, 1995: str. 225-244.
4. Cerňák M., Rusko M.: An Evaluation of Synthetic Speech Using the PESQ Measure, Proceedings of Forum Acusticum, Budapest, 2005, pp. 2725-2728, ISBN 963 8241 68 3.
5. MDRT - standard ANSI S3.2-1989
6. William D. Voiers, Diagnostic Evaluation of Speech Quality, in Speech Intelligibility and Recognition, Mones E. Hawley, ed., Dowden, Hutchinson & Ross, Inc. (Stroudsburg, PA), 1977.
7. Pauliny E.: Slovenská fonológia, SPN 1979
8. Kráľ A., Sabol J.: Fonetika a fonológia, SPN 1989, pp. 273-287
Škvareninová O.: Paronymický slovník, SPN 1999
9. The Slovak National Corpus. Bratislava: Jazykovedný ústav L. Štúra SAV 2003. Available from WWW: <http://korpus.juls.savba.sk>.

Bulgarian and English Semantic Dictionaries for the Purposes of Information Retrieval*

Max Silberztein¹ and Svetla Koeva²

¹ Université de Franche-Comté

² Bulgarian Academy of Sciences

Abstract. The paper presents broad conception for stemming - as a mutual correspondence between word-form paradigms of all literals belonging to the synonymous sets constituting a given WordNet relation. The implementation includes the association of literals from the Bulgarian and English WordNets with the corresponding super-lemmas and inflection types in Semantic Dictionaries. The Semantic Dictionaries have been constructed with the NooJ linguistic development environment. NooJ dictionaries contain indistinctly simple or compound words thanks to an inflection system that can process both simple and compound words' inflectional morphology in a unified way. Moreover, NooJ can provide linking of all word forms associated with an equivalent super-lemma. When the super-lemma corresponds to a given semantic relation between words, a semantic stemming can be accomplished.

1 Introduction

The goals of the presented investigation are directed to the implementation of natural language semantic relations in Information retrieval systems. This involves broad conception for stemming - as a mutual correspondence between word-form paradigms of all literals belonging to the synonymous sets constituting a given WordNet relation. The “semantic” stemming requires working out of the following tasks:

- to provide complete formalization of the inflection of simple and compound literals included in the Bulgarian and English WordNet structures;
- to create specialized Semantic Dictionaries for Bulgarian and English based on WordNet semantic relations.

Both tasks have been implemented with the NooJ linguistic development environment [7].

* The reported work is part of the Joint research RILA project *Information retrieval based on semantic relations* between LASELDI, Université de Franche-Comté, and Department of Computational Linguistics, IBL, Bulgarian Academy of Sciences.

2 NooJ Dictionaries

NooJ dictionaries contain indistinctly simple or compound words thanks to an inflection system that can process both simple and compound words' inflectional morphology in a unified way.

2.1 Merging Simple and Compound Words

For instance, the two following lexical entries:

academic program,N+FLEX=APPLE

window,N+FLEX=APPLE

inflect the same way (they take an 's' in the plural). Therefore, they are both associated with the same inflectional class: APPLE. The class APPLE is defined by the following expression:

APPLE = $\langle E \rangle$ /singular + s/plural;

that states that if one adds nothing to the lexical entry ($\langle E \rangle$ is the empty string), one gets the singular form ("singular"); if one adds an "s" to the end of the lexical entry, one gets the plural form ("plural"). NooJ's inflectional engine is equivalent to a stack automaton. It uses a dozen default commands that operate on the suffix of each lexical entry:

- $\langle B \rangle$: equivalent to the keyboard key "Backspace"
- $\langle D \rangle$: Duplicate current character
- $\langle E \rangle$: Empty string
- $\langle L \rangle$: equivalent to the keyboard key "Left arrow"
- $\langle N \rangle$: go to end of Next word form
- $\langle P \rangle$: go to end of Previous word form
- $\langle R \rangle$: equivalent to the keyboard key "Right arrow"
- $\langle S \rangle$: equivalent to the keyboard "delete" key

Users can override these commands, and add their own.

NooJ is capable of inflecting compounds. For instance, the class "ACTOFGOD" is defined by the following expression:

ACTOFGOD = $\langle E \rangle$ /singular + $\langle P \rangle \langle W \rangle$ s/plural;

The operator $\langle PW \rangle$ stands for: "go to the end of the first component of the lexical entry" Note that the following three entries are associated with this class, even though their length is different:

bag of tricks,N+FLX=ACTOFGOD

balance of payment deficit,N+FLX=ACTOFGOD

member of the opposite sex,N+FLX=ACTOFGOD

In the same manner, even though the lexical entries: *blank piece of paper*, *last line of defence*, *family history of cancer*, *sexual harassment in the work place*, *etc.* have different lengths, they can be associated with a unique inflectional class because the inflection is carried by the same component (the second one). Agreements between components of a compound can be described as well. For instance, the following inflectional expression formalizes the agreement between

the two components of compounds such as journeyman carpenter:

$\langle E \rangle / \text{singular} + s \langle P \rangle \langle B \rangle 2 \text{en} / \text{plural}$

To get the plural form, add an “s” to the end of the compound, then go back to the previous ($\langle P \rangle$) component, delete the two last characters ($\langle B \rangle 2$), and add the suffix “en”. In conclusion, NooJ can process simple and compound words’ inflection completely automatically. This has allowed us to unify the description of simple and compound words in WordNet type dictionaries.

2.2 Dictionary and Inflection

NooJ dictionaries are directly compiled into a Finite-State Transducer, in which all the relevant inflectional paradigms are stored as well. This characteristic is very important for our project: it gives NooJ the ability to perform morphological operations during parse time. NooJ can link any inflected form, to any other inflected form represented in its transducer. Thus, NooJ can perform complex transformations within texts. For instance, it is now possible to replace a certain conjugated verb with its past participle form, and vice-versa, within a particular text:

John eats the apple \leftrightarrow the apple is eaten by John

Using this new functionality will enhance several current NLP applications, such as automatic translation and information retrieval applications.

2.3 Property Definition and Types

NooJ dictionaries can be displayed either in list (“free”) form, or in table (“typed”) form. This requires that features of the dictionary be typed, so that all the values of a common property can be regrouped in one column. We also need to state the number of relevant properties for each category of word (e.g. Tense for Verbs, Number for Nouns, etc.), in order to distinguish absent default values, from irrelevant ones. This is done via a “Property Definition” file that contains rules such as:

NDistribution = Hum + Conc + Abst ;

NGender = m + f ;

NNumber = s + p ;

...

VTense = Present + Futur +... ;

VPers = 1 + 2 + 3 ;

VNumber = s + p ;

...

The next figure displays a NooJ dictionary in table form: Note that all features in a NooJ dictionary do not have to be typed; if NooJ does not know the type of a feature, it will simply display it as a column header, and enter the “+” (if the feature is present) and “-” (if absent) values accordingly. A given property may be associated with more than one category (e.g. Number is relevant both for nouns and verbs). But NooJ checks that one feature (e.g. “+p”) does not

Entry	Category	CR	Genre	Nombre	Pers	Sem	Temps
a	N		m	p		-	
a	N		m	s		-	
avoir	V		-	s	3		P
à	PREP						
abaissier	V		-	-	-		W
abandon	N		m	s		-	
abandonner	V		-	s	3		P
abandonner	V		-	s	2		Y
abandonner	V		-	s	1		P
abandonner	V		-	s	1		S
abandonner	V		-	s	3		S
abandonner	V		m	s	-		K
abandonné	A		m	s			
abandonné	N		m	s		-	
abandonner	V		f	s	-		K

Fig. 1. Table view for a NooJ dictionary

correspond to more than one property (e.g. “Gender” and “Tense”). NooJ distinguishes default properties from irrelevant ones. Moreover, associating NooJ lexical features with typed properties opens up possibilities for implementing unification mechanisms in the future.

2.4 Bulgarian Grammatical Dictionary

The grammatical information included in the Bulgarian Grammatical Dictionary (BGD) is divided into three types [2]: category information that describes lemmas and indicates the words clustering into grammatical classes (Noun, Verb, Adjective, Pronoun, Numeral, and Other); paradigmatic information that also characterizes lemmas and shows the grouping of words into grammatical subclasses, i. e. - Personal, Transitive, Perfective for verbs, Common, Proper for nouns, etc.; and grammatical information that determines the formation of word forms and shows the classification of words into grammatical types according to their inflection, conjugation, sound and accent alternations, etc. The BGD is a list of lemmas where each entry is associated with a label [4]. The label itself represents the grammatical class and subclass to which the respective lemma belongs and contains a unique number that shows the grammatical type. All words in the language that belong to the same grammatical class, subclass and have an identical set of endings and sound / stress alternations are associated

with one and the same label. Each label is connected with the corresponding formal description of endings and alternations.

The inflectional engine used is equivalent to a stack automaton. Despite the existence of some differences in the format, the BGD represents a kind of DELAS dictionary, and it is compiled into a Finite-State Transducer. The BGD which already contains over 85000 lemmas has a parallel version in NooJ format where dictionary labels are transliterated and formal descriptions are transformed into the NooJ formal apparatus.

3 WordNet

The global WordNet [1]; [6] is an extensive network of synonymous sets and the semantic relations existing between them, enabling cross-language references between equivalent sets of words in different languages [9]. The Bulgarian wordnet (BulNet) has been initially developed in the framework of the project *BalkanNet – a multilingual semantic network for the Balkan Languages* which has been aimed at the creation of a semantic and lexical network of the Balkan languages [8].

	Bg N	Bg V	Bg Adj	Bg Adv	Bg Total
Synsets	15 508	4 421	4 027	449	24 405
Literals	27 772	15 701	7 017	1 094	51 584
Graphic-words	22 381	8 860	5 040	817	37 098
ILR	25 577	11 143	6 931	815	44 466

Table 1. The distribution of Bulgarian synsets into parts of speech

	En N	En V	En Adj	En Adv	En Total
Synsets	79 689	13 508	18563	3 664	115 424
Literals	141 691	24 632	31 016	5 808	203 147
Graphic-words	114 649	11 306	21 437	4 660	152 052
ILR	129 983	36 457	34 880	3 628	204 948

Table 2. The distribution of synsets in English Wordnet 2.0

The Bulgarian WordNet [3] models nouns, verbs, adjectives, and (occasionally) adverbs, and contains already 24405 word senses (towards 1.09.2005), where 51584 literals have been included (the ratio is 2,11). The distribution of Bulgarian and English synsets across different parts of speech is shown in Tables 1 and 2.

3.1 Wordnet Structure

Every synset encodes the equivalence relation between several literals (at least one has to be present), having a unique meaning (specified in the SENSE tag value), belonging to one and the same part of speech (specified in the POS tag value), and expressing the same lexical meaning (defined in the DEF tag value). Each synset is related to the corresponding synset in the English Wordnet2.0 via its identification number ID. There has to be at least one language-internal relation (there could be more) between a synset and another synset in the monolingual data base. There could also be several optional tags encoding usage, some stylistic, morphological or syntactical features, etc.

3.2 Inflecting Wordnet

Literals included in the wordnet structure can be either simple words or compounds i.e. the English synset *car:2, railcar:1, railway car:1, railroad car:1* with the definition “it a wheeled vehicle adapted to the rails of railroad” corresponds to Bulgarian synset *vagon:1*; the Bulgarian synset *hol:1 salon:1 balna zala:1* with the definition “the large room of a manor or castle” corresponds to the English one *manor hall:1, hall:5*. Comparing to lemmas compounds have their own inflective rules. In order to merge the language data existing in BulNet and BGD it was decided to assign an additional grammatical note to each literal thus linking it with the BGD lemma’s label [5]. All labels for BGD entry forms that are found in the BulNet have been entered as values of the LNOTE (lexical note) grammatical tag in the XML format. Most of the literals which were not recognized are either specialized terms that have no place in a grammatical dictionary of the common lexis (often written in Latin) or compounds. The contradictory cases where two or more labels were associated with one and the same literal are solved manually. The classification of compounds according to different inflectional types is under development.

3.3 WordNet Relations

The major part of the relations encoded in the Bulgarian WordNet is semantic relations. There are also some morpho-semantic relations, some morphological (derivational) relations, and some extralinguistic ones. WordNet relations of equivalence, inheritance, similarity, and thematic domains affiliations are of interest to the Information retrieval purposes. Those are: synonymy; hypernymy, meronymy, similar to, verb group, also see, and category domain.

Synonymy

Synonymy is a semantic relation of equivalence (reflexive, symmetric, and transitive) between literals belonging to one and the same part of speech. In Princeton WordNet the substitution criteria for synonymy is mainly adopted: “two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value” [6]. Thus the relation implies that one synonym may substitute another (synonym) in a context and vice versa.

The consequences from such an approach are at least two - not only the exact synonymy is included in the data base (a context is not every context). Second, it is easy to find contexts in which words are interchangeable, but still denoting different concepts (for example hypernyms and hyponyms), and there are many words which have similar meanings and by definition they are synonyms but are hardly interchangeable in any context due to different reasons - syntactic, stylistic, etc. (for example an obsolete and a common word).

Hypernymy

Hypernymy (Hyponymy) is an inverse, asymmetric, and transitive relation between synsets, which correspond to the notion of class-inclusion between synsets belonging to one and the same part of speech. The relation implies that the hypernym may substitute for the hyponym in a context but not the other way round. Hypernymy is a transitive relation: e.g. being a kind of *cvete* (*flower*), *roza* (*rose*) has inherited not only all semantic features of *cvete* (*flower*), but also those of its superordinates: *rastenie* (*plant*), *zhiv organizam* (*organism*), etc.

Meronymy

Meronymy (Holonymy) is inverse, asymmetric, and transitive relation which link synsets denoting wholes with those denoting their parts. The Part meronymy typically relates components to their wholes. In BulNet we restrict the Part relation to the components that are topologically included one in the other with physical attachment: *book* is a part of *library*, *library* is a part of *building*, **book* is not a part of *building*, only *library* - *building* relation is encoded as Part meronymy. The Member meronymy is a relation between sets and their members i.e. *football player* - *football team* - *football league*, *football player* is a member of a *football team*, *football team* is a member of *football league*, as well as a *football player* is a member of *football league*. The Portion meronymy is between wholes and their portions i.e. *crumb of bread* - *slice of bread* - *loaf of bread*; *crumb of bread* is a portion of a *slice of bread*, *slice of bread* is a portion of *loaf of bread*, as well as *crumb of bread* is a portion of *loaf of bread*.

Relations of equivalence

Similar to is a symmetric relation between similar adjectival synsets. i.e. the synset *nice:1* with a gloss "pleasant or pleasing or agreeable in nature or appearance" is in a Similar to relation with the synset *good:7* defined "as agreeable or pleasing". Verb group is a symmetric relation between semantically related verb synsets: the synset *wash:9*, *wash out:4*, *wash off:1*, *wash away:2* with meaning "remove by the application of water or other liquid and soap or some other cleaning agent" is in a Verb group relation with the synset *wash:1*, *rinse:2* with a definition "clean with some chemical process". Also see is a symmetric relation between synsets - verbs or adjectives, that are close in meaning i.e. *beautiful:1* defined as "delighting the senses or exciting intellectual or emotional admiration" is in a relation Also see with *attractive:1* defined as "pleasing to the eye or mind especially through beauty or charm" **Category domain**

Category domain is an asymmetric extralinguistic relation between synsets denoting a concept and the sphere of knowledge it belongs to i.e. the synset *alibi:1* with the definition "a defence by an accused person purporting to show that he

or she could not have committed the crime in question” is in a Category domain relation with the synset *law:2, jurisprudence:2* defined as “the collection of rules imposed by authority”.

4 Semantic Stemming

4.1 Multi-fields Dictionaries

The number of fields of NooJ dictionaries is no limited to one. NooJ dictionaries can contain entries associated with a “super-lemma” that can be an orthographical variant, the translation in another language, a synonymous entry or an hyperonym. For instance, consider the following lexical entries:

U.N.,United Nations,N+Org

czar,tsar,N+FLX=Pen

The first entry (U.N.) is associated with super-lemma “United Nations”; it does not inflect. This entry is similar to a DELACF entry. The second entry (czar) is associated with super-lemma “tsar”; it inflects according to the paradigm “Pen” (i.e. takes an ‘s’ in the plural). Being able to associate words with super-lemmas, i.e. words that do not necessarily correspond to their inflectional lemma (“czar” is czars’s lemma, not “tsar” opens up a new range of applications.

4.2 Semantic Dictionaries

The Semantic Dictionaries are designed using the WordNet structures (enumerated relations), on the one hand, and the respective inflectional dictionaries, on the other hand. There are two types of relations in the wordnet - symmetric (as synonymy) and asymmetric (as hypernymy) which determine the two approaches with super-lemma association. For the symmetric relations the super-lemma in Semantic Dictionaries is considered as the IDentification number of a given synonymous set.

author,ENG20-10090311-n,N+FLX=APPLE

writer,ENG20-10090311-n,N+FLX=APPLE

For the asymmetric relations the formalization is in the direction from the more concrete to more global concept (thus the super-lemma is the ID of the highest synonymous set in the hierarchy), but the other way is also possible. The main applications of the Semantic Dictionaries are directed towards Information retrieval by means of: semantic equivalence with synonymy dictionaries, semantic specification with hyperonymy and meronymy dictionaries, Information retrieval by means of similarity and thematic domains affiliations. The Semantic Dictionaries provide retrieve of all word-forms of all literals belonging to the synonymous sets constituting a given WordNet relation (Figure 2).

5 Conclusions and Future Directions

The Multi-fields dictionaries can provide Information retrieval by means of semantic equivalence with synonymy dictionaries, by means of semantic specification with hyperonymy and meronymy dictionaries, by means of similarity

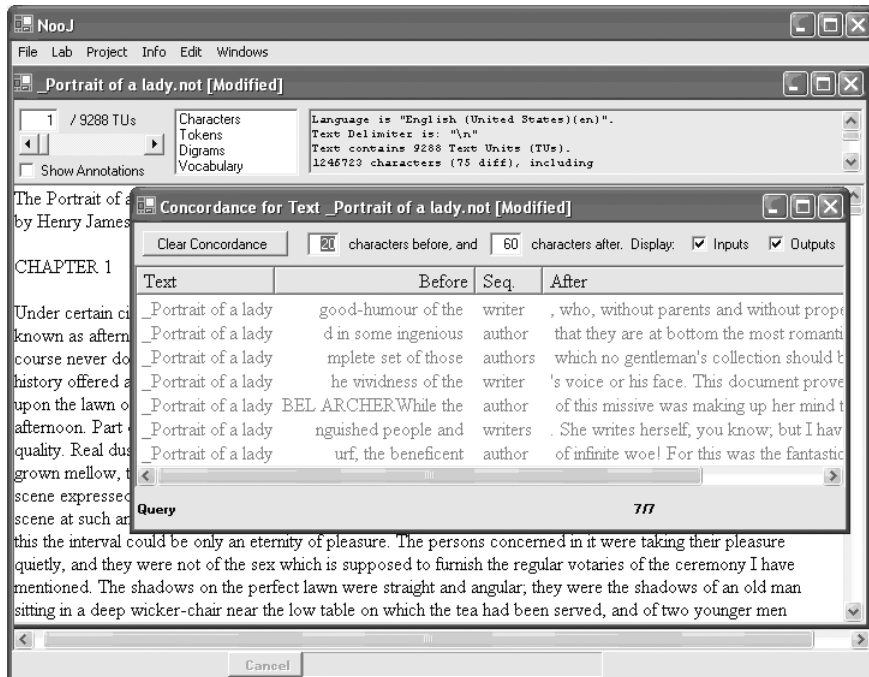


Fig. 2. Semantic stemming

relations, and by means of thematic domains affiliations. Future work is directed to the extensions and enhancements of the Semantic Dictionaries:

- Extension of the dictionaries coverage;
- Addition of other semantic relations;
- Inclusion of additional information to the entries.
- Integration of multilingual semantic extraction with NooJ using the Inter-Lingual-Index relation.

References

1. Fellbaum, C. (ed.). WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press, 1998.
2. Koeva S., Bulgarian Grammatical Dictionary. Organization of the Language Data, Bulgarian language, 1998, vol. 6: 49-58.
3. Koeva S., T. Tinchev and S. Mihov Bulgarian Wordnet-Structure and Validation in: Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004: 61-78.
4. S. Koeva Modern Language Technologies – Applications and Perspectives, in: Lows of/for Language, Hejzal, Sofia, 2004, 111- 157
5. S. Koeva Validating Bulgarian WordNet Using Grammatical Information in: Proceedings from Joint International Conference of the Association for Literary and

- Linguistic Computing and the Association for Computers and the Humanities, Gteborg University, 2004, 80-82.
6. Miller G. A. Introduction to WordNet: An On-Line Lexical Database. In "International Journal of Lexicography", Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K.J. Vol. 3, No. 4, 1990, 235-244.
 7. Silberztein M. NooJ: an Oriented Object Approach. In INTEX pour la Linguistique et le Traitement Automatique des Langues. Les Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comte: Besancon, 2004.
 8. Stamou S., K. Ofazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou, BALKANET: A Multilingual Semantic Network for the Balkan Languages, Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002, 12-14.
 9. Vossen P. (ed.) EuroWordNet: A Multilingual Database with Lexical Semantic Networks for European Languages. Kluwer Academic Publishers, Dordrecht. 1999.

Slavic Text Taggers Project

Danko Šipka

Department of Languages and Literatures
Arizona State University
PO Box 870202
Tempe, AZ 85287-0202
danko.sipka@asu.edu

Abstract. Responding to national needs for more efficient instruction of less commonly taught languages (LCTLs), utilizing its technological potential and social environment, Arizona State University Department of Languages and Literatures Slavic Section in cooperation with Critical Languages Institute is developing Bosnian/Croatian/Serbian (BCS), Polish, and Russian text taggers. The taggers are available at <http://www.asusilc.net/cgi-bin/newtepajgu.pl> They allow the user to paste in a text, copied from an on-line newspaper or acquired in another manner, and have it tagged with English glosses and equipped with the option of displaying the inflection of each wordform in the text. The present paper summarizes achievements of this project hitherto, identifies its major problem areas, and outlines its envisaged development hence.

1 Introduction

One of the hallmarks of the early twenty-first century is the shifting locus of educational training from the traditional fixed classroom to the Internet and real-life immersion. Language instruction is not an exception in this respect. There is a demand for on-line and immersion language learning which will enable the student to act as an independent performer of various tasks in the target language (e.g., understanding authentic materials, engaging in on-line and face-to-face interactions, capturing cultural differences, etc.) In contrast to these needs, most available textbooks remain limited to traditional in-class instruction where rote learning detached from real life and learners as objects rather than subjects is perpetuated. Similarly, courses of foreign languages as a rule do not incorporate immersion or task-based e-learning. In effect, learning outcomes are reduced to passive knowledge of grammatical structures and vocabulary, rather than to linguistic and cultural literacy in multiple genres of the target language. This unfortunate state of the affairs is a frequent subject of concern in the academe (see for example Brecht, R. D. and W. P. Rivers, 2002[1]). Despite the widespread I-just-want-to-speak student attitude, the most valuable professional skill to be acquired in Slavic language classrooms is the ability to understand authentic written and spoken texts in the target language. It is therefore imperative that authentic materials are included in the curriculum as early as possible. However, attempting to address this imperative leads to the following dilemma. On the one hand, students with limited command of the vocabulary cannot be expected to process “raw” authentic texts as constant dictionary lookup would be overly time consuming and frustrating. On the other hand, instructors do not command sufficient financial and temporal resources to

manually gloss these texts with English equivalents. Being able to confront authentic materials early in the instructional process bears a two-fold importance. First, the student acquires the sense of achievement, the feeling that she/he can utilize the language in a sensible manner, which in and of itself wields a beneficial effect on the cognitive and affective factors in language acquisition. Secondly, the early inclusion of authentic materials creates a solid cognitive basis for later phases of the process in which authentic texts become ineluctable. The Slavic text tagger projects are an attempt to resolve the dilemma between pressing instructional needs and limited resources. Providing tools which facilitate the comprehension of written texts (newspaper articles, short stories, public and corporate web sites, etc.), meets the aforementioned instructional needs with little or no resources required by the instructors. More information about the general framework of the project can be found in Šipka (2004) [2].

2 Design of the Taggers

Taggers for Bosnian/Croatian/Serbian (BCS), Polish and Russian are generally available and located at <http://www.asusilc.net/cgi-bin/newtepajgu.pl>. The taggers are a part of a wider project titled Learner-centered Task-oriented Language Instruction, presented at <http://www.asusilc.net/lctli>. Each of the three taggers accept electronic texts in various formats (UTF-8, cp-1250, cp-1251, etc.), either typed into the text window or copied from an internet page and pasted into the window.

The taggers return the text tagged with the English glosses as can be seen at

- <http://www.asusilc.net/lctli/exbcs.htm> (BCS),
- <http://www.asusilc.net/lctli/expol.htm> (Polish),
- <http://www.asusilc.net/lctli/exrus.htm> (Russian).

By clicking at any of the underlined word forms in the text, the user can get their respective English gloss. Thus, clicking at the BCS word *sahrana* will yield the following gloss: *sahrana, e f [I] funeral n, sepulture n, interment n, inhumation*. Pressing the I (inflection) button will expand the word *sahrana* in all its forms (Nominative Singular *sahrana*, Genitive Singular *sahrane*, etc.). The design of the taggers is intended to facilitate the cumbersome and time-consuming process of looking up the English equivalents and determining the morphological category of the wordforms in texts.

The following technologies were utilized to implement the taggers. The entire knowledge base, with dictionary forms, inflectional forms and the English equivalents is stored into a relational database in MySQL. Perl scripts with HTML forms as their GUI are used to query the database and tag the text, while the resulting tagged text is implemented in a form of HTML, DHTML with a limited use of Java Script. Central to this design was the idea that all operations are performed serverside, which stipulates minimum requirements on the part of the user, coupled with the transferability of the resulting HTML page with the tagged text.

The taggers accept the input text in various code pages (Unicode, Windows, ISO, etc.) and the output text is always rendered in Unicode. It is important to note that the BCS tagger can handle both Latin and Cyrillic script.

The resulting tagged page can be saved and post-edited. An example of a post-edited text created using a previous version of the BCS tagger and incorporated into the BCS 201-202 course can be found at:

<http://www.public.asu.edu/~dsipka/bcs202a1.html>.

The background on methodology and technology in on-line delivery of BCS can be found in Šipka (2003)[3]

Presently, the most advanced is the BCS tagger (the access is available at: <http://www.asusilc.net/cgi-bin/newtepajgu.pl?lang=sr>), which covers approximately 95% of an average newspaper text.

The Polish tagger (<http://www.asusilc.net/cgi-bin/newtepajgu.pl?lang=pl>) covers some 85% of an average newspaper text, while the Russian tagger (<http://www.asusilc.net/cgi-bin/newtepajgu.pl?lang=ru>) covers over 60%.

In the present version of the taggers, the text is tagged automatically, and all possible tags are listed. Thus if we have the BCS form *je*, which can be either accusative or genitive singular of the personal pronoun *ona* 'she' and the third person singular of the verb *biti* 'to be', both these glosses are listed. This manner of tagging replaced the previous version of the tagger, in which the text could be tagged automatically or interactively. In the former case the most frequent solution was deployed in case of ambiguous forms (e.g., in the example above, the verbal meaning would be selected). The interactive manner of tagging was prompting the user to resolve all cases of ambiguous forms (e.g., to tell if the form *je* found in the text belongs to the verb *biti* 'to be' or to the pronoun *ona* 'she'). However, it turned out in the course of longitudinal testing that manual tagging is overly time-consuming and annoying while automatic selection of only one tag bears a risk of being inaccurate. The present solution allows swift tagging and concurrently provides the user with the opportunity of selecting the right equivalent while using the tagged text.

An average newspaper text of 350 word forms is tagged in six seconds.

3 Major Challenges

The major challenge currently addressed within the Slavic text taggers project is that of appending and amending the knowledge bases. At present, the team members are working on formatted text files, which are then transferred into the databases using PERL scripts.

Given that this solution does not allow computational lexicographers and grammarians to immediately introduce a change and see its result, the development of a knowledge base authoring tool is underway. The completion of this tool is envisaged for the last quarter of 2005 and in its completed form it will allow addition of new items as well as modification of the inflectional forms and the English equivalents. Central to the development of this resource is to allow concurrent use of various updating

techniques – automatic (using the available resources) and manual, modification of one peculiar form of inflection and assigning a completely different paradigm, etc.

Two major techniques are being deployed in appending and amending the knowledge bases. First, the databases are filtered using available monolingual and bilingual lexical lists or dictionaries and the missing entries are added. Second, the tagger itself is used as an important tool of testing the comprehensiveness and accurateness of the database. As can be seen in the Appendix, if a word form is not lemmatized, it will not be underlined, if it is lemmatized yet its English equivalent is missing, it will not be in the bold type. Those forms which are lemmatized and with the English equivalent will be bold and underlined. Testing the taggers with the texts from the most popular newspapers is used to train the knowledge bases.

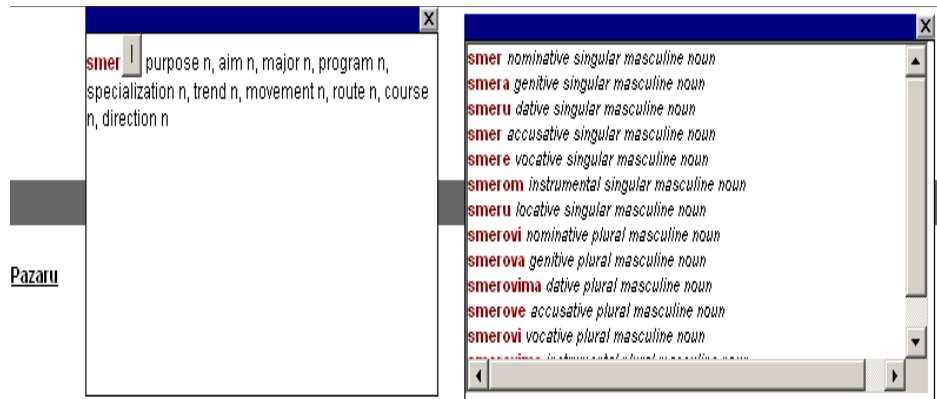
Inasmuch as the three languages exhibit a varied degree of development, the concrete tasks performed on the databases expose considerable differences. In the case of BCS, some low-frequency lexemes and irregular forms are still being added to the knowledge base. At present, the process is geared toward appending the knowledge base with the items from a 230,000-entry word list, which served as the bases for Šipka (2002). Even upon full inclusion of the aforementioned list, there will still be some 1-2% unrecognized forms, owing primarily to the creativity of the journalists and other authors. For example, the negative particle *ne-* can be added to practically any noun or adjective, as in *neodgovarajući* ‘unfitting, unsuitable’ in the Appendix below. This problem will be tackled by tagging parts of the word form if the tagging of the form in its entirety fails. Obviously, proper names will not be tagged at all.

The problems being solved in the Polish and Russian tagger are quite different than those in the BCS tagger. Here, the goal is to form a solid foundation of approximately 80,000 lexemes (i.e., canonic forms) with their corresponding equivalents and inflections. The synergy of tagger training of the knowledge base and its appending using the available electronic resources is deployed here in a manner akin to the BCS tagger project.

4 Further Development

The goal of the project is to reach 95% converge for the Polish and Russian taggers and 98% coverage for the BCS tagger by October 2006. All activities in the intervening period will be subordinated to that overarching goal. A further step will include the expansion of the resources to other languages beyond the Slavic linguistic realm. Finally, a number of spin-off projects is envisaged. Most notably, a tool will be created for semi-automatic collections of neologisms. A net-bot will be implemented to randomly search and tag a predetermined set of web pages and extract unattested lexemes along with their contexts for a subsequent human lexicographic treatment.

Appendix



ipi u međuljudskim odnosima koji prete da ugroze normalan početak nove školske godine . Posle sednice Nastavničkog veća , početkom ove amoupravi , prosvetnim vlastima , republickom ministru prosvete Slobodanu Vuksanoviću i ministru za ljudska i manjinska prava Rasimu Ljajiću o imenovanja novog školskog odbora .

ene Huseina Hanića koji je više od dve decenije bio direktor . Posle toga , Ministarstvo prosvete za v. d. direktora imenovala je profesora te šljiljski nadzor . Kontrolu su obavljali republicki prosvetni inspektori Svetlana Filipović i Rodoljub Petrović i opštinski inspektor Šulsuma Hodžić . Or pozitivnim zakonskim propisima , neodgovarajuća stručna zastupljenost u nastavi , nekompletni dosjeji zaposlenih , prevođenje vanrednih u re e broja učenika po odeljenjima bez pravnog osnova , nastavi su pratili učenici koji nisu upisani u školu , matične knjige su neuredno vođene ,

elo je vanredne mere . Za v. d. direktora imenovana je Radomirka Rajović , profesor biologije i imenovan je novi Školski odbor na godinu dana . nedavno , stigla je nova odluka imenovan je novi Školski odbor u kojem su pet članova Srbi , četiri Bošnjaci .

utim imenovanim školskim odborom " uutili su donje nomenim institucijama i pojedincima i zatražili " da se na demokratski način hitno ob-

Appendix 1: Screen caption of the BCS tagger

References

1. Brecht, R. D. and W. P. Rivers. "The Language Crisis in the United States: Language, National Security and the Federal Role" In: Baker, Stephen J (ed.) *Language Policy: Lessons from Global Models*, Monterey: MIIS, (2002), 76–90
2. Šipka, D. "On-line Delivery for Serbo-Croatian (Bosniac, Croatian, Serbian)", *Journal of the NCOLCTL*, Volume 1, (2003), 95–118
3. Šipka, D. "Content-centered resources for the West Balkans" In R. Jourdenais, & S. Springer (Eds.). *Content, tasks and projects in the language classroom: 2004 conference proceedings*. Monterey, CA: Monterey Institute of International Studies, (2004), 123-129
4. Šipka, D. *Enigmatski Glosar. Dio 1. Osnovni oblici*, Beograd: Alma (2002), 1–1315

Multi-Words Named Entity Recognition in Polish Texts

Dominika Urbańska¹ and Agnieszka Mykowiecka^{1,2}

¹ Polish-Japanese Institute of Information Technology,
Warsaw, Koszykowa 86, Poland

`dominika.urbanska@pjwstk.edu.pl`

² Institute of Computer Science, Polish Academy of Sciences,
Warsaw, Ordonia 21, Poland

`Agnieszka.Mykowiecka@ipipan.waw.pl`

Abstract. This article describes one of the first attempts to implement recognition methods of personal and institutional names in informal Polish texts. This application uses a hybrid method based on searching dictionaries enriched with some heuristic methods. The recognition process is divided into four stages: marking potential NE, marking the “key words”, creating the full Name Entity and finally creating the base form of all the recognized Name Entities. The results are placed into external text files and can be used as a starting point for other applications.

1 Introduction

Huge amounts of information that are widely available via web browsers, causes a growing demand for applications that are able to automatically process data expressed in natural languages. However, such analysis needs a collection of linguistic tools and resources. One of many problems which has to be solved concern named entities. Many types of texts, especially newspaper texts, include a lot of proper names (Named Entities, NE) which are generally not present in existing lexicons. The first idea to do NE recognition is to create special gazetteers of, for instance, names of people, places and organizations. But this solution is not sufficient. Although some Named Entities come from well defined data collections (e.g. country names), more frequently they come from open, non restricted sets. Such Named Entities can appear and disappear in everyday usage of natural languages. For examples, it is possible to construct lists of known organizations, but new companies are created daily, and their names need to be recognized as well. A second example are surnames of currently popular people. That is why during the construction of an NER application it is not recommended to base it only on dynamically updated lists. Another negative aspect of using such lists is that exploration of huge data sets is time consuming. The next problem is that proper names can be complex entities, consisting of several words and these words may occur in texts in various forms so the lexicons should contain all of them. And even if we want to have such NE lists, it will be good to create them automatically.

The main difficulty in NER task is the fact that many proper names are multi-words strings and it is not easy to recognize their limits. In Polish, proper names are generally capitalized, but they can contain words beginning with lowercase letters – conjunctions and prepositions. What is more, frequently two proper names can appear one after another so some methods of splitting them are necessary.

The interesting solutions to NER problem were found in statistical and AI methods like decision trees, Hidden Markow Models, the maximum entropy model or neural networks, e.g. [3]). But it is still quite common to use “hand-made rules”, which are customized for specific tasks (e.g. [2] and [1]). Their preparation is time consuming and they are not easily portable but they usually give good results.

For many applications, such as Information Extraction (IE), Automatic Text Summarizations or Question Answering (QA), the Multi-words Named Entity Recognition (NER) is a crucial and preliminary task. How important this task is, was recognized already at Message Understanding Conference (MUC 6, <http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html>), where NER was defined as an important separate task. After this conference, it was stated that the worst and most difficult entities to classify are the names of companies and more importantly, people.

In Polish computational linguistics the Name Entity Recognition is still not a well explored subject, although some important works in this field exist, e.g. (Piskorski, 2004). This article describes one of the first attempts to implement an application for person and institution names recognition in Polish texts based on a hybrid approach. The program is using a number of dictionaries, containing lists of selected types of proper names (for example first names) as well as some heuristic methods, which allow the program to recognize complex names. The lack of annotated Polish corpora made it impossible to use a pure statistical approach.

2 Application Characteristics

Our goal is Name Entity Recognition for the most difficult name type – we recognize names, surnames and complete person referring phrases as well as institution names. One can search for one of the NE types or all of enumerated types at the same time. Besides of finding the NE limits, the application finds the base form of the phrase. This feature is very important for a highly inflectional language as Polish, as knowing the base form of a name helps us to recognize names referring to the same object and to create a standardized list of recognized names. The application produces results being two text files – one contains text with tags delimiting all recognized Name Entities and the second contains only Name Entities together with their base forms.

Usually the Name Entity Recognition task is divided into few separate, smaller sub-tasks. In our approach, we firstly mark all capitalized words (they are potentially Name Entities). To reduce the processing time, we begin with

using small names dictionary (names are divided into two groups: the basic dictionary contains the list of most common names, the expanded one consists of all formally registered Polish names). Proper NER process is divided into two stages. First, the “key words” are marked. For searching person names, the “key words” are all names from the dictionary. For institutions, we have chosen several nouns like *bank*, *school* or *university*. This stage can be called “dictionary search stage”, because both names and institution types are placed in dictionaries. The last processing stage is using the designed algorithm (precisely describes in next section) to recognize the limits of the Name Entities. The general idea consists in adding the capitalized words from the key word neighborhood but some special situations are distinguished and dealt with.

Fig. 1 presents schematically all described processing stages while Fig. 2 show the application window.

I – inputting the data	<i>Piotr Koter studiuje na Uniwersytecie Warszawskim.</i>
II – marking potential NE	<u>Piotr Koter</u> studiuje na <u>Uniwersytecie Warszawskim.</u>
III – marking the keywords	<u>Piotr</u> <u>Koter</u> studiuje na <u>Uniwersytecie</u> <u>Warszawskim.</u>
IV – creating NE	<u>Piotr Koter</u> studiuje na <u>Uniwersytecie Warszawskim.</u>

Fig. 1. Processing stages



Fig. 2. Program window

The second challenge for the application is creation of the base form of all recognized Name Entities. The dictionaries contain all inflectional word forms, so the problem concern mainly those words which are not included in them.

The application tries to guess their correct form on the basis of some stored endings and morphological features of those NE parts which are stored in the dictionaries.

2.1 Person Names Recognition

The application recognizes single names and phrases containing names, surnames, initials and titles. The first and middle names recognition is easy as it is done using a long list of Polish names, which was taken from an official government source. The constructed dictionary contains all inflectional forms of these names obtained by an heuristic algorithm based on the suffix forms, e.g. for words ending with 'ka' the forms ending with 'ki', 'ce' 'ką', 'kę' are generated (for *Dominika* it gives us *Dominiki*, *Dominice*, *Dominiką* and *Dominikę*). Recognition of surnames is not so trivial, as it is impossible and impractical to have a list of all potential surnames. We defined the following heuristics helping us to judge whether names are followed by surnames:

- if the word after a first name is capitalized, and it's not a name, it's probably a surname,
- if the word before a first name is capitalized, and it's not a title or degree and after the name there is no capitalized word, the word before it's probably the (a) surname,
- if the word after a title or degree is capitalized, and it's not a name, it's surname,
- if a string is once judged to be a surname, it is treated as a surname within the entire text; the application also finds its (automatically constructed) inflected forms.

2.2 Institution Names Recognition

Institutions' names recognition algorithm starts with searching for selected keywords. In the first step, the name limits are the beginning and the end of the sentence containing a keyword, it can be for example the following entire sentence:

- (1) [Dzisiaj Rektor Uniwersytetu Kardynała Stefana Wyszyńskiego Adam Górski wygłosi wykład.]
 [*Today, the Dean of the Cardinal Stefan Wyszyński University Adam Górski will give a lecture.*]

The only exception from the above rule is when one sentence contains more than one keyword – then the next keyword recognized becomes the border for the current phrase:

- (2) [W tym roku studenci Uniwersytetu Warszawskiego i studenci] [Akademii Medycznej organizują konferencje.]
 [*This year, students from Warsaw University and students from*] [*Medical Academy are organizing the conference.*]

Starting from the key word and using the following rules, we add next words to the potential institution name (the initially established maximal borders cannot be crossed). The next word belongs to the institution name if (this decision scheme is presented in Fig.3):

1. the word after the key word (or the latest added word) is capitalized,
2. the word after the key word (or the latest added) is not capitalized and it is:
 - (a) the word: “im.” or “imienia” and after it the next word is capitalized,
 - (b) “w” or “we” (in) and after there is capitalized word,
 - (c) “i” (and) and after there is capitalized word, and it is not the latest word in the interval the word
 - (d) “nr”, “numer” (numer) and after it the is capitalized word or the digit,
3. if the word after the key word (or the latest added) is a name or surname in genitive
 - (a) and it’s not preceded by the word “im.” or “imienia”
 - (b) and before the key word there is no word like “Rektor”, “Dyrektor” (Dean, Manager).

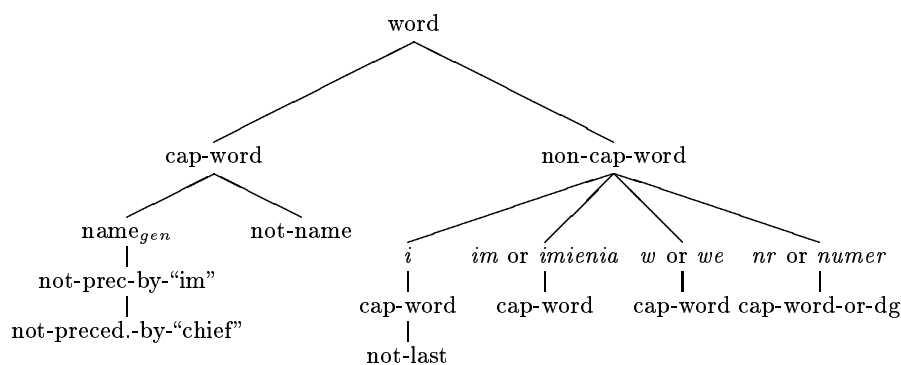


Fig. 3. “Decision tree”

2.3 Base Forms

Determining the base form of the recognized name entity is based on two kinds of information. The first one is the morphological data about the key words stored in the dictionary. The second source of data are two lists of suffixes characteristic for two genders. One list consists of the female suffixes (e. g. *ska*, *cka*, *owa*) the other one stores males (e.g. *ski*, *cki*, *owy*). For a given word, we search the suffixes and if we find one we postulate the gender and use stored suffix for the base form. If the word does not match with any of suffixes, we leave the word untouched.

2.4 Output Format

As the final result we obtain two output files. One of them is the input text with the beginning and the end of all the identified Name Entities marked. As tags we use those suggested at the MUC-6 conference. Each NE is marked with the ENAMEX tag of one of two subtypes: PERSON and ORGANIZATION.

For example:

- (3) <ENAMEX TYPE="PERSON">Dominika Urbańska</ENAMEX> studiuje w <ENAMEX TYPE="ORGANIZATION">Polsko-Japońskiej Wyższej Szkole Technik Komputerowych</ENAMEX>
<ENAMEX TYPE="PERSON">Dominika Urbańska </ENAMEX> is studying in <ENAMEX TYPE="ORGANIZATION">Polish-Japanese Institute of Information Technology</ENAMEX>

Second output file stores all the Name Entities, which were recognized in the texts and their base forms. For the sentence above we will get:

- (4) Dominika Urbańska:: base form Dominika Urbańska
:: Dominika Urbańska -lp -rz -M
Polsko-Japońskiej Wyższej Szkole Technik Komputerowych
:: base form Polsko-Japońska Wyższa Szkoła Technik Komputerowych
:: Polsko-Japońskiej Wyższej Szkole Technik Komputerowych -lp -rz -C¹
Polsko-Japońskiej Wyższej Szkole Technik Komputerowych
:: base form Polsko-Japońska Wyższa Szkoła Technik Komputerowych
:: Polsko-Japońskiej Wyższej Szkole Technik Komputerowych -lp -rz -N

3 Evaluation

Recognition of Polish first names and surnames is quite satisfactory. If the application was provided with the first name and surname of certain person, it is able to select correct answers and shows them. Entering e.g. "Dominika" we could cogitate as about a nominative of feminine name, but it could be also a genitive (possessive) or accusative of male's name form "Dominik". When the application finds e.g. "Dominika Urbańska", a surname is being interpreted as feminine, which conducts also set feminine gender to "Dominika" for all over the text. It means that this first name, which was underspecified for gender before analyzing this pair of words, becomes unequivocally specified as feminine for the entire text.

¹ Although in the sentence "Dominika Urbańska studiuje w Polsko-Japońskiej Wyższej Szkole Technik Komputerowych, the organization name appears in locative but the applications shows the two results: in locative and in dative. But please notice that the ablative form and dative form is the same. And because application does not know the contexts of following word in sentence it's impossible for it to predict which of the following form is the correct one.

The analysis of institution names' is definitely more complicated. Most problems occur when an institution name appears together with a person name. But the application correctly recognizes following examples: Rektor Akademii Jan Dąbrowski 'The Dean of Academy Jan Dąbrowski' and Rektor Uniwersytetu Adama Mickiewicza 'The Dean of University of Adam Mickiewicz'.

Another difficult situation solved by the program concerns the conjunction "i" (and). This word can be either a part of the institution name or can coordinate two different named entities e.g Wyższa Szkoła Handlu i Prawa i Akademia Medyczna 'Higher School of Trade and Law and Medical Academy' (first "i" is a part of an institution name, but the second one joins two different names).

The preliminary tests showed pretty good functionality of the application. For testing purposes, about one hundred short citations downloaded from internet or made by testers were used. Most of them contained between two and fifteen sentences. The results evaluation is showed in (5). It is divided into two parts: we evaluated pointing out the NE and separately the assessment of base forms.

(5)

	NE type	precision ²	recall
NER	persons	.98	.89
	organizations	.85	.73
base forms	persons	.92	.85
	organizations	.80	.70

What is interesting, compared to a commercial product based on the GRAM dictionary (<http://gram.neurosoft.pl/>), the application showed better results. For example, in the sentence *Profesor Kazimierz Subieta wykłada w PJWSTK*. 'Professor Kazimierz Subieta lectures at PJIT' the surname Subieta wasn't recognized by GRAM and in the sentence *Bolesław Prus napisał powieść "Faraon"* 'Bolesław Prus wrote a novel "Faraon"' Prus instead of being recognized as a surname was recognized as an genitive form of "Prusy" (a former country name). Both examples were correctly recognized by the application.

4 Conclusion

Summarizing, the application proved to be of a practical value. Implemented algorithms are capable of recognizing many person and institution names in Polish texts. The application is easy to modify or update. Obvious suggestions for further work is to include more types of proper names, namely location, time, monetary units and titles. It is also planned to combine it with an existing morphological analyzer. As it was emphasized at the very beginning, the process of the Name Entity Recognition is a key to many tasks in natural language processing and the presented program is planed to be used in IE applications.

² Precision is defined as usual as a ratio of applicable ones to all given answers, while recall is a proportion of provided positive answers to all positive answers.

References

1. Bick, E.: A Named Entity Recognizer for Danish, Proc. of LREC (2004)
2. Farmakiotou, D., V. Karkaletsis, J. Koutsias, G. Sigletos, C. D. Spyropoulos and P. Stamatopoulos: Rule-based named entity recognition for Greek financial texts. Proc. of the Workshop on Computational lexicography and Multimedia Dictionaries, COMLEX (2000)
3. G. Petasis, V. Karkaletsis, C. Grover, B. Hachey, M.T. Paziienza, M. Vindigni, J. Coch: Adaptive, Multilingual Named Entity Recognition in Web Pages, ECAI 2004 (2004)
4. Piskorski, J.: Extraction of Polish Named-Entities. Proceedings of LREC 2004, (2004)
5. Report from Human Communication Resource Center <http://www.hcrc.ed.ac.uk/AnnualReport/Text/game.html>, (1998)
6. Urbańska, D.: Analizator nazw własnych. Master's thesis. Polish-Japanese Institute of Information Technology, Warszawa (in preparation), (2005)

Creating of Slovak Electronic Phonetic Dictionary for Use in Speech Recognition

Pavol Vančo and Marek Nagy

Comenius University, Mlynská dolina, Bratislava 84225, Slovakia,
pvanco@gmail.com

Abstract. A phonetic dictionary is important for working with sound side of a language. Speech recognition is put in front of a task to create words from basic units - phonemes. A conversion mechanism that can attach words to sequences of phonemes is necessary for this task. A printed version of the Slovak phonetic dictionary is the only accessible source. The conversion from this printed version to an electronic version was needed. The transformation of old phonetic transcription to the new SAMPA phonetic symbols was needed as well. Problems that appeared during conversion process were solved either by adding new symbols or by using new words.

1 Introduction

In the recent years the need for a faster and more comfortable way to communicate with computers became a problem that can be solved by a human interface using recognition (also known as digital signal processing or DSP) and synthesis. The most important is speech and pattern recognition. As many of the computer related problems the speech recognition is trying to copy the way people recognize speech. The problem is complexity of the recognition process. In a simplified way the recorded sound is transformed and then compared to the known patterns. These patterns can be either whole words (for example represented as sound samples) or phonemes. Whole words approach is effective only if the task is to recognize words that are always the same – these are usually basic commands.

The approach based on phonemes is more complicated and more effective. Phoneme is a basic, theoretical unit of sound. Every word consists of one or more phonemes. The same letter can have different sound and therefore different phoneme if it's in different context - for example the letter "v" in slovak words "vedieť" and "včera". Words are sorted in alphabetical order in classic phonetic dictionary and they have their phonetic transcription assigned. In speech recognition alphabetical order of phonetic transcription is needed. That means we need to use reversed phonetic dictionary. That is why it is necessary to have electronic version of phonetic dictionary. It can be used not only for speech recognition but also for speech synthesis.

SAMPA	Dictionary	SAMPA	Dictionary
a	a	n	ɲ
a:	á	N	ɳ
ɤ	ä	r	r
E	e	r=	ɾ
E:	é	r=:	ɽ
I	i	l	l
I:	í	l=	ɭ
O	o	l=:	ɮ
O:	ó	J	n
U	u	L	ɭ
U:	ú	f	f
I_ ^a	ia	v	v
I_ ^E	ie	f_v	w
I_ ^U\	iu	U_ ^	ũ
U_ ^O	uo	s	s
p	p	z	z
b	b	S	š
t	t	Z	ž
d	d	x	x
c	t	h\	h
J\	d	G	ɣ
k	k	j	j
g	g	I_ ^	í
m	m	ts	c
F	ɲ	dz	dz
n	n	tS	c
		dZ	dž

Table 1. Slovak SAMPA

2 Making of Phonetic Dictionary

International phonetic alphabet (IPA) is used for electronical transcription of phonemes. All electronic phonetic texts should be written in this alphabet. The problem is that it is using symbols not available in computer ASCII (American Standard Code for Information Interchange) code that is most universal. ASCII code has only 95 printable characters and therefore IPA with ASCII cannot be used together (at least not straightforward). The solution is either by using UNICODE (this encoding type has 65 535 characters) or by using other transcription scheme. The one I used is SAMPA [4] as this type of transcription is international as well as it has slovak version. It also uses only ASCII encoding for phonemes. However when I was transcribing phonetic dictionary I found some characters not present in the basic set (table 1).

Different symbols than the ones that can be found in slovak SAMPA were mostly found in words from foreign languages. The remaining symbols were properties of phonetic dictionary I was working with (stress symbols, arcs, spaces,

Symbol used	Symbol found in dictionary
[?]	half of reversed question mark
[z=]	ž
[z=:]	ẓ̌
[N\]	mirrored ɛ
[Q\]	ř
[y:]	ů
[y]	ü
[2]	ö
[2:]	ő
[@]	mirrored e
[_.]	arc under dash
[=]	arc under space
[sp]	space between words
["]	main stress
[%]	secondary stress
[t_>]	double letter
a[;o]	ä
e[;:]	ē
e[;)]	ẹ̄
o[;)]	ọ̄
é[;)]	ẹ̣̄

Table 2. New rules for slovak SAMPA

etc). The symbols from other languages were created according to SAMPA for that particular language (French, German and Hungarian symbols). The remaining symbols were added by myself (table 2). I used existing rules to create new ones so that my version is consistent with old rules.

2.1 Transcription

The only available phonetic dictionary is Slovak phonetic dictionary [3]. To create electronic phonetic dictionary I had to transcribe this book first. It was too difficult to transcribe the book into SAMPA symbols immediately. Therefore I created my own symbol vocabulary based on $\text{T}_{\text{E}}\text{X}$ rules. This vocabulary could be used as its own phonetic transcription but as it was not compatible with other encodings I transformed it to SAMPA version. There were also other problems based on the way the original phonetic dictionary was composed. Some words were doubled (used on two places). There were different versions of the same word on the same line. For speech recognition it was better to save all forms of the word; and not just the basic form. That made transcribing this book quite difficult.

I programmed small and useful scanners to do manual minor repairs and changes as well as preprocessing for next step. The steps were as follows:

- Changing two column style to one column style

- Removing special characters that were used in book and were not relevant to phonetic transcription
- Creating new lines for multiple words on the same line
- Manual checking of errors
- Creating phonemes for newly created words
- Manual checking of errors

The very last step was changing encoding to SAMPA version.

2.2 Statistics

The phonetic dictionary is quite big - it has 420 pages with words and their phonetic transcriptions. Information about the quantity of pages per letter are in table 3. This information can be used for statistical purposes as the information about how many words are there for one letter. Final electronic version has 66675 words with their phonetic equivalents. Text file using UTF-8 encoding has more than 2.1 megabytes. I used 49 rules to change my encoding to SAMPA encoding. It took me more than two months.

Letter	A	B	C	D	E	F	G	H	CH	I	J	K	L	M	N	O	P	Q	R	S	Š	T	U	V	W-Y	Z
Pages	13	14	10	24	8	8	5	12	4	10	4	24	13	19	24	28	56	1	24	33	9	20	9	29	3	32

Table 3. Pages per letter

3 Conclusion

I used my knowledge of electronic text transliteration to convert book version of Slovak phonetic dictionary into its electronic version with few changes. Electronic version is more complex has a few more words and also can be reversed and sorted alphabetically according to phonemes. That makes it ready to use for speech recognition.

Current version is ready to use. But still it can be updated with various information such as word classes, noun cases, verb tense, etc. This information is valuable when using for more difficult tasks like speech understanding. The next step is creating speech recognition system to be used with this dictionary.

References

1. Mistrík J.: Jazyk a reč. Mladé letá (1999)
2. Kolektív autorov: Pravidlá slovenského pravopisu. VEDA (2000)
3. Kráľ, Á.: Pravidlá slovenskej výslovnosti. SPN, Bratislava (1983)
4. SAMPA homepage: <http://www.phon.ucl.ac.uk/home/sampa/home.htm> (2005)
5. Slovak SAMPA homepage: http://www.ui.savba.sk/speech/sampa_sk.htm (2005)

Russian Historical Corpora of the 18th and 19th Centuries

Victor Zakharov ^{1,2}

¹ Department of Mathematical Linguistics
Philological Faculty, St. Petersburg State University
Universitetskaja emb., 11, 199034 St. Petersburg, Russia

² Institute for Linguistic Studies, The Russian Academy of Sciences
Tuchkov st., 9, 199053 St. Petersburg, Russia
vz@l1z1168.spb.edu

Abstract. The paper deals with a corpus of the Russian language of the 18–19th century. It covers the period of both the 2nd half of the 18th century and the 19th century. The whole 18th century is to be included into the corpus on the next stages of the project. The corpus is intended for compiling historical dictionaries of the above periods. Additional metadata tags were contrived that correspond to lexicographical labels. Experiments in morphological tagging are described. The results obtained provide a baseline allowing to improve graphematical (premorphological) processing of texts and to work out linguistic and algorithmic tools to improve morphological analysis.

1 Introduction

Recently the creation of different text corpora has been at the cutting edge of applied linguistics. In Western countries corpus linguistics formed a separate linguistic universe in the early 1990s [1, 2], even though the concept of *corpus* and the first electronic corpora had been known long before. One of the most important notions used in creating a corpus is a mark-up (tagging, annotation). We are inclined to see it as a borderline separating a linguistic corpus from an electronic collection of texts. The whole corpora universe can be broken down into two major groups: specialized and general corpora, the latter applying to the language in general (normally, to one of its historical periods). A lot of corpora of different languages have been created for the last years. So far Russian linguistics lags behind its Western counterparts in corpus studies. The most renowned Russian corpus is the Uppsala Corpus of Russian Texts created in the remote 60s and outside Russia. By now its linguistic material is neither up to date in volume (one million word occurrences), nor complies with modern conceptions of a national corpus at all. And furthermore, this corpus lacks in linguistic annotation, a feature which is absolutely indispensable in a state of the art in creating a corpus today. In two decades during the 80s and 90s in Russia, the Russian language databases were built at the Institute of the Russian Language within the Computerized Russian Language Fund Program [3]. Unfortunately, the accumulated results were either abandoned or lost and did not become public heritage. Meantime quite a number of research teams are doing their best in continuing the program on the modern basis. The results of these efforts were presented at conferences and published (see references in [4]). The project of the National Corpus of the Russian Language was started in 2002 to form a representative

corpus of modern Russian with over 100 million words. The present state of corpus of 65 mln. words (as of fall 2005) can be found on the Internet at the address <http://ruscorpora.ru>. The description of project can be found in [5].

2 The Projects of Historical Corpora

Within the framework of the National Corpus of the Russian Language there will be a corpus based on the 19th century Russian texts. Collection and preprocessing of the texts for the National Corpus of the Russian Language are carrying out at the Institute for Linguistic Studies of the Russian Academy of Sciences in Saint-Petersburg, being a part of work in compiling a new dictionary of the Russian language of the 19th century [6]. The dictionary is actually planned as a historical one since the lexical items appeared in the 19th century or underwent any changes during that period will be included in it. What really differentiates Russian of the 19th century from other periods is strong dynamics of its lexical and semantic systems [7]. The corpus of the 19th century is expected to be both a tool and a source for multifaceted works in lexicography with the final goal of compiling a historical dictionary of the period.

When compiling a dictionary of a certain time (in our case it is the 19th century) a scholar constantly faces the necessity of comparing the material being analyzed with the lexis of the periods that precede or follow the time of interest. It sounds reasonable that not only texts of the 19th century but also texts (and/or dictionaries) of the preceding and the following time periods can be useful as the empirical base for the dictionary. And if the 20th century material is adequately covered by the National Corpus of the Russian Language together with texts posted on the Internet, there are very few texts of the 18th century there and they are far from being representative. All these reasons made us start building a corpus that will incorporate the second half of the 18th century as well. This project began with the creating a corpus of texts by Mikhail Lomonosov – the great Russian scientist and a man of letters. The further stages will engulf the whole textual mass of the 18th century.

Chronological boundaries prove to be one of the problems when one is trying to create a corpus of a certain period. The corpus of the 18th century will start with the texts originating in 1708, the year when the Civil alphabet was introduced into Russia. The boundary between texts of the 18th and 19th century may be set at 1810–1812. This period marks a turning point in development of lexical structure of the Russian language, e.g. wider penetration of abstract words and colloquialisms into literary texts. This can be accounted for by significant social and cultural factors. Replacement of former Colleges by Ministries, conquest of Bessarabia, creation of Anti-Napoleon European coalition, the Patriotic War of 1812, that triggered the global makeover of Russian social mentality, presence of Russian troops in Europe, introduction of constitutional governments into Finland and Poland, exploration of the Far East and Kamchatka by the Russians – this is a brief list of the most important events in 1810s – 1820s. It should be noted that Russian lexicographers Y. S. Sorokin and L. L. Kutina, working on the “Draft Dictionary of the XVIII century Russian Language” [8], set the upper boundary at 1803-1805 with a partial capture of some later material.

The upper boundary of 19th century is then suggested to move up to 1904 – 1905. It could enable us to reflect to the most extent rapid replenishing of lexical and phraseology vocabulary, semantic reorganization of many words accompanying with their stylistic shift and determining new correlation between the literary and colloquial layers in various styles appearing at the end of the century. The Russian language of this period represents the paramount stock of words and word combinations being heterogeneous both in their origin and semantic/stylistic characteristics.

As a foundation for the corpus a proportional representative selection was made, including:

- 1) original fiction – prose and poetry;
- 2) translations;
- 3) social and political journalism;
- 4) popular science essays;
- 5) letters, personal journals and memoirs;
- 6) business and official documents;
- 7) religious literature.

It is of major importance to ensure isomorphism between our selection of texts and the real genre/style breakdown in the Russian texts.

3 Morphological Tagging

All corpus texts are kept in three forms, namely: a) the text archive (texts in the original form); b) the corpusoid[9] (processed and marked-up XML-texts); c) the database of a corpus manager.

In general, there are three levels of text description and, thus, three types of metadata, namely, quasi-bibliographic, structural and linguistic ones. The first level of texts description includes a set of standard bibliographic data elements and a set of elements of literature and book science characterizing genre, style, the history of writing and publishing, etc. [10]. At the second level the following elements are added to the document structure: text, chapter, section, paragraph, sentence, phrase, word. Syntactically only the identification of the some multiword units (e.g., "potomu chto", "tak kak", "iz-za", "v techenie") is fulfilled. In this stage some special tags are produced for the beginning and for the end of a sentence (<sent>...</sent>) and for the rest of the punctuation signs (<pun>...</pun>) that will be used by the procedures of the morphological disambiguation. The morphological tagging, i.e. lemmatization and assigning morphological characteristics to all tokens is carried out at the third level.

The modules of Dialing translation system [11] are used for automated tagging. Morphology interpretation of each text token constitutes a triplet: <Lemma, Part of Speech, Grammmemes>. A grammeme is an elementary descriptor attributing a token to a certain morphological class, e.g. *камнем* (*by a stone*) with lemma КАМЕНЬ will be described by the following set of grammemes: "masculine, singular, instrumental, inanimate". All possible lemmas and sets of grammemes are given for each token. This morphological ambiguity will mainly be solved at next steps of analysis. But the morphological pro-

cessor applies heuristic context-dependent rules that resolve some simple cases of ambiguity.

A set of possible tags and their attributes was defined. XML was chosen as a mark-up language which allows effective processing of tagged texts by different standard applications, both today and in future. XML-tags and their attributes are as follows: **<text>** - text, **<p>** - paragraph, **<s>** - sentence, **<w>** - word (token), **<ana>** - common tag for morphology (attributes lemma, pos (part of speech), gram (grammeme)), **<pun>** - punctuation mark, **<dg>** - tag for numerals, **<frgn>** - tag for foreign fragments.

Attributes " <i>pos</i> " of the tag <i><ana></i>	Attributes " <i>gram</i> " of the tag <i><ana></i>
С - noun, П - adjective, Г - verb, ПРИЧАСТИЕ - participle ДЕПРИЧАСТИЕ - adverbial participle ИНФИНИТИВ - infinitive МС - pronoun, МС-П - pronoun-adjective, МС-ПРЕДК - pronoun- predicative, ЧИСЛ - numeral, ЧИСЛ-П - numeral-adjective, Н - adverb, ПРЕДК - predicative, ПРЕДЛ - preposition, СОЮЗ - conjunction, МЕЖД - interjection, ЧАСТ - particle, ВВОДН - parenthesis, СРАВН - comparative, ФРАЗ – idiom.	мр, жр, ср – gender: masculine, feminine, neuter; од, но - animate, inanimate ед, мн – number: singular, plural; им, рд, дт, в н, тв, пр – case: nominative, genitive, dative, accusative, instrumental, locative; св, не – aspect: perfective, imperfective; пе, ни – transitive, intransitive verb; дст, стр – voice: active, passive; нст, прш, буд – tense: present, past, future; пвл – mood: imperative; 1л, 2л, 3л – person: first, second, third; кр – short form (for adjectives and participles); сравн – comparative (for adjectives); 0 – indeclinable; имя, фам, отч - names: personal, family, patronymic; лок, орг - place, organization; кач - qualitative adjective; вопр,относ – relative mode, interrogative mode (for adverbs); дфст – a word usually has not plural; опч – frequent misprint or error; жарг – slang.

Fig. 1. Meanings of attributes of the tag *<ana>*

The question now arises of whether the contemporary Russian tagger can be used to automatically tag the old-time texts? The morphological dictionary of Dialing system created on the basis of A. Zaliznyak's "Russian Grammar Dictionary" [12] contains more than 130 thousand lemmas that cover about 3 million word forms, however it turns to be insufficient for real texts. One ought to remember that this dictionary describes the standard language of the second half of the 20th century. What it lacks is many types of words such as colloquial, dialect, slangy, professional etc. And quite naturally, the obsolete lexis of the 18–19th centuries is poorly presented in it as well. When

tagging a text morphologically one is to face a number of phenomena which are not present in the current morphological system. This includes variant inflexions of genitive singular masculine forms (nouns) (*у берегу – у берега, блеска – блеску, дому – дома*¹), and of nominative plural masculine and neutral forms (nouns) (*веки – века, снега – снега, кольца – кольца, блюда – блюда*), superlative forms of adjectives (*самый нежнейший, самый юнейший* (F.M. Dostoyevsky "The Double")), subjunctive forms of participles and adverbial participles (*обретший бы, вытупавшись бы*), monosyllabic intransitive adverbial participles (*ждя, шья, бья (бия), чтя*), the usage of archaic forms connected with the Church Slavonic language due to the long history of Russian, e.g., the nominative-accusative dual form². The following examples from the texts by Gogol illustrate peculiarities of the morphological system of the 19th century: *"два русские мужика"* (modern usage: два русских); *"двое какие-то мужчин"* (двое каких-то); *"влача за два деревянные кляча изорванный бредень, где видны были два запутавшиеся рака и блестела попавшаяся плотва"* (деревянных, запутавшихся); *"у ней деревушка"* (у неё), *"я с вами расстаюсь не долее как на два дни"* (дня).

The analysis of the tagged texts has highlighted various problems to be solved. The main reasons for failures in morphological tagging of old Russian texts are as follows: incompleteness of the morphological dictionary; failure to tag word-formative derivatives; failure in tagging because of inflexion problems; insufficiency of graphematical analysis. For a detailed evaluation of morphological tagging see [13].

When tagging such inflectional languages as Russian it is strongly recommended to base your work on a morphological dictionary. The solution of the problem consists in processing of significant amount of texts, revealing unknown words, describing these words and adding them to the dictionary. This will entail its replenishment.

The Dialing tagger has a prediction algorithm for unknown words which searches in the dictionary for the most conterminous word form to an entrance word [14]. When tagging, a "predictor" assigns to an unknown token such a set of possible tags which is determined by the similar dictionary entry. The accuracy of prediction within the universe of contemporary Russian current-news texts is between 85 and 90%. The prediction is dubbed accurate if it produces at least one correct answer. This algorithm, though, when applied to the 18-19th century vocabulary, does not prove efficient enough, yielding but 70%, and thus needs further improvement. The expansion of lists of prefixes and suffixes as well as the use of more complicated linguistic algorithms are probably required.

However, the completed experiments do testify that the texts of the 18th/19th centuries can be successfully subjected to automatic procedure of morphological tagging. The specific lexis of the appropriate sublanguage must be identified and added to the morphological dictionary of the system. As new material becomes available, the module of morphological tagging will be also adapted to the language of the 18th/19th centuries.

Special attention must be paid to the presence of proper names in the automatic dictionary. Moreover, it does not apply only to automatic dictionaries. This issue ap-

1 «Сделали десять шагов от **дому** к крыльцу» (А.Ф. Вельтман, "Саломея", 1846).

2 «Боги праведные дали Одинакие **крилѣ**» (Е.А. Баратынский, 1835)

pears to be a lexicographical one: should a historical dictionary of the language contain proper names? If so, to what extent? I'm inclined to say, yes. The new edition of the "Russian Grammar Dictionary" includes about 8,000 proper names. But these are mostly modern names used in contemporary texts. Texts of old periods contain very specific names both semantically and morphologically³.

4 Corpora and Lexicography

It seems that to perform lexicographical tasks would require a special kind of metadata. I have an impression that the National Corpus of the Russian Language and all modern existing corpora are hardly suitable for practical needs of **dictionary** compiling and **diachronic lexicography**. For these tasks lexical and semantic tagging, fixation of lexical and semantic variants are required. It is a single concrete meaning or, say, shade of meaning that is so important for lexicographers and lexicologists. And not a lexeme but such a meaning does become the subject of annotation. So, the word *дело* has from 11 to 15 senses in different dictionaries. How quickly and effectively can one find, for example, «дело» in the contexts with meanings «бой, сражение» (battle, fight)? An other meaning «деятельность, действия, поступки» (activity, work, behavior) opposes to the meaning «мысль, идея» (thoughts and words)⁴. Today it is quite difficult to choose and to define appropriate meaning from 6150 contexts in the National Russian Corpus.

Special lexical metadata tagging with a set of parameters seems necessary for all this. These parameters will have the meanings: original – adopted, normative – off-normative, contemporary – becoming obsolete – obsolete (*аполитизм, бражник, десть, селадон* etc.), common – special (*архитрав, антаблемент, болометр, безунок*), direct sense – figurative sense (*автомат, ковер, тюфяк*) and so on. This lexical tagging is to be made automatically according to the pre-established coordinates. At present these coordinates can only be represented by an authoritative dictionary, for instance, «Dictionary of Contemporary Russian» in 17 volumes or «Explanatory Dictionary of Russian» (edited by S. Ozhegov, N. Shvedova).

There is another way available – manual content tagging of quite a small proportional archive of sample texts and using this archive as a model for automated content marking. The algorithm of our actions in this case will be as follows: manual lexical tagging of some representative texts – creation of checklist on the basis of this subcorpus – then automatic tagging of next texts on the basis of checklist – verification of the results obtained and modification of the dictionary, etc.

This content (or semantic) corpus is to comprise elements of an expert system which, apart from the corpus, would include as its components, electronic quotation index, various dictionaries, and means of linguistic and statistic processing of corpus or index data. The main task of the system is to supply lexicographers with lexical data

³ There is an example to be given from historical toponymy. Texts of the 18–19th century are known to have obsolete place names used for stylistic purposes: *Гиперборейское море, Азвония*. Some of them without fail ought to be included into the dictionary.

⁴ «Обломов засмеялся этому, как *делу* совершенно невозможному» И.А.Гончаров. "Обломов".

archive and tools to gain unbiased information about word and its relations to others, to classify word contexts, etc – in other words, to compile materials for dictionaries. In short, a special program and a linguistic system should be designed to allow a user to operate efficiently with corpus data in combination with citation card files. The latter, which are available at the Institute for Linguistic Studies in Saint-Petersburg, count more than 12,000,000 cards. The third part of this system is a set of existing dictionaries of Russian presented in a digital form.

Finally, I argue that future corpora and dictionaries must be structured in a different way. A corpus should be an indispensable supplement to a dictionary. A dictionary always represents but the main nomenclature of lexical items in their general meanings. Dictionaries and grammars hardly ever take into consideration the unstable and probabilistic nature of the language. There are a lot of lexical and semantic variants, lexical items, and set expressions usually ignored when compiling a dictionary. Synonymy is also accounted in dictionaries but partially. The same situation is observed with lexical collocations. A corpus presents means for discovering and describing of such cases. Thus, a corpus must be viewed both as a tool for compiling a dictionary and as its integral component.

5 Conclusion and Further Work

In this paper we have presented the work that has been done in developing a corpus of the Russian language of the 18th and 19th centuries. The corpus is intended for compiling historical dictionaries of named periods. Special metadata tags necessary for lexicographers were invented. The corpus has to be combined with existing dictionaries of Russian and with citation card files. The experiments in automatic morphological tagging of the 18th and 19th centuries texts were described. The experiments showed that texts of the 19th century and even the 18th century can be processed by our morphological tagger. In practice, however, the program dictionary and algorithms have to be tweaked for Russian of the cited time.

Another important task of the project is presentation of corpus texts in its historical alphabet. The implementation of UNICODE simplifies the problem. Nevertheless, there remains the problem of font presentation on users' computers and at the same time we want to allow users to QUERY in modern alphabet, thus, we have to ensure compatibility of words in old and new alphabets within a corpus manager.

In the following stages of the project it is planned to cover the text data archive of the whole 18th century. We do have an opportunity to create the corpus that will include all the texts of this century (or at least most of them). The joint electronic catalog of civil press books, designed by the National Library of Russia, counts about 12,500 bibliographic records. Theoretically, we have an opportunity to digitize and annotate all the texts in civil alphabet (approximately 150,000,000 tokens). Finally this material will form the historical corpus of the Russian language of modern history.

Acknowledgements

This research has been partly supported by the grant of the Russian Humanitarian Scientific Fund No. 03-04-00232a (The Vocabulary of M.V. Lomonosov). I'm grateful to the students of the Mathematical Linguistics Department of Saint-Petersburg State University for their manual work on analyzing tagged documents and namely to one of them, Maria Khokhlova, for her support in this work. My heartfelt thank goes also to Sergei Volkov for his deep thoughts and inspiring discussion concerning issues of analyzing Russian texts of previous centuries.

References

1. Leech G. The State of Art in Corpus Linguistics // *English Corpus Linguistics* / Aijmer K., Altenberg B. (eds.) London, Longman, 1991. P. 8–29.
2. Fillmore Charles J. Corpus linguistics' vs. 'Computer-aided armchair linguistics. *Directions in Corpus Linguistics* / Svartvik Jan (ed.). Mouton de Gruyter, 1992. (Proceedings from a 1992 Nobel Symposium on Corpus Linguistics, Stockholm). P. 35–60.
3. Andrjushhenko V. M. *Koncepcija i Arkhitektura Mashinnogo Fonda Russkogo Jazyka.* / Ershov A.P. (ed.). M., 1989.
4. *Doklady Nauchnoj Konferentsii "Korpusnaja Lingvistika i Lingvisticheskie Bazy Dannykh"* / Gerd A.S. (ed.). St. Petersburg, 2002; *Trudy Mezhdunarodnoi Konferentsii "Korpusnaja Lingvistika–2004"* / Belyaeva L.N., Gerd A.S., Zakharov V.P., Koval S.A., Mitrofanova O.A. (eds.). St. Petersburg, 2004; *Proceedings of the International Conference «Megaling'2005. Applied Linguistics at Crossroads»* / Dikareva S.S., Zakharov V.P. (eds). St. Petersburg, 2005; *"Nauchno-tehnicheskaya informatsiya", Special Issues, 2003, N6, 2005, N3.*
5. *Korpusnaja Lingvistika: Natsionalnyi Korpus Russkogo Jazyka* // *"Nauchno-tehnicheskaya informatsiya", Special Issue, 2005, N7.*
6. *Proekt Slovarya Russkogo Jazyka XIX veka.* / Volkov S.S. (ed.). St. Petersburg, 2002.
7. Volkov S.Sv. *Ob "Istoricheskom slovare russkogo yazyka XIX veka".* // *Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskikh issledovanii RAN.* Tom I. Chast' 3. SPb, 2003, S. 85-94.
8. *Slovar' Russkogo Jazyka XVIII veka.* Proekt. / Sorokin Yu.S (ed.). Leningrad, 1977.
9. Garabík R. *Corpus Construction Tools* // *Proceedings of the International Conference «MegaLing'2005. Applied Linguistics at Crossroads», Meganom, Crimea, Ukraine, June–July 2005* / Zakharov V.P., Dikareva S.S. (eds). P. 25-32.
10. Volkov S.Sv., Zakharov V.P., Dmitrieva E.A. *Metarazmetka v istoricheskom korpuse XIX veka.* // *Trudy mezhdunarodnoi konferentsii "Korpusnaja lingvistika–2004"*. SPb., 2004. S. 86-98.
11. Sokirko A. A short description of Dialing Project URL:
<http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html>
12. Zaliznjak A. A. *Grammaticeskij slovar' russkogo jazyka.* Moscow, 1977, ..., 2003.
13. Zakharov V., Volkov S. *Evaluating Morphological Tagging of Russian texts of the 19th Century* // *Text, Speech and Dialogue. Proceedings of the 7th International*

Conference TSD 2004, Brno, Czech Republic, September 2004 / Petr Sojka, Ivan Kopeček, Karel Pala (eds.). – Springer-Verlag, Berlin, Heidelberg, 2004. P. 235-242.

14. Sokirko A. Morphological modules on the site www.aot.ru: Prediction of unfound words // URL: <http://www.aot.ru/docs/sokirko/Dialog2004.htm>

Building a Pilot Spoken Corpus

Jana Zemljarič Miklavčič and Marko Stabej

Filozofska fakulteta, Univerza v Ljubljani
jana.zemljarič@ff.uni-lj.si

1 Introduction

A linguistic corpus is a collection of texts which have been selected and brought together for studying the language. The texts of a well-designed corpus should be sampled in a way to represent, as far as possible, the language they are chosen from. Therefore, spoken corpora are built as resources for linguistic research of spoken language. Spoken language here means any language whose original presentation is in oral form. “If such a text is later presented in written form, without change except for the transcription, it should be classified as spoken” (Eagles Preliminary Recommendations on Corpus Typology, 1996).

Research potential of spoken corpora is large and heterogeneous. They can be used for examination of theoretically settled grammatical rules or for new language descriptions, based on empirical data, especially in lexicography and grammar studies. They are also of significant importance as reference sources in language learning, pragmatic studies, discourse analysis and contrastive studies. Recently, spoken corpora have also been recognized as important language resources even in speech technologies.

Spoken language data are, as it is commonly known, very difficult to work with. The words that appear in an orthographic transcription of a speech event constitute only a partial representation of the original speech event. To supplement the lack of information, the analyst can capture other features, by making either a prosodic or phonetic transcription, and can also record contextual features. Clearly, the level of transcription will depend upon the purposes for which the corpus is being developed. For a linguist whose interest is in the patterning of language and in lexical frequency over large quantities of data, “there will be little need for sophisticated transcription, and the main consideration will be the quantity and speed of transcription work” (Thompson 2004). The phonetician, on the other hand, requires less data, but a high degree of accuracy and detail in the phonetic transcription of recordings, with links, where possible, to the sound files. For a discourse analyst, richly detailed information on the contextual features of the original events will be needed. However, the more detailed the transcription is, less readable it becomes.

The project of building a pilot spoken corpus of Slovene was aimed at theoretical and practical foundation on building a spoken corpus of Slovene language, which is planned to complement the 300 million word FidaPLUS¹ corpus as its spoken component. The purpose of a pilot corpus was to redefine the criteria for the collection,

¹ <http://www.fida.net/slo/index.html>

selection and documentation of spoken materials, to develop and test transcription criteria and mark-up conventions, and finally to show some possibilities for the use of a corpus for linguistic purposes and language analysis. The aim of the project was to have a pilot spoken corpus, available in searchable form, with transcriptions linked to sound files.

2 Pilot Corpus Design

Slovene language is spoken by 2 million speakers in Republic of Slovenia and in Slovene minorities in Italy, Austria and Hungary, partly also in diasporas around the world. Corpus of adult native speakers of Slovene should gain representativeness through a combination of demographic and contextual sampling. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other. For the criteria of demographic sampling, 4 categories were proposed: sex, age, region and education, and for contextual sampling structure of a text (mono-, dialog), setting (public, private), relation between speakers and hearers (formal, informal), media (face to face, telephone, radio, TV) and genre of a text. All criteria have been taken into consideration while building a pilot corpus.

The pilot corpus consists of 7 digital recordings in total length of 89 minutes. All texts were recorded in year 2004. The specification of the recordings is shown in the following table:

ID	Duration (min.)	No. of speakers	Place of recording	Surreptitious	Genre
R01	2.17	2	University	No	interview
R02	54.50	6	Studio	No	round table
R03	3.58	2	Home	No	interview
R04	7.31	5	Office	No	spont. convers.
R05	3.23	5	Skate-park	No	interview
R06	11.54	3	Workplace	No	spont. convers.
R07	5.12	2	Home	Yes	spont. convers.

Table 1. Pilot corpus recording's documentation

All information about speakers has been collected on speaker's identity lists. The data are represented in following table:

ID	Sex	Year of Birth	Age	Education	Region
G01	F	1963	41	University	Central
G02	M	1965	39	University	Central
G03	F	1966	38	University	Central
G04	F	1967	37	University	Central
G05	F	1968	36	University	Central
G06	F	1968	36	University	Central
G07	M	1970(?)	34(?)	University	Other
G08	M	1933(?)	71(?)	University	Central
G09	F	1979	25	University	South-east

ID	Sex	Year of Birth	Age	Education	Region
G10	F	1967	37	High school	North-west
G11	M	1987(?)	17(?)	Primary sch.	Central
G12	M	1987(?)	17(?)	Primary sch.	Central
G13	M	1987(?)	17(?)	Primary sch.	Central
G14	F	1976	28	University	South-east
G15	F	1979	25	University	Central
G16	M	1978	26	High school	Central
G17	M	1978	26	High school	Central
G18	F	?	?	?	?
G19	F	1969	35	University	North-west
G20	M	1948	56	High school	North-west

Table 2. Pilot corpus speakers' documentation

The sample of 20 speakers is representative according to the sex of the speakers but not according to other demographic criteria. The actual spoken corpus should consist of texts representatively taken from 5 areas that represent 5 dialectal groups of Slovene language. Furthermore there should be 3 age classes and 3 educational classes. The rather opportunistic nature of a pilot corpus should be taken into consideration while analyzing the corpus.

Pilot corpus is better designed in the concern of contextual criteria: different structure types, settings, speaker's positions, genres and media are represented among the texts. However, the telephone conversations and some other text genres should necessary be added to the planned spoken corpus. The final design of the pilot corpus according to contextual criteria is presented in following table:

Contextual criteria	Proportion
Dialogue (or multilogue) vs. Monologue	94 % : 6 %
Private vs. Public	19,5 % : 80,5 %
Informal vs. Formal	35,5 % : 64,5 %
Media vs. Face to face	31 % : 69 %
Surreptitious vs. Nonsurreptitious	5,6 % : 94,4 %

Table 3. Texts according to selected contextual criteria

3 Transcription and Mark-up

The actual transcription work has been carried out with two transcription tools, *Transcriber* and *Praat*. *Transcriber* is a tool for segmenting, labeling and transcribing speech, whereas *Praat* is a program for speech analysis and synthesis. Both programs allow automatic alignment of transcriptions and audio files and are available as free software on the internet. We found *Transcriber* more user-friendly for transcribing than *Praat*, however, it doesn't enable transcribing overlapping speech for more than two speakers. *Praat*, on the other hand, is less suitable for transcribing and works very slowly for longer recordings (more than 30 minutes) but it allows transcribing overlapping speech of more than two speakers which is often the case with spontaneous speech.

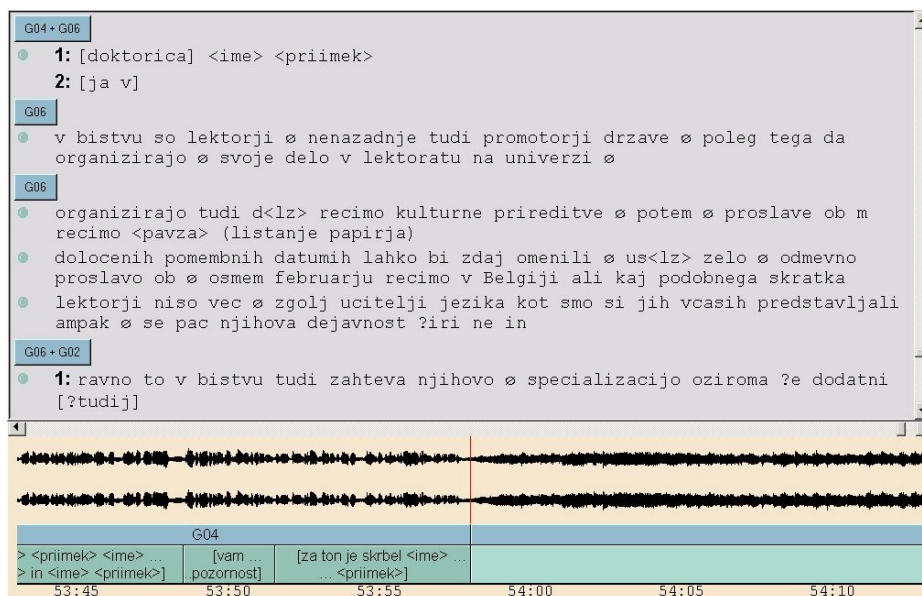


Fig. 1. Transcriber working platform transcription of the spoken corpus of Slovene

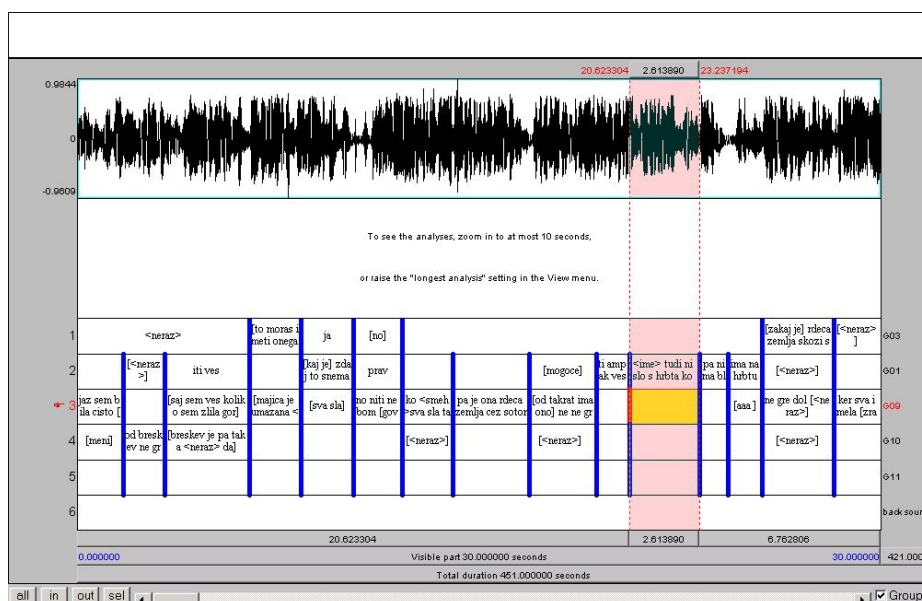


Fig. 2. Praat working platform, transcription of the spoken corpus of Slovene

During actual transcription work, the transcription principles for pilot corpus have been decided. Some of the adopted tags later proved to be inadequate and will be a subject to change in the future; however, at this stage they can not be changed. We were following the TEI and EAGLES recommendations on transcribing and annotating spoken texts, but also looking for an individual format, corresponding to the nature of Slovene language. A modified orthographic transcription is usually used when spoken corpora are concerned and so it is the case in pilot spoken corpus of Slovene. Basic unit of a speech is an utterance, defined by a short pause or a speaker turn. No punctuation is used and capital letters are used for proper names only.

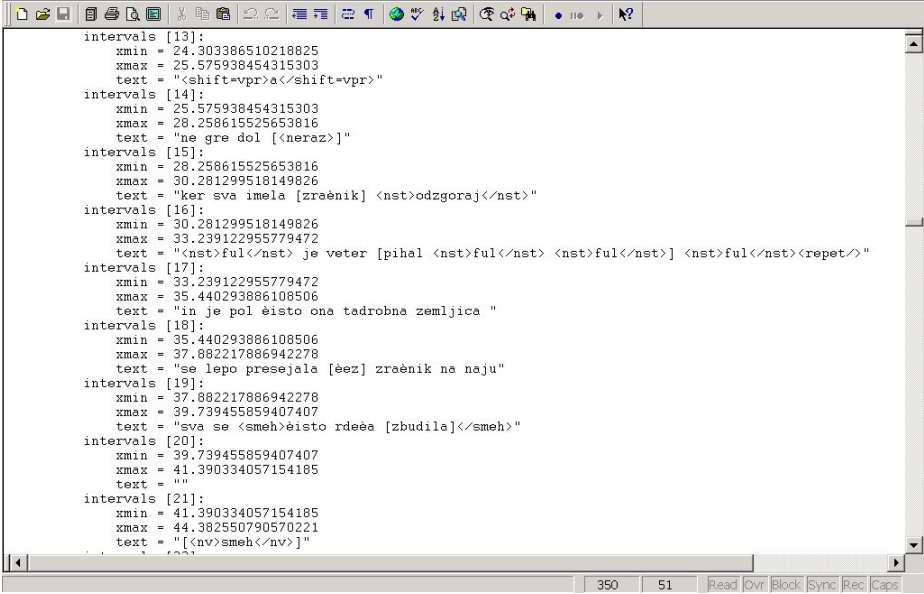
The adopted transcription standard for pilot corpus of Slovene language is presented on the following scheme:

Tag	Meaning
<pavza>	short pause (app. 1 sec)
<pravza>(5)	pause (5 sec)
<ime>	personal name
<priimek>	family name
<priimek><f>	family name, feminine form
<neraz>	unintelligible
<neraz> (5)	unintelligible (5 sec)
<?> text </?>	uncertain transcription
<lz>	false start, truncated word
<repet>text</repet>	repetition
<nst>word</nst>	non-standard word or form
<okr>word</okr>	acronym or abbreviation
[text]	overlapping speech
<singing>text</singing>	paralinguistic markers
<shift=vpr>text</>	part of the text with interrogative intonation
<shift=poud>text</>	emphasised, stressed
<tj: norv>text</tj>	a word or a text, spoken in foreign language
<nv>laughing</nv>	nonverbal events
(description)	non communicative background sound
<??> text</??>	speaker unknown or uncertain

Table 4. Mark-up conventions, used in Pilot Spoken corpus of Slovene

4 Conversion

The conversion of transcriptions into a searchable corpus has been made by Knut Hofland at the The Department of Culture, Language and Information Technology (Aksis) at the University of Bergen. Words are not morpho-syntactically neither syntactically annotated. Transcripts are linked to the digital audio recordings. To align audio and transcript, programs such as *Transcriber* and *Praat* place markers in the transcript that point to precise timings within the sound files.



```

intervals [13]:
  xmin = 24.303386510218825
  xmax = 25.575938454315303
  text = "<shift-vpr>a</shift-vpr>"
intervals [14]:
  xmin = 25.575938454315303
  xmax = 28.258615525653816
  text = "ne gre dol [<neraz>]"
intervals [15]:
  xmin = 28.258615525653816
  xmax = 30.281299518149826
  text = "ker sva imela [zraènik] <nst>odzgoraj</nst>"
intervals [16]:
  xmin = 30.281299518149826
  xmax = 33.239122955779472
  text = "<nst>ful</nst> je veter [pihal <nst>ful</nst> <nst>ful</nst>] <nst>ful</nst><repet>"
intervals [17]:
  xmin = 33.239122955779472
  xmax = 35.440293886108506
  text = "in je pol èisto ona tadbna zemljica "
intervals [18]:
  xmin = 35.440293886108506
  xmax = 37.882217886942278
  text = "se lepo presejala [èez] zraènik na naju"
intervals [19]:
  xmin = 37.882217886942278
  xmax = 39.739455859407407
  text = "eva se <smeh>èisto rdeea [zbudila]</smeh>"
intervals [20]:
  xmin = 39.739455859407407
  xmax = 41.390334057154185
  text = ""
intervals [21]:
  xmin = 41.390334057154185
  xmax = 44.382550790570221
  text = "[<nv>smeh</nv>]"

```

Fig. 3. Time intervals in transcript (WordPad format)

Pilot corpus of Slovene became a part of Corpus Work Bench at Bergen University;² for scientific purposes and noncommercial use it is also available on FidaPLUS web site.³

² <http://torvald.hit.uib.no/talem/jana/s9.html>

³ <http://www.fida.net/slo/index.html>

Fig. 4. Pilot Spoken Corpus of Slovene, search form

5 Corpus Analysis

Building a pilot spoken corpus involved a lot of transcription and annotation work. For 89 minutes of recordings about 100 hours of actual transcription work were needed. Additional time has been spent for many revisions while deciding about transcription standard. The size of the corpus is about 15.000 tokens – words, prosodic and non-linguistics tags. The size is rather limited but it allows to show some examples of corpus use.

To relate the kind of query to the size of corpus, it is best to start with a list of most frequently used objects, such as word forms or tags. The frequencies follow Zipf's Law,⁴ which basically means that about half of them occur only once, a quarter only twice, and so on. There are 3118 word forms in pilot corpus and 2100 among them appears only once, 400 twice, and about 600 word forms have three or more appearances; the most frequent word however has almost 500 appearances. There is very little point in studying words with one occurrence, except in specialized research. If the research is about events which are more complex than just word occurrence, then a limited corpus size will probably disable further investigation.

4 Zipf, G. K. 1935. *The psychobiology of language*. New York: Houghton Mifflin.

1	498	35.422	je
2	425	30.230	ne
3	358	25.464	ə
4	313	22.263	pa
5	297	21.125	in
6	284	20.201	se
7	270	19.205	da
8	268	19.063	to
9	265	18.849	ja
10	264	18.778	v
11	186	13.230	na
12	143	10.171	tudi
13	130	9.247	za
14	115	8.180	ki
15	106	7.540	so
16	105	7.469	tako
17	105	7.469	mhm
18	98	6.971	kaj
19	88	6.259	a
20	86	6.117	še
21	84	5.975	če
22	78	5.548	zdaj

Table 5. Frequency list of a Pilot Spoken Corpus of Slovene

The frequency list shows most frequently used words in pilot spoken corpus together with their absolute and relative⁵ frequency. The most frequent word is “je”, 3rd person singular form of a verb *to be (is)*. The second is a negation word “ne”, meaning *no*, which can also be a discourse marker with no negative connotation. The third most frequently used word is a hesitation voice with mouth half open, “ə”. Among 40 most frequent words in pilot spoken corpus we can find mostly grammatical words, discourse markers and filled pauses. All words need further study for a definition of their (contextual) meanings and discourse functions.

Ten most frequent words from Slovene reference corpus Fida (www.fida.net) appear also on the list of 20 most frequent words in pilot spoken corpus. This can be investigated despite the fact that Fida is a lemmatized corpus and pilot spoken corpus isn't. The fact that the remaining ten words from the top twenty frequency list in pilot corpus do not have an outstanding position on Fida's frequency list can hardly be taken as a surprise, considering that they are mainly discourse markers and filled pauses. The spoken texts of the pilot corpus are accessible in two ways: as whole texts and through concordancer. Sometimes, the whole texts are important for linguistics analysis, however one will search for concordances more often. In pilot spoken corpus we can search for one, two or three words, or part of the words:

⁵ Appearance within 1000 words.

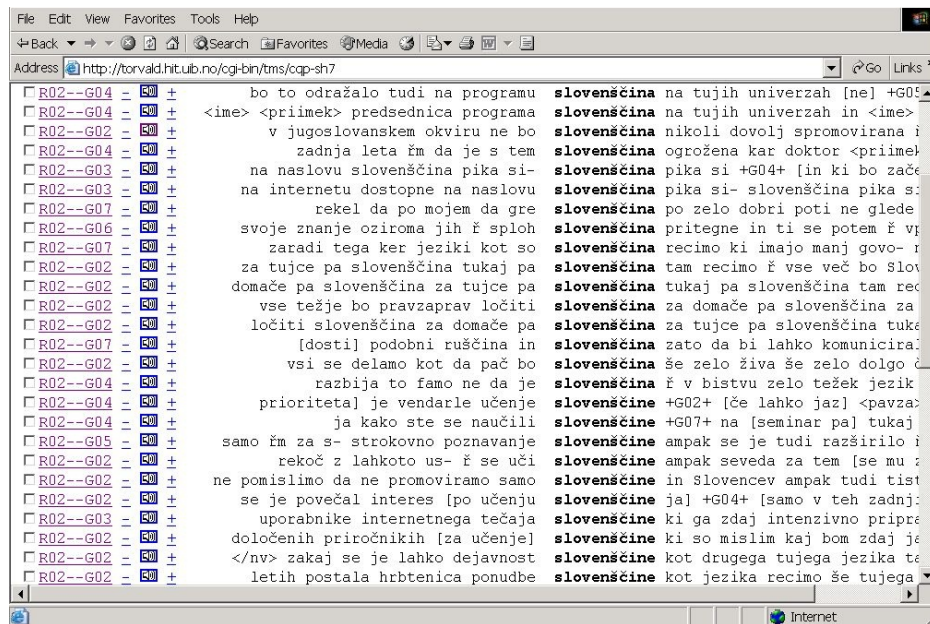


Fig. 5. Pilot Spoken Corpus of Slovene, concordance of a word form “slovenščina”

On fig. 5, the concordance of the word “slovenščina” (Slovene language) can be observed. Each utterance is linked to the actual sound file and attributed with speaker and record identification (i.e. G02, R02). The three special Slovene characters (č, ž, š) presented a problem at one stage of conversion but have finally been properly used in search form.

Collocations are also an important source for language studies, especially in lexicography and language learning. A small sized corpus doesn't enable adequate research comparing to large corpora, however even there some interesting language behavior could be observed. Following case shows a typical syntactic pattern from spoken language. When searching for the word form “vem”, *I know*, which is the most frequent verb word form in pilot corpus, in 90 % of appearances we found it in collocation with “ne”, *no*, what would mean *I don't know*. That doesn't mean that all speakers of the corpus were ignorant – we are rather dealing with weakened meaning of collocation “ne vem” which often appears as a filled pause in a speech.

f0d0v0ju </okr> ali kje že saj	ne vem	ja 0 če bi Slovenci vo
ogorčeni <nst> pizda </nst> to pa	ne vem	ja čisto so ogorčeni k
reče 0 </shift=vpr> pač <pavza>	ne vem	jaz zdaj teh izrazov [
je večno vprašanje in nikoli	ne vem	kaj bi odgovoril [ampak
to tudi z muziko <neraz> +G13+	ne vem	kaj poslušáš ne tako si
+G12+ [<nv> smeh </nv>] +G11+	ne vem	kaj še ja to je to +G13
v sili uporabili tudi jaz	ne vem	kaj študent biologije
aj potem ko ti izpiski hodiyo jaz	ne vem	kako je <nst> potlej
[enaintrideseti] +G16+ [saj	ne vem	kako je] sploh [petek]
ur a veš +G17+ [<neraz>	ne vem	kako jim to uspeva ej]
je zraven tudi poglavje recimo	ne vem	kako se učiti jezik ne
vrste sp- <pavza> ponuja recimo	ne vem	kako se v slovenščini
na <nst> faksu </nst> +G15+ ne vem	ne vem	kako to [izgleda] +G03+
kruh kupit pa bom rekel	ne vem	kako z rokami sem si t
ne more biti ogrožena n-	ne vem	kakšen bi bil 0 razlog
so tukaj navade kakšni 0m	ne vem	kakšen je delovni čas
ker pač nisem +G16+ [ja]	ne vem	kakšna [delitev] je to
<neraz> kazalo še drugačne oblike	ne vem	kakšne tedenske delavn
so <neraz>] +G20+ ja enkrat pa	ne vem	kdaj je bila je pa nise
ja] 0m predsednik uprave jaz	ne vem	kdo <neraz> saj <nst> n
ja <tj> deck </tj> [to] stane	ne vem	koliko sedemnajst
se je malo zajebaval ah kaj jaz	ne vem	ne saj je možno da ima
edino] na <nst> faksu </nst> +G15+	ne vem	ne vem kako to [izgled
kar se literature tiče ampak	ne vem	ne vem <repet> potem im
smeh </nv> to je tako <neraz>	ne vem	no on je meni <pavza>
<pavza> pa tudi drugače kaj pa	ne vem	no <pavza> <shift=vpr>
mene vprašáš] +G17+ [ej	ne vem	o čem <neraz>]
hočeš <neraz> vse imeti +G11+ iz	ne vem	od <lz> iz [<?> Obsess
Rupel je car Ru- Rupel je	ne vem	on je tak svetovljan on
mislim v0- primerljiv seveda z	ne vem	s francoskim inštitutom
vsaj kar se tiče jezika	ne vem	s te strani se mi zdi [
Irak <nv> pihne skozi nos </nv> saj	ne vem	saj to bo ne pa malo bo
gori po hribih mogoče je kaj jaz	ne vem	samo +G03+ [aha] +G01+
[no že ampak] [kolikor jaz	ne vem	se deli] v slovenščini
[ane] <shift=vpr> kako je	ne vem	Shakespeare </shift=v
</tj> si pač +G13+ oprijet pa to	ne vem	srajca pa to <nst>
vem kaj poslušáš ne tako si potem	ne vem	te <tj> punk </tj> si p
še tiho v slovenščini ali	ne vem	v katerem jeziku so bili

Fig. 6. Part of collocations list “ne vem”, I don't know

The discourse marker “mhm” has, as expected, very high absolute frequency (105) comparing to its absolute frequency in ten thousand times bigger corpus Fida (156). With the pilot corpus we could argue the explanation of a meaning of the word “mhm” in Slovene standard dictionary: it is explained as a word of hesitation or a word of restrained agreement. In the pilot corpus, we couldn't find but one example to prove that explanation among 105 mhms, however some highly represented meanings should be added to the explanation in the dictionary.

[mhm]	<nv> smeh </nv>	[mhm]	+G07+ [ampak] ø druga
[mhm]	v katalonščini pa ne morem	[mhm]	+G07+ [dobesedno] sede
[mhm]	državi kar se tega tiče ne	[mhm]	+G07+ [s] tega vidika
[mhm]	v ozadju) +G03+ [øm] <pavza>	[mhm]	+G09+ [hitro] eno kart
[mhm]	kot vidite na Atlantiku ne) +G01+	[mhm]	+G09+ [mhm] +G10+ [mhm]
[mhm]	ne moreš v bistvu ti dobiti +G15+	[mhm]	+G09+ in plače in
[mhm]	ne] +G01+ [mhm] +G09+	[mhm]	+G10+ [mhm] +G15+ [mhm]
[mhm]	[mhm] +G09+ [mhm] +G10+	[mhm]	+G15+ [mhm] +G03+ obil
[mhm]	časa za [druženje] +G04+	[mhm]	[ampak tista prioritete
[mhm]	v Slovenijo preko Moskve ne	[mhm]	[ampak ø] +G02+ [<nv>
[mhm]	vidika [ne] hø +G02+	[mhm]	[ja <smeh> doktor
[mhm]	[bo že torej mogoče]	[mhm]	[ja] +G02+ [ø]
[mhm]	mi zdi [da] +G02+	[mhm]	[mhm] [ø doktorica <im
[mhm]	[s Slovenci] +G07+	[mhm]	[mhm] øm kaj vam je to
[mhm]	usposobljena za slovenščino +??+	[mhm]	[mhm] øm <ime>
[mhm]	[da] +G02+ [mhm]	[mhm]	[ø doktorica <ime>
[mhm]	na seminarju [ane] +G03+	[mhm]	ampak res predvsem v
[mhm]	kaj bom zdaj jaz +G04+	[mhm]	bo mogoče <smeh> še o
[mhm]	čez [štirideset] let +G04+	[mhm]	bog ve kaj bo ne jaz
[mhm]	funkcijo] čeprav je res +G04+	[mhm]	da je potreba po [neče
[mhm]	dejansko je pa res +G04+	[mhm]	da je- ø bo treba po
[mhm]	mi včasih smo to +G04+	[mhm]	doživljali tako ne do
[mhm]	[dežela] za nas +G04+	[mhm]	drugače jaz sem študir
[mhm]	[perspektive] predstaviti +G04+	[mhm]	in to je uspelo ne tud
[mhm]	naredil in tako naprej +G04+	[mhm]	in zato pripravljamo
[mhm]	konkretno lepo ja +??+	[mhm]	ja hvala [<nv> smeh
[mhm]	<repet/> patetičen [ne] +G04+	[mhm]	ker je pa tako ne kar
[mhm]	in [drugi] ne +G04+	[mhm]	ker v bistvu ø en del
[mhm]	še dodatni [študij] +G02+	[mhm]	mhm doktor <ime>
[mhm]	v Kataloniji [ne] +G04+	[mhm]	mi smo se srečali s t
[mhm]	<ime> +G15+ je pa obilo dežja <pavza>	[mhm]	ne v Bergnu ne +G03+ a
[mhm]	in slovenščini [ne] +G04+	[mhm]	no kot je bilo že prej
[mhm]	<repet/> naj bi bil razlog	[mhm]	potem ø če lahko govo
[mhm]	[drugi] del publike +G04+	[mhm]	recimo ki študira
[mhm]	govo- n- [govorcev] +G04+	[mhm]	saj v končni fazi ne
[mhm]	posebej [dopovedovati] +G04+	[mhm]	torej seminar je zasn
[mhm]	[ljudi] v +G04+	[mhm]	tujih državah da bodo
[mhm]	od ponedeljka do petka ne ja	[mhm]	<neraz> pa naši semen
[mhm]	življenje [<neraz>] +G04+	[mhm]	<pavza> [Slovenija] in
[mhm]	kot za eno državo +??+ [mhm]	[mhm]	<pavza> mislim in to
[mhm]	precej na delovnem mestu +G04+	[mhm]	ø doktor <ime> <priim
[mhm]	razsežnosti ø [Slovenije] +G04+	[mhm]	ø je pa razlika
[mhm]	kulture +G07+ [seminarja]	[mhm]	ø ki prihaja iz
[mhm]	oziroma] morajo biti +G04+	[mhm]	ø na razpolago tako da
[mhm]	se mi zdi] ne +G04+	[mhm]	ø no morda v tem
[mhm]	</nv>] [ne] +G04+	[mhm]	ø no saj ø zanimivo bi

Fig. 7. Pilot Spoken Corpus of Slovene, concordance of a word "mhm"

6 Perspectives

Despite many deficiencies, which derived mostly from limited time, human and financial resources, the pilot spoken corpus of Slovene serves its original purpose. That was mainly to set the criteria for the collection, selection and documentation of spoken materials, develop and test transcription and mark-up conventions, and finally show some possibilities for a corpus use. Pilot spoken corpus, available in searchable form with transcriptions linked to sound files, is available for language research community. It will now serve as a resource for redefinition of sampling methods and transcription and mark-up criteria. Improvements will be taken into considerations when, in the indefinite future, a construction of 5 million word spoken corpus of Slovene becomes a reality.

References

1. BIBER, Douglas, 1993: Representativeness in Corpus Design. *Literary and Linguistics Computing* 8/4. 243–257.
2. CROWDY, Steve, 1993: Spoken Corpus Design. *Literary and Linguistics Computing* 8/4. Oxford University Press. 259–265.
3. CROWDY, Steve, 1994: Spoken Corpus Transcription. *Literary and Linguistics Computing* 9/1. Oxford University Press. 25–28.
4. Developing Linguistic Corpora: a Guide to Good Practice, 2005. Edited by Martin Wynne.
<http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm>
5. Eagles Preliminary recommendations on Corpus Typology, 1996.
<http://www.ilc.cnr.it/EAGLES96/corpusstyp/node15.html>
6. GORJANC, Vojko, 2005a: *Uvod v korpusno jezikoslovje* (Zbirka Zrenja). Domžale, Založba Izolit.
7. LEECH, Geoffrey, Greg MYERS in Jenny THOMAS (ur.), 1995: *Spoken English on Computer. Transcription, mark-up and application*. New York: Longman Publishing.
8. SINCLAIR, John, 2004: Corpus and Text – Basic Principles. In *Developing Linguistic Corpora: a Guide to Good Practice*. Edited by Martin Wynne.
9. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>
10. THOMPSON, Paul, 2004: Spoken language corpora. In *Developing Linguistic Corpora: a Guide to Good Practice*. Edited by Martin Wynne.
11. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter5.htm>

Appendix



Attribution-NoDerivs 2.5

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN “AS-IS” BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE (“CCPL” OR “LICENSE”). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

- a. **“Collective Work”** means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.
- b. **“Derivative Work”** means a work based upon the Work or upon the Work and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a

Collective Work will not be considered a Derivative Work for the purpose of this License. For the avoidance of doubt, where the Work is a musical composition or sound recording, the synchronization of the Work in timed-relation with a moving image (“synching”) will be considered a Derivative Work for the purpose of this License.

- c. **“Licensor”** means the individual or entity that offers the Work under the terms of this License.
- d. **“Original Author”** means the individual or entity who created the Work.
- e. **“Work”** means the copyrightable work of authorship offered under the terms of this License.
- f. **“You”** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

2. Fair Use Rights. Nothing in this license is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

- a. to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;
- b. to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works.
- c. For the avoidance of doubt, where the work is a musical composition:
 - i. **Performance Royalties Under Blanket Licenses.** Licensor waives the exclusive right to collect, whether individually or via a performance rights society (e.g. ASCAP, BMI, SESAC), royalties for the public performance or public digital performance (e.g. webcast) of the Work.
 - ii. **Mechanical Rights and Statutory Royalties.** Licensor waives the exclusive right to collect, whether individually or via a music rights society or designated agent (e.g. Harry Fox Agency), royalties for any phonorecord You create from the Work (“cover version”) and distribute, subject to the compulsory license created by 17 USC Section 115 of the US Copyright Act (or the equivalent in other jurisdictions).
- d. **Webcasting Rights and Statutory Royalties.** For the avoidance of doubt, where the Work is a sound recording, Licensor waives the exclusive right to collect, whether individually or via a performance-rights society (e.g.

SoundExchange), royalties for the public digital performance (e.g. webcast) of the Work, subject to the compulsory license created by 17 USC Section 114 of the US Copyright Act (or the equivalent in other jurisdictions).

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats, but otherwise you have no rights to make Derivative Works. All rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

- a. You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this License, and You must include a copy of, or the Uniform Resource Identifier for, this License with every copy or phonorecord of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this License or the recipients' exercise of the rights granted hereunder. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this License Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this License. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any credit as required by clause 4(b), as requested.
- b. If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or Collective Works, You must keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or (ii) if the Original Author and/or Licensor designate another party or parties (e.g. a sponsor institute, publishing entity, journal) for attribution in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; the title of the Work if supplied; and to the extent reasonably practicable, the Uniform Resource Identifier, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE MATERIALS, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

- a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Collective Works from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.
- b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

- a. Each time You distribute or publicly digitally perform the Work, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.
- b. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this

agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

- c. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
- d. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, neither party will use the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time.

Creative Commons may be contacted at <http://creativecommons.org/>.

Computer Treatment of Slavic and East European Languages

Editor
Radovan Garabík

Návrh obálky: Vladimír Benko
Zodpovedný redaktor: Emil Borčín
Technický redaktor: Peter Luciak
Prvé vydanie. Vydala VEDA, Vydavateľstvo Slovenskej akadémie vied,
v Bratislave roku 2005 ako svoju 3580. publikáciu z tlačových podkladov
Jazykovedného ústavu Ľ. Štúra SAV. 246 strán.

ISBN 80-224-0895-6