Practical Applications of Language Corpora
2014

University of Lodz

20-22 November 2014
Contents

1 Abstracts 7

A Corpus-Based Study on Attitudinal Stance in Doctoral Disserta-
tions of Native and Non-Native Speakers of English

Reyhan Ağcan, Kahramanmaras Sütçü İmam University . . . 7

A learner corpus investigation on the use of negative and positive
vocabulary by EFL learners

M. Pınar Babanoglu, Kahramanmaras Sütçü İmam University
Cem Can, Çukurova University  . . . . . . . . . . . . . . . . . . . 8

Questions of context, culture, and genre in identifying semantic
prosody

Shala Barczewska, Jan Kochanowski University  . . . . . . . . 9

Researching translation with comparable corpora: the Eurofog project

Łucja Biel, University of Warsaw . . . . . . . . . . . . . . . . . . 11

Challenges in Processing Real-Life Speech Corpora

Łukasz Brocki, Danijel Koržinek, Krzysztof Marasek, Polish-
Japanese Institute of Information Technology . . . . . . . . . . . 13

Annotation tools and interfaces for managing transcription of old
Polish texts

Renata Bronikowska, Institute of the Polish Language, Polish
Academy of Sciences
Włodzimierz Gruszczyński, Warsaw School of Social Sciences
and Humanities
Dorota Adamiec, Institute of the Polish Language, Polish Academy
of Sciences
Michał Lenart, Institute of Computer Science, Polish Academy
of Sciences
Małgorzata Majewska, Institute of the Polish Language, Polish
Academy of Sciences
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy
of Sciences . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

The position of objects in Old English and Old High German trans-
lations

Anna Cichosz, University of Lodz  . . . . . . . . . . . . . . . . . . . 18

Time-discretising adverbials in reference corpora: distributional evi-
dence of conceptualisation patterns

Mikołaj Deckert, Piotr Peziik, University of Lodz . . . . . . . 19

Measuring the distribution of mass and count nouns. A comparison
between a rating study and a corpus based analysis.

Francesca - Chiara - Giorgio Franzon - Zanini - Arcara,
Università degli Studi di Padova . . . . . . . . . . . . . . . . . . . 20

Evidence from a Learner Corpus: Can a Test of Vocabulary Recog-
nition Predict Vocabulary Use?

Roger Gee, Holy Family University . . . . . . . . . . . . . . . . . . . 22
Corpus-Based Study of Polish Public Procurement Law in English Translation

**Justyna Giczel-Pastwa**, University of Gdansk  
24

A Corpus-driven study of the use and discourse functions of lexical bundles in Polish patient information leaflets

**Łukasz Grabowski**, Lazarski University  
**Paweł Bechler**, Warsaw University  
26

Mapping corpora: Approaches to understand the geographies in texts

**Ian Gregory**, Lancaster University  
28

Temporality and language contact. A corpus-based study of the language of the Russian-speaking immigrants in Germany.

**Christine Grillborzer**, Slavisches Seminar Universität Freiburg  
29

Semantic changes in English affixes – a corpus-based study

**Marta Grochocka-Sidorowicz**, WA AMU  
30

No need for rules in the -ez/-eza Spanish nominalization

**Matías Guzmán Naranjo**, Universität Leipzig  
32

Valency: Towards a comprehensive valence dictionary of Polish

**Elżbieta Hajnicz**, Agnieszka Patejuk, Adam Przepiórkowski, Marcin Woźniński, Institute of Computer Science, Polish Academy of Sciences  
34

Comparing tense and aspect forms across the world’s languages

**Martin Haspelmath**, Max Planck Institute for Evolutionary Anthropology, Leipzig  
36

The NOW and Consciousness: a Comparison

**Dwight Holbrook**, Adam Mickiewicz University in Poznan  
37

Towards a road map for exploring evaluative meaning

**Susan Hunston**, University of Birmingham  
39

Evaluating threats in the press

**Jane Helen Johnson**, University of Bologna  
40

Heuristic Detection of Zero Subjects for Polish

**Adam Kaczmarek**, University of Wroclaw / Wroclaw University of Technology

Michał Marcińczuk, Wroclaw University of Technology  
42

Corpus-based Analysis of Czech Units Expressing Mental States and Their Polish Equivalents. Identification of Meaning and Establishing Polish Equivalents Referring to Different Theories (Valency, Pattern Grammar, Cognitive Linguistics)

**Elżbieta Kaczmarska**, University of Warsaw  
44

CL for CDA: What can we do better?

**Victoria Kamasa**, Adam Mickiewicz University in Poznan  
45

Corpus tools and resources in lexicography

**Mariusz Kamiński**, Panstwowa Wyższa Szkoła Zawodowa w Nysie  
47
Automatic keyphrase extraction for documents in Polish - comparison of supervised and unsupervised approaches  
**Paweł Kędzia**, Henryk Maciejewski, Marcin Oleksy, Błażej Ryśnik, Wroclaw University of Technology 48

A corpora-driven study of social class in Britain (a case-study of abstract nouns)  
**Natalia Khokhlova**, MGIMO-University 50

CorpoGrabber: The Toolchain to Automatic Acquiring and Extraction of the Website Content  
**Jan Kocoń**, Wroclaw University of Technology 52

Avenues for Corpus-based Research on Informal Spoken Czech  
**Marie Kopřivová**, David Lukeš, Zuzana Komrsková, Petra Klimešová, Charles University 53

Distance, knowledge and axiology in TV news discourse  
**Monika Kopytowska**, Distance, knowledge and axiology in TV news discourse 55

Time as a line and as a clock in signed languages: a view from cognitive linguistics  
**Krzysztof Kosecki**, University of Lodz 57

Feminine derivatives in National Corpus of Polish  
**Katarzyna Krasnowska**, Polish Academy of Sciences 58

Language Change in Video Game Instructions? - Testing Frequency Effects in Corpus and Experimental Data  
**Anne Krause**, University of Freiburg 59

Hyperspace representations and free associations: comparison of two methods in a study of loan-word and native-word synonyms  
**Adrianna Kuciuńska**, University of Warsaw 61

Temporal and aspectual properties of events: degrees of deverbalization in Polish and English modifiers  
**Barbara Lewandowska-Tomaszczyk**, University of Lodz 65

The persistence of stereotypes. A corpus-based analysis of the representation of rheumatoid arthritis in press discourse  
**Maria Libura**, Lazarski University 67

Evidentiality in Mandarin narrative discourse: A corpus-driven approach  
**Lu Lu**, SOAS 69

Standardization in safety data sheets? A corpus-based study into the problems of translating safety documents  
**Aleksandra Makowska**, University of Lodz 71
A corpus perspective on the football played in Brazil and the U.K  
Sabra Matuda, University of São Paulo .............................. 72

Evaluating the Libya crisis: through the eyes of journalists writing in English and Italian  
Amanda C. Murphy, Universita Cattolica del Sacro Cuore .... 74

Studying ellipsis alternation in English corpus data  
Joanna Nykiel, University of Silesia ................................. 76

Linguistic analysis of weblog genres  
Marcin Oleksy, Jan Kocóñ, Wrocław University of Technology  
Maciej Maciej, Institute of Literary Research, Polish Academy of Sciences  
Maciej Piasni, Wrocław University of Technology ............... 79

Temporal Expressions in Polish Corpus KPWr  
Jan Kocóñ, Michał Marcińczuk, Marcin Oleksy, Tomasz Bernañ, Michal Wolski .................................................. 82

Light Verb Constructions in TICLE and PICLE: How should we accelerate their lexical priming in EFL Context?  
Katarzyna Papaja, University of Silesia  
Cem Can, Abdurrahman Kilimci, Cukurova University ........... 83

Chronological Corpus of Polish Press Texts (1945–1954)  
Adam Pawłowski, Wrocław University ............................... 85

Exploring prosodic patterns of discourse devices  
Piotr Peñzik, University of Lodz  
Ewa Guz, John Paul II Catholic University of Lublin ............. 87

CLARIN-PL – a Polish Language Technology Infrastructure for the Users  
Maciej Piasni, Wrocław University of Technology ............... 89

A Multilayer System of Lexical Resources for Language Technology Infrastructure  
Maciej Piasni, Marek Maziarz, Michał Marcińczuk, Pawel Kędzia, Adam Radziszewski, Wrocław University of Technology . 91

Literary Criticism of Cartography and Indexing of Spatial Data and Semantic Relations  
Maciej Piasni, Wrocław University of Technology ............... 94

Adapting the Louvain Error Tagset for the annotation of the written component of the PELCRA Learner English Corpus  
Agata Pieszczek, University of Warsaw ................................ 96

PoS-grams: an extra tool for the corpus linguist’s kit?  
Antonio Pinna, David Brett, Università degli Studi di Sassari .... 98

The concept of a combinatorial dictionary and its multi-stage verification  
Adam Przepiórkowski, Institute of Computer Science, Polish Academy of Sciences .................................................. 100
A morphosyntactic approximation of semantic roles in a parser of Polish

Adam Przepiórkowski, Wojciech Jaworski, Institute of Computer Science, Polish Academy of Sciences .............................. 101

A Lexicographic Tool for Recognising Polish Adjective Classes

Radosław Ramocki, Mieszko Służewski, Marek Maziarz, Wrocław University of Technology .............................. 104

Brazilian food names in English: a corpus-based study

Rozane Rebechi, University of São Paulo .............................. 106

Verbs and argument structure constructions in specialized languages

Anna Riccio, University of Naples "L'Orientale" .............................. 108

Introducing a corpus of non-native Czech with automatic annotation

Alexandr Rosen, Charles University .................................. 110


Stanisław Goźdż-Roszkowski, University of Lodz .............................. 112

Large Polish-English Lexico-Semantic Resource Based on plWordNet - Princeton WordNet Mapping

Ewa Rudnicka, Maciej Piasecki, Wrocław University of Technology

Wojciech Witkowski, University of Wroclaw .............................. 113

A corpus-based comparative study of lexical choices in Middle English dialects

Paulina Rybińska, University of Lodz .............................. 115

Report on Teacher Trainees’ use of their own Graded Reader Corpus

James Thomas, Masaryk University .................................. 117

Evaluative folk metalanguage in the Polish corpus

Jerzy Tomaszczyk, University of Lodz .............................. 119

Time and space in distance representations: a corpus based study

Jacek Walliński, University of Lodz .............................. 120

Effective architecture of web services for Polish language processing

Tomasz Walkowiak, Marcin Pol, Wroclaw University of Technology .............................. 122

Corpus-based approach to finding targets of evaluations in Polish

Aleksander Wawer, Institute of Computer Science, Polish Academy of Sciences .............................. 124

Large Scale Processing of Polish Political Tweets

Aleksander Wawer, Institute of Computer Science, Polish Academy of Sciences

Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Science

Mateusz Kopeć, Institute of Computer Science, Polish Academy of Science .............................. 126
Morfeusz — the new prospects

**Marcin Woliński**, Michał Lenart, Witold Kieraś, Małgorzata Marciniak, Piotr Rychlik, Institute of Computer Science, Polish Academy of Sciences .................................................. 128

Verbalization patterns in learners’ constructed responses to German listening comprehension items

**Magdalena Wolska**, Eberhard Karls Universität Tübingen .................................. 131

Towards an Events Annotated Corpus of Polish

**Michał Marcińczuk**, Marcin Oleksy, Jan Kocoń, Tomasz Bernaś, Michał Wolski ................................................................. 132

Force Dynamics and Time

**Jacek Wozny**, Wrocław University ................................................................. 134

Poliqarp 2: a corpus lookup system for multiple annotation layers

**Bartosz Zaborowski**, Dominika Pawlik, Aleksander Zabłocki, Institute of Computer Science, Polish Academy of Sciences ........................................ 135

Surviving across centuries: investigating the use of the enclitic -t’ in the 19th century Czech

**Anna Zitová**, Charles University ................................................................. 137
Abstracts

A Corpus-Based Study on Attitudinal Stance in Doctoral Dissertations of Native and Non-Native Speakers of English

Reyhan Ağçam, Kahramanmaraş Sütçü İmam University

Although objective voice is extensively considered among the fundamentals of academic writing, author stance, which is defined as “the way academics annotate their texts to comment on the possible accuracy or creditability of a claim, the extent they want to commit themselves to it, or the attitude they want to convey to an entity, a proposition or the reader” (Hyland, 2005), seem to be involved in the production of scientific research papers. That is, no matter how authors, especially those who work in the field of social sciences, endeavour to employ impersonal expressions for the sake of objectivity, they inevitably adopt a certain stance while reporting their findings and sometimes even making suggestions for further research. This study investigates the use of attitudinal stance devices in Academic English by native and non-native academic authors of English. Being corpus-based in design, it will report the results of the Contrastive Interlanguage Analysis (Granger, 1996) administered to a total number of 136 doctoral dissertations written by Turkish-speaking, Spanish-speaking and native academic authors of English. Frequencies of attitudinal stance devices were separately calculated for each corpus and a log-likelihood test was applied to see whether native and non-native academic authors of English significantly differ with respect to the use of these devices.
A learner corpus investigation on the use of negative and positive vocabulary by EFL learners

M. Pinar Babanoglu, Kahramanmaraş Sütçü İmam University
Cem Can, Çukurova University

Relationship between gender and language has been a challenging inquiry as being closely related to social issues. Particular interest has mainly been centered around whether women and men use the language differently. As a factor of sociolinguistics, most of earlier literature on gender and language studies has concentrated on the identification and characterization of male/female language and how language users speak and write in ways that reflect their genders. In the literature, relationship between gender and language has mostly concentrated on the identification and characterization of male/female language and how language users speak and write in ways that reflect their genders. For instance, Lakoff (1975) suggests some diversifications such as women use more tag questions, more correct grammar, more intensifiers and even more empty adjectives than man such as ‘charming’ and ‘nice’. Holmes (1998) draws attention to a range of sociolinguistic universal keynotes differing in gender such as; women and men develop different patterns of language, women tend use linguistic devices that stress solidarity more often than men do and women are stylistically more flexible than man.

This study investigates the lexical uses of positive words such as ‘happy’, ‘believe’, ‘joy’, etc. and negative words like ‘awful’, ‘ugly’ and ‘fail’ etc. in written productions of female and male EFL learners. Major aim is to examine whether there are quantitative (frequency) and qualitative (choice of words) differences in the use negative and positive words between male and female EFL learners and also between male and female native English speakers. Two types of comparisons as of male/female and learner/native speaker are emphasized to highlight the possible differences between and within gender and speaker groups. Methodology of the study depends on Contrastive Interlanguage Analysis that involves comparing learner data with native speaker data or comparing different types of non-native speaker or learner data. TICLE as learner corpus and LOCNESS as native English speaker corpus are main data sources of the study. Data collection is based upon identification of selected words throughout corpora by Wordsmith tools. Frequency counts and log-likelihood measurement are utilized for quantitative analysis process.
Questions of context, culture, and genre in identifying semantic prosody
Shala Barczewska, Jan Kochanowski University

The role of collocation in determining a word’s connotational meaning, or semantic prosody, has a long history in corpus linguistics and discourse analysis. For many researchers, collocation has seemed almost a magical tool, revealing that words such as dealings (Partington, 1998, pp. 72-74) or regime (Sinclair, 2003, pp. 17-18) could reflect disapproval or irony (cf. Louw 2000).

Others, however, questioned the universality of its application (cf. Bednarek, 2008). One of the concerns is how to identify ‘positive’ or ‘negative’ evaluation in an objective manner. Dilts and Newman suggest using the 76 adjective pairs evaluated in a 1950’s study by Osgood, Suci and Tannenbaum (1957:47-64). In their study, these adjective pairs were given correlation scores based on respondents’ assessments along different evaluative scales, among which was the notions of ‘good’ and ‘bad’ (Dilts and Newman 2006, p. 33).

While on the surface such a strategy may seem to provide an objective standard, it is still based on subjective opinion. The only difference is that in using the 76 word pairs of Osgood et al. the evaluations are provided by several subjective participants, in this case students (1957). Furthermore, their project did not take into consideration culture, context and genre, all of which play an important role in determining a word’s connotation in a particular situation (Bednarek 2008:122-125).

To illustrate the limits of such an approach, this study looks believing - sceptical, which Dilts and Newman cite Osgood, et al. as correlating with ‘good’ - ‘bad’ respectively (with a score of 0.38, or in twentieth position out of the seventy-six pairs). Modern use of these adjectives is analyzed in the context of a corpus focused on evolution education in the United States. It shows that the connotation of the words is highly dependent on context and genre and may even have shifted since the 1950s. Hence it argues for a cautious approach to applying such scales to questions of semantic prosody in discourse analysis.

References

Researching translation with comparable corpora: the Eurofog project
Łucja Biel, University of Warsaw

The paper discusses the potential and limitations of the comparable corpus methodology in translation studies. The discussion is illustrated by the Eurofog project, a first large-scale empirical study into the Polish eurolect and its impact on the Polish language of the law.

The comparative-corpus methodology was proposed by Mona Baker in her pioneer paper (1993), to study the distinctive features of translated language against nontranslated language. The method involves the comparison of two monolingual corpora, one with translated texts and the other with corresponding nontranslated text. The existing studies have raised questions as to the legitimacy of research into translations without any reference to underlying source texts. Recently Baker’s method has been modified into multilingual comparative corpora (cf. Hansen-Schirra and Teich, 2009: 1162), which - in addition to the fundamental monolingual comparative corpora - also includes a parallel corpus of source and target texts.

The paper presents an application of Baker’s modified method in the Eurofog project. The study was built around two research questions: (1) how does the language of translated EU law differ from the language of Polish law (the textual fit of translation) and 2) are the differences correlated with the post-accession changes in the language of the Polish national law (2011) compared to its pre-accession version (1999) (the Europeanisation of the Polish language of the law). The study was conducted on translation corpora (JRC Acquis corpus of regulations and directives translated into Polish), the English version of the JRC Acquis corpus, legal comparable corpora (the Polish Law Corpus) and general reference corpus (NKJP).

The paper will first present the type of data and findings that may be obtained via this method:

- In respect of the textual fit, the data show significant differences in the distribution of lexicogrammatical patterns in the corpora, including realization of deontic modality, over-representation of passives and participles, realization of the prototypical patterns of legal reasoning, asymmetrical distribution of synonymous pairs, underrepresentation of some complex adverbials. The low textual fit of translated legislation creates a distinct, more ‘European’ genre of legal Polish.

- In respect of Europeanisation, the large inflow of translated law and the harmonisation of Polish law with EU law have overall had a relatively little effect on the grammatical layer of national legislation. The generic features of national law tend to be relatively stable; the changes usually involve an increase/decrease of frequency. The general trend is a shift towards a flatter sentence structure with a smaller depth of embeddings due to the increased sentence length and the reduced number of sentences, accompanied by the reduced number of
finite hypotactic clauses on the one hand, and an increase of present adverbal participles and some complex prepositions which require nominalisations, on the other hand. The most spectacular change is the increase of textual mapping patterns for intra- and intertextual references. The layer which has been more profoundly affected is terminology.

The paper will conclude with the assessment of usefulness of this method to study translations and with suggestions for further research.

References

Challenges in Processing Real-Life Speech Corpora
Łukasz Brocki, Danijel Koržinek, Krzysztof Marasek, Polish-Japanese Institute of Information Technology

This paper describes various tools for speech processing created in the CLARIN-PL project for the purpose of augmenting the existing speech corpora with additional levels of annotation. The tools in question are split into three categories: text processing tools, simple signal analysis tools and more sophisticated model-based speech analysis tools. All the tools are based on well known techniques using state-of-the-art, well established algorithms. Unfortunately, all of them have serious performance issues when faced with real-life noisy data. This paper describes the tools and their limitations, while also discussing some of the techniques that can be used to mitigate the problems as much as possible. As a demonstration, the tools are applied to some of the databases developed in the project by other team members: a collection of historic articles and the Spokes database containing recordings of real-life spontaneous conversations. Finally, possible uses of the annotation derived using these tools is discussed. Following is a short description of the tool categories. Before analyzing speech, all the related text has to be phonetized. Phonetization is a process that converts text from the written into its spoken form. While it can seem like a straightforward process, it often turns out to be quite challenging, even for Slavic languages. First of all, the text needs to be normalized, meaning that all the numbers and abbreviations need to be expanded into their full spoken form. This step is especially complicated in inflected languages like Polish, where the expansion is highly dependent on the grammatical context. After that, we are still faced with many exceptions in pronunciation, for example in words of foreign origin. Normalization is not a widely discussed topic and there are few tools available. Our approach relies on rules for expansion and Statistical Language Models for disambiguation. Our phonetization is based strictly on a rule-based system, but experiments with a statistical system show that it could be more reliable with names and words of foreign origin. There are several basic signal processing methods useful for analyzing speech data. Pitch, tempo and energy are the three basic parameters affecting prosody, which is an important concept in analyzing both the low-level syntactic as well as higher level semantic and pragmatic content of speech. Pitch and energy are usually derived using direct signal processing algorithms, while the tempo is measured using information described in the following section. While the aforementioned signal processing algorithms are deterministic in nature, they can still underperform when faced with real-life noisy data. Speech modelling tools are based around Large-Vocabulary Continuous Speech Recognition and Speech Alignment. Speech corpora can be divided into three groups: speech data with an accurate transcription of its contents, noisy data or data with an inaccurate transcription and data without any form of transcription. For all types of corpora, we would like to derive an accurate time alignment of text description on several levels (phrases, words, syllables, phonemes), in other
words “what is spoken when”. This allows us to easily index and search through data, which would otherwise be impossible. It also provides us with a mean of computing the tempo measure as mentioned in the previous section. For the first kind of corpora, this task is much easier to solve by using well known time-alignment algorithms. If we are lacking transcriptions, we need to apply complete speech recognition, which will obviously perform much worse. To deal with the intermediate cases, a special combination of speech-recognition and text-text alignment is used.
Annotation tools and interfaces for managing transcription of old Polish texts
Renata Bronikowska, Institute of the Polish Language, Polish Academy of Sciences
Włodzimierz Gruszczyński, Warsaw School of Social Sciences and Humanities
Dorota Adamiec, Institute of the Polish Language, Polish Academy of Sciences
Michał Lenart, Institute of Computer Science, Polish Academy of Sciences
Małgorzata Majewska, Institute of the Polish Language, Polish Academy of Sciences
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences

1. Introduction
Despite many other available digital work environments for transcription, editing and publication of old texts, such as Oxygen-based Ediarum (http://www.bbw.de/en/telota/software/ediarum) or an interesting Web-based Transcription Desk of Transcribe Bentham Initiative (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham), their excessiveness and/or on-line character made them unsuited to a simple task of transcription and rough annotation of structure and language in an old Polish text, undertaken currently in a project of diachronic extension of the National Corpus of Polish [1]. This led to development of a new approach, more convenient to an average text editor users and at the same time reusing familiar editing software while making it independent on being on-line all the time. Based on previous experience, a word-processor-based setting customized with style definitions and keyboard shortcuts has been prepared.

2. Transcription management
The transcription process involves three user roles: editor — initiator and manager of the process, transcriber — providing a transcribed, structurally and lexically annotated texts, and proofreader — verifying the annotation. It starts with creation of a minimal metadata description of a new text and assigning it to a transcriber to maintain relationship with a regular set of experts for the whole life of the text. Transcribers prepare the transcription using Microsoft Word, save resulting documents in RTF (Rich Text Format, a popular cross-platform format for document interchange) and upload them via Web browser to a dedicated site. Editors forward the text to proofreading or send it back to transcription if errors are found. The workflow ends when the text is ready to be included in the corpus.

The environment GUI is modelled after Drupal content management platform with two basic views:

- list view showing all texts being processed and providing interface for executing actions (changing text workflow state, dis-
playing its metadata and current content etc./)

• item view presenting interface to managing metadata of an individual text.

On upload texts are converted to TEI P5 XML [2]. HTML presentation is also available in the system for easier review.

3. Style-based markup
Microsoft Word was selected as on-line editing environment due to its popularity in the humanities in Poland, but simplicity of the editing styles- and shortcuts- based environment makes it available for integration with word processing components from competitive office suites. Both character and paragraph styles were used to produce pseudo-XML, with markers (such as illustration placeholders) and structural markup (such as section structure) represented with pseudo- tags (texts in square brackets such as [EQUATION] or [DOCUMENT BEGIN] and [DOCUMENT END]) and single-level structures represented with character styles. Languages, being vertical to other types of markup, are represented with text colour (thus making it possible to mark up e.g. a Latin interjection in a text title). Formatting is used to visually distinguish different styles.

4. XML representation format
Text representation format follows the NKJP structure, which intends to support future stand-off annotation of the corpus, e.g. of the morphosyntactic level, and possible application of tools developed around NKJP (document indexers, search engines, taggers etc.). Structural elements are modelled explicitly which preserves all content as it appeared in the document. For example, the title of the document is represented in place of its occurrence on the title page. Document metadata often repeat these values, but they are reconstructed independently. In order to facilitate discovery of occurrence of a particular old-Polish term in the content, a new <pb> (page break) tag has been used, currently absent from NKJP. Apart from separating pages, its n attribute is used to preserve the number (or other label) of the following page. Page breaks can occur inside the paragraph, logically separating pagination and structural divisions. An independent annotation structure is represented with inline content such as marginal notes. Other structural elements such as <figure/> or <formula/> placeholders are preserved in place. Foreign interjections are embedded in <foreign> tag with language codes preserved in xml:lang attribute. For languages from outside the official ISO 639 code list, such as Ruthenian, new codes are invented using the x-code scheme reserved for private use.

5. Error detection and XML conversion
A Web application for managing upload of transcribed texts was integrated with a RTF-to-TEI converter, detecting potential errors in markup. Several types of errors are currently detected, mostly sig-
nalling structural inconsistencies or missing vital information about the work: broken or misplaced tags, out-of-page text fragments, inconsistent foreign language marking etc. Another useful function of the managing application is capability of displaying a simplified HTML preview version of the converted text to enable error detection.

6. Summary
The application proves very effective in the course of text transcription. On one hand it enables a precise structural annotation of a text. The markup allows to represent the most important elements of a text structure providing many possibilities of specialized corpus searching. On the other hand the possibility of using a well-known MS Word environment makes the system friendly for users less acquainted with IT. Thanks to automatic XML conversion the ultimate text format is consistent with the widely accepted TEI standard. The application also organizes a fluent workflow between all users in the text transcription process and facilitates error correction. Finally, it can be easily customized to handle other types of texts which makes it suitable for other corpus building projects

References

The position of objects in Old English and Old High German translations
Anna Cichosz, University of Lodz

It is generally assumed that the position of objects in Old Germanic languages, including OE and OHG, is influenced by two main factors: clause type and object type. As reported by Fischer et al. (2000), Allen (1995) and Mitchell (1985) for OE and Lockwood (1968) and Axel (2007) for OHG, pronominal objects tend to appear before the verb while nominal objects, especially very long ones, would rather follow the verb. On the whole, both SOV and SVO patterns may be found in main clauses (with pronominal objects often intervening between the subject and the verb and thus violating the V-2 constraint characteristic of this clause type), while subordinate clauses tend to demonstrate the SOV order, especially (but not exclusively) with pronominal objects. Interestingly enough, numerous studies show that deviations from these patterns are relatively frequent, which suggests that other factors may also be at work. It has been noted that information structure may exert some influence on the position of (especially nominal) objects (cf. Petrova 2012 and Taylor & Pintzuk 2014), while Koopman (2005) suggests that in translated material, which is a substantial part of the available OE and OHG textual records, the order of the Latin source text may also influence the relative frequency of the OV and VO patterns, though the degree of this influence has never been checked. The aim of this study is to fill in this gap and investigate the influence of the Latin source texts on the position of objects in selected OE and OHG translations. The following research question have been formulated: a) what are the possible and most frequent positions of objects in various types of clauses in the analysed texts?; b) how does object type (i.e. pronominal vs. nominal) influence its position?; c) which of the object positions follow the Latin original and which are clearly independent of the source text?; d) are there any significant differences between the orders employed in the translations into the same target language?; e) do any of these differences stem from the source text interference? The study is based on a parallel corpus containing syntactically annotated samples of translated texts and their originals (Tatian, Isidor and Physiologus for OHG, The Book of Genesis, The West Saxon Gospels and Bede’s Historia Ecclesiastica for OE, ca. 21,000 clauses altogether), created for the purpose of the project “The influence of Latin syntax on word order of selected Old English and Old High German translations”.
Time-discretising adverbials in reference corpora: distributional evidence of conceptualisation patterns

Mikołaj Deckert, Piotr Pęzik, University of Lodz

This paper looks at the distribution of selected adverbials used to discretise and quantify time in units of minutes, seconds and hours in reference language corpora of Polish and English. We carry out an exploratory analysis of the distributional patterns of such expressions and come up with three main findings. Firstly, we observe that the discretisation intervals for minutes and seconds diverge from the overall frequency distribution of cardinal numerals in the two corpora. Secondly, a significant level of correlation has been found between frequency distributions of n-seconds and n-minutes time adverbials in Polish and English data. Finally, the distribution of salient intervals for n-hours adverbials differs considerably from the discretisation patterns observed for n-seconds and n-minutes expressions. We relate these findings to general cognitive aspects of quantification and further link them to the iconic clock dial-based time conceptualisation. This is supplemented with a qualitative analysis of samples from the National Corpus of Polish and the British National Corpus as we attempt to position discretising construals of time along what could be referred to as the formulaicity-propositionality continuum.
Measuring the distribution of mass and count nouns. A comparison between a rating study and a corpus based analysis.
Francesca - Chiara - Giorgio Franzon - Zanini - Arcara, Università degli Studi di Padova

1. Introduction
Despite the huge amount of literature involving theoretical linguistic and experimental approaches concerning the mass and count issue (for a review: Bale & Barner, 2011), few studies have dealt with measuring the distribution of mass and count nouns in the actual use of language (for English: Katz & Zamparelli, 2012). As a first step to fill this gap, a survey on subjective frequency of plural and singular word forms was conducted. Results of the survey were compared to frequencies collected from different corpora of written Italian.

2. A complementary approach
In this study we focused on a list of 478 nouns (239 nouns inflected both at the singular and at the plural), that according to theoretical definition and to a preliminary corpus study are spaced across the range of mass and count usage. The list included even the plural of nouns for which only singular occurrences would be expected on a theoretical basis (pure mass nouns). In the first part of the study, a questionnaire on subjective frequency was designed. Participants were asked to rate (on a 5-point Likert scale) the subjective frequency of a list of nouns. The available ratings ranged from “never heard or seen” (score: 0) to “I hear or see this word more than once a day” (score: 4). The questionnaire was administered online. A total of 126 participants took part to the study. Participants varied widely in age and education. In the second part of the study, we collected the frequency values of the experimental stimuli from several corpora of written Italian: Colfis, La Repubblica, Subtlex (tagged), Subtlex (ignoring tags), it-WAC. These corpora differ from each other as for the strictness of the criteria concerning the texts they selected as sample. Results on estimated subjective frequency and on corpora frequency were then compared, by means of correlations. We considered that the subjective frequency was a better estimate of the actual acceptability of the singular/plural inflection of a noun. Eventually, by investigating these correlations, it was possible to measure the reliability of corpus-based estimates on the actual possibility of occurrence of singular and plural forms of nouns. Observed correlations with the set of experimental nouns ranged from 0.70 to 0.78. The Subtlex (ignoring tags) showed the higher correlation, whereas the Colfis showed the lowest correlation. In a further analysis we focused on nouns for which the mean subjective estimates of frequency had a score equal or greater than 2. We assumed that such threshold points to nouns for which there is a consistent judgment of plausibility across the participants of the questionnaire. Interestingly, focusing on these nouns some corpora
showed zero occurrences (Colfis = 70, Subtlex = 23, Subtlex – no tag = 10).

3. **Conclusions**

The results of the rating questionnaire showed that for almost every word in our list the plural was heard or read by the participants at least once in their life. In line with recent studies, this confirms the elasticity of mass and count usage in natural language (Katz & Zamparelli, 2012). Some corpora seem to capture better than others the distribution of possibilities that are largely accepted by the native speakers. This suggests that corpus studies on mass and count nouns should be interpreted cautiously, in the light of the corpus characteristics. Indeed, these results suggest that different selection criteria of the corpus could lead to less precise estimates of the possibility of observing a given word form in the language usage.

**References**


There are many facets to vocabulary knowledge, including basic dictionary meaning, pronunciation, likely grammatical constructions, collocations, semantic associations, pre-fabricated chunks, and textual and genre restrictions (Hoey, 2003), but one way of organizing vocabulary knowledge is distinguishing between receptive and productive vocabulary (Lee & Muncie, 2006; Meara & Olmos Alcoy, 2010). Receptive knowledge of vocabulary is thought to be easier to acquire. As Nation and Webb (2011, p. 41) point out, “learning vocabulary productively takes more time and effort than receptive learning.” The distinction is useful for instructional purposes as it can provide some indication of students’ listening and reading proficiency versus students’ speaking and writing proficiency (Webb, 2008). However, the relationship between the two types of knowledge is not completely clear, not only because of possible “distinct constructs of vocabulary knowledge, but also the degree to which various forms of vocabulary knowledge differ from one another in complexity and difficulty for learners” (Stewart, Batty, & Buvée, 2012, p. 696-697). Also, it is not clear how receptive and productive vocabulary are related, the relative rate at which they grow, and whether there is a threshold level receptive vocabulary (Meara & Olmos Alcoy, 2010, p. 223).

Testing for productive vocabulary size is complicated by a number of factors. These include individuals who have a rather simple speaking or writing style, situational factors that influence vocabulary production, the use of frequency levels with assumptions that words are learned by level, issues in counting types with different size texts, and the unit for counting words (Nation & Webb, 2011). Not surprisingly, it “has proved surprisingly difficult to develop simple and elegant tests of productive vocabulary size that have any degree of face validity” (Meara & Olmos Alcoy, 2010, p. 223).

The main purpose of this presentation is to describe the preliminary results of an analysis of a small Vietnamese L1 English learner corpus to determine if some features of productive vocabulary use are related to a receptive vocabulary test. The presenter will begin by briefly describing the construction of the learner corpus. It consists of 16, 589 words produced by 147 post-university students as part of placement testing for an English language program in Vietnam. The students wrote after taking a receptive vocabulary test, the Bilingual Vocabulary Size Test for Vietnamese Learners (Nguyen & Nation, 2011). Given the complicated nature of measuring vocabulary use, a number of features are being investigated. The table below presents the features being investigated along with the instrument used for each analysis:
<table>
<thead>
<tr>
<th>Feature for Analysis</th>
<th>Instrument for Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>Vocabprofile</td>
</tr>
<tr>
<td>Frequency Level</td>
<td>Vocabprofile; WordAnd-Phrase</td>
</tr>
<tr>
<td>Academic Words</td>
<td>Vocabprofile; WordAnd-Phrase</td>
</tr>
<tr>
<td>N-grams</td>
<td>Antconc</td>
</tr>
<tr>
<td>Type-token ratio</td>
<td>Vocabprofile</td>
</tr>
<tr>
<td>Lexical Density</td>
<td>Vocabprofile</td>
</tr>
</tbody>
</table>

**References**


Corpus-Based Study of Polish Public Procurement Law in English Translation

Justyna Giczela-Pastwa, University of Gdansk

Corpora have been used by translation scholars for over two decades now and have allowed to gain an insight into a number of aspects which according to the Holmes-Toury map (Munday 2008) fit into the categories of descriptive and applied translation studies.

With the use of corpus methodologies fruitful research has been conducted into translationese, translation universals, terminology, stylistics of translation and translator training, to name but a few (cf. Xiao and Yue 2009; Kruger et al. 2011). It has been suggested that corpus-based translation studies have been evolving into a new coherent paradigm (Laviosa 1998: 474).

The paper aims at presenting an ongoing project design which corresponds with the aforementioned research trend. The translation phenomenon under research is inverse translation of legal texts. Despite the fact that translation into a non-native language is commonly commissioned not only in Poland but in other countries where languages of limited diffusion (cf. Mikkelson 1999) are spoken, inverse translation has rarely been acknowledged by translation scholars, Pokorn (2005) being one of the notable exceptions. Further contributions have also been made by Pavlović (2007, 2009) and Wang (2009, 2011).

In the project corpus linguistics methods are used in order to analyse English translations of Polish public procurement law. Available translations, published by three different publishing houses, are compared to one another and, most importantly, to non-translated English statutes regulating the field of public procurement in England and Wales. The research carried out with the use of monolingual comparable specialized corpora fits into the category of descriptive translation studies.

Special attention is paid to a certain linguistic circularity inherent in the object of study: Polish public procurement law implements relevant EU regulations, the Polish official versions of which have been drafted on the basis of English texts and thus reflect qualities of English language (in-depth investigation in Biel 2014). In this way the language of Polish public procurement law may possess unique attributes which are reproduced in the translation into English. Corpus methods may allow to observe it in more detail.

References

A Corpus-driven study of the use and discourse functions of lexical bundles in Polish patient information leaflets
Łukasz Grabowski, Lazarski University
Pawel Bechler, Warsaw University

So far little attention has been paid to corpus linguistic explorations of recurrent phraseologies found in Polish texts, in particular texts representing specialists registers of language use. As a consequence, one may note the lack of corpus linguistic studies of lexical bundles (Biber et al. 1999) found in texts originally written in Polish. Conducted from a register perspective (Biber and Conrad 2009), this descriptive and exploratory study (Grabowski 2014) is intended as a first step towards a comprehensive corpus-driven description of the use and discursive functions of the most frequent lexical bundles found in patient information leaflets (PILs), one of the most commonly used text types in the healthcare sector in Poland. The research material includes 100 PILs written originally in Polish, extracted from internet websites of ten pharmaceutical companies operating on the Polish market, compiled in a purpose-designed corpus of circa 197,000 words. Based largely on the methodology proposed by Biber, Conrad and Cortes (2004), Biber (2006), Hyland (2008) and Gozdż-Roszkowski (2011), which makes possible an analysis of the use and discursive functions of lexical bundles, the study is also meant to provide methodological guidelines for future research on lexical bundles in Polish texts. This appears to be important since so far lexical bundles have been studied predominantly in texts originally written in English. The results of this preliminary research reveal salient links between the frequent occurrence of lexical bundles on the one hand, and situational and functional characteristics of the PILs under scrutiny on the other. More specifically, it was revealed that the PILs are by lexical bundles expressing stance, notably obligation/directive bundles and epistemic stance bundles that are used to instruct patients on how to properly use medicines and to alert them to potential side-effects. Finally, a few methodological guidelines and suggestions are provided that pertain to the criteria used to identify the bundles as well as to the methods used to avert the problem of overlapping bundles, among other findings reported in the presentation.

Keywords: corpus linguistics, phraseology, register analysis, corpus-driven approach, lexical bundles, patient information leaflets

References

Mapping corpora: Approaches to understand the geographies in texts
Ian Gregory, Lancaster University

Many texts carry a wealth of information about places and geographies, however, traditional corpus analysis techniques struggle to summarise and analyse this information. If we are to use corpora effectively in the social sciences and humanities, subjects where geography is frequently important, this presents a serious limitation. Geographers make extensive use of digital technologies through geographical information systems (GIS) and virtual globes such as Google Earth, however traditionally these have concentrated on quantitative or cartographic sources. This paper explores methods of bringing these two approaches together to allow us to better understand the geographies within corpora. This involves going through several stages: firstly place-names within the text need to be automatically identified and associated with a coordinate-based location. Doing this allows the corpus to be mapped. From here, more sophisticated analyses can be undertaken that combine GIS-based spatial analysis techniques that allow us to ask ‘where is the corpus talking about?’, with approaches from corpus linguistics that allow us to ask ‘what is the corpus saying about these places?’ The paper draws on a range of different examples including a relatively small corpus on the literature of the English Lake District, and Early English Books Online (EEBO) a much larger corpus from which we explore the geographies of prostitution in the seventeenth century and before.
Temporality and language contact. A corpus-based study of the language of the Russian-speaking immigrants in Germany.
Christine Grillborzer, Slavisches Seminar Universität Freiburg

The central aim of the study is the empirical analysis of the German-Russian language contact situation and particularly its influence on the syntactical structure of the temporal adjuncts in the Russian spoken in the diaspora.

The study consists of two parts. The first one is the analysis of the syntactical changes based on the interview corpus composed at the Slavonic Department of the Regensburg University. Although this analysis yields a great number of syntactical particularities, all of them appear with a comparatively low frequency. Therefore, a further empirical study implementing the observed tendencies seems to be of avail in order to generalize the conclusions.

Thus, the second part of the study is the coherent evaluation of the questionnaire concerning the narrow field of temporal adjunct that has been completed by thirty Russian-speaking immigrants. The results, compared with the data of a commensurate control group, allow us to demonstrate the certain explicit contact-induced tendencies of the syntactic change of the Russian spoken in Germany.

The analysis, based on the authentic linguistic data from the corpus, allows us to formulate the following hypotheses:

- Generalization of the accusative constructions: The accusative offers numerous possibilities to mark time in German. An extension of the narrower functional domain of the accusative in Russian can be expected, i.e. its use instead of the more complex instrumental as well as prepositional constructions.

- Dominance of a certain preposition: Most German prepositions have several correspondences in Russian. These are used contextual and, as a general rule, cannot be exchanged. It is supposed that the use of a certain preposition will be preferred.

- Changes in the instrumental constructions: In regard to the Russian instrumental constructions (which find no equivalent in German) one can suppose various possible directions of change.

- Levelling of divergences: Though the changes are predictable, it is clearly more difficult to estimate which preposition will dominate.

- Incorrect government: The analysis of corpus makes evident that the government of prepositions is noticeably concerned by the transfer phenomena.

- Different behavior in regard to the production and reception of the contact-induced peculiarities or mistakes.
Semantic changes in English affixes – a corpus-based study
Marta Grochocka-Sidorowicz, WA AMU

The aim of the study was to offer an insight into the semantics of some productive derivational affixes in English by means of examining nonce formations and neologisms comprising thereof, as well as to find out how these affixes are accounted for in learner’s dictionaries. To begin with, it is crucial to make a distinction between the notion of nonce formation and neologism as the terms are understood differently by different authors (see e.g. Štekauer 2002, Hohenhaus 1998, 2005). In a nutshell, every new creation is born as a highly context-dependent nonce formation which may later become a neologism if used frequently, in various contexts and over a certain period of time (Fischer 1998). Next, for the purposes of the study, a database of nonce formations and neologisms was created using a web-based application called NeoDet. First, the application was used to compile a study corpus out of which new words could be extracted. The study corpus consisted of fourteen million words and comprised articles from the most widely read British broadsheets and tabloids. Furthermore, the application was used to semi-automatically extract new lexical items from the study corpus following a procedure based on the exclusion principle, i.e. every word from the study corpus which is absent from the exclusion list constitutes a nonce formation or neologism candidate subject to further verification. The exclusion list consisted of words deemed as established in the language and included a few online dictionaries, slang dictionaries, the British National Corpus, and a wordlist of proper names and geographical names. Only when a lexical item was absent from all the exclusion sources was it regarded as a potentially new word. Each nonce formation and neologism candidate was then verified manually and the whole noisy output was discarded. In the process of verification, whenever an affixed nonce formation or neologism was encountered, the NeoDet search engine was used to check if the study corpus contains more examples of coinages with that affix. A semantic analysis of all nonce formations and neologisms containing a given affix was then performed. Additionally, it was decided to compare the results of the analysis with the way pedagogical dictionaries treat the affixes in question. For this reason, the online versions of the so-called Big Five were consulted. The Big Five are the most popular British monolingual dictionaries with advanced learners of English as their target users. It is sensible to expect the dictionaries to provide the most up-to-date information concerning affixes as strong market competition forces publishing houses to be particularly sensitive to language novelties. At the same time, it stands to reason that if a learner comes across a new coinage which they fail to understand, a general dictionary will prove of no help, as its purpose is to provide information about words which are already established in the language. Nonetheless, as observed by Moon (2008), creative linguistic behaviour is often systematic, and as such it deserves to be accounted for, especially in learner-oriented dictionaries. Hence, it was decided to examine how current pedagogical dictionaries manage to meet learners’ expectations when it comes to pro-
viding information about derivational affixes and compare these findings with the data obtained from the study corpus. All in all, very interesting findings emerged in the case of the following affixes: -gate, -esque, -mania, and ultra-. It was possible to observe certain changes that appear to be taking place in their semantics and draw conclusions that may be used to provide better treatment of the investigated affixes in pedagogical dictionaries.

References


No need for rules in the -ez/-eza Spanish nominalization
Matias Guzman Naranjo, Universität Leipzig

The alternation in Spanish between the allophones -ez and -eza (eg. rápido-rapidex ‘fast-speed’, natural-naturaleza ‘natural-nature’) has been characterized (Aranovich and Argun 2006) as being governed by the number syllables of the base adjective: adjectives with two syllables will take -eza, and adjectives with three or more syllables will take -ez. On the basis of corpus evidence, both historical and synchronic, and experimental results, I argue against this account and in favour of a completely usage-based explanation.

There are two crucial problems with the rule-based account. The first one is that -ez(a) is not the only nominalizer of adjectives in Spanish, other forms like -ura (largura ‘length’) and -i-dad (velocidad ‘speed’) are also possible and there is no systematic rule that can account for which adjective nominalizes with which form. This means that speakers already need to link each adjective to each form, which makes the rule superfluous, specially because -ez(a) is unproductive in modern Spanish, and only about 100 adjectives can nominalize with it. The second problem is that there are multiple exceptions (historical and synchronic) which strongly argue against a deterministic rule. The first kind of exception are words that when first entered the paradigm were used with both forms, but later in the evolution of the language regularized. The second kind of exception we observe are words which typically appear with the form that follows the syllable counting rule, but occasionally are found in texts with the other form (eg. rojez instead of rojeza ‘redness’), these are usually very low frequency nominalizations. And finally, we find very interesting cases like ñoño (‘nerd’) which has only recently been added to the paradigm as ñoñez (‘nerdiness’), not respecting the syllable counting rule. What these examples show is that in productive use of -ez and -eza speakers do not follow the supposed rule, but instead these are coined on the basis of analogy to other existing forms. This is specially clear with the case of ñoño-ñoñez which was coined in analogy to another exception: niño-niñez (‘child-childhood’).

As an alternative, I propose that the pattern arose diachronically through frequency and analogy. High frequency words got to keep their form and attracted phonological neighbors to their pattern, while low frequency forms adapted through analogy to the pattern of higher frequency neighbors. This is not a synchronic rule but a path of language change. There are several ways in which we can test this hypothesis. I present evidence in the form of a diachronic corpus study of this phenomenon, and a statistical model that shows how this analogical process could occur. Using the Corpus del Español I show that there are examples of both forms for a large number of adjectives the century they first entered the paradigm, but in subsequent centuries one form becomes completely dominant and often unique. For testing whether there is a basis for an analogical process I fitted a Naive Discriminative Learning model to 96 adjectives using as predictors certain
word endings, the frequency of the adjective and the number of phonological neighbors (cues which speakers could follow). The model achieved an accuracy of 0.9375 and C score of 0.96, with only 6 errors (vs 12 by the syllable counting rule).

The conclusion is that the rule based account is unfit to explain the -ez/-eza alternation, and instead we need a diachronic, usage-based explanation.

References

Walenty: Towards a comprehensive valence dictionary of Polish
Elżbieta Hajnicz, Agnieszka Patejuk, Adam Przepiórkowski, Marcin Woliński,
Institute of Computer Science, Polish Academy of Sciences

The aim of this paper is to present Walenty, a comprehensive valence dictionary of Polish developed at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS), currently with the CLARIN-PL project. The dictionary is meant to be both human- and machine-readable; in particular, it is being employed by two parsers of Polish, Świgr (Woliński, 2004) and POLFIE (Patejuk and Przepiórkowski, 2012). As these parsers are based on two rather different linguistic approaches, the valence dictionary must be sufficiently expressive to accommodate for the needs of both – and perhaps other to come. For this reason, Walenty exhibits a number of features which are rare or absent in other valence dictionaries.

Each lexical entry contains a number of valence schemata, and each schema is a set of specifications of an argument. But are the noun phrase (NP) “a republican” in “Pat became a republican” and the adjective phrase (AdjP) “quite conservative” in “Pat became quite conservative” two different arguments, belonging to separate valence schemata of BECOME, or two different realisations of one argument specified in a single schema? Walenty is explicit about what counts as the same argument, and it employs the coordination test to resolve such doubts: if two phrase types can be felicitously coordinated in the same sentence, they are different realisations of the same argument; this is the case here, cf. “Pat became a republican and quite conservative” (Sag et al., 1985, p. 142, ex. (67a)). Hence, arguments are specified disjunctively, by listing phrase types which may occupy a given syntactic position and may in principle be coordinated.

Specification of nominal arguments includes information about their case. However, this is insufficient in Polish to uniquely determine grammatical subjects and objects: the former do not have to be nominal at all, but – depending on the verb – may sometimes be clausal or infinitival, the latter do not have to be accusative, as there are passivisable verbs taking instrumental or genitive objects. For this reason, subject and object arguments are explicitly marked as such.

Moreover, some nominal arguments cannot be assigned a unique morphological case, as it may depend on syntactic context. One such phenomenon is the so-called Genitive of Negation, where the normally accusative complement of an affirmative verb must bear the genitive case when the verb is negated. The case of such arguments is marked as structural, following the generative tradition. Another related couple of phenomena widely discussed in the generative literature are control and raising; for example, valence schemata of verbs such as PROMISE and ORDER are the same at the purely morphosyntactic level – they combine with an NP subject, an NP object and an infinitival phrase, e.g., “John promised/ordered Tom to come” – but they differ in their control properties: with “promised”, the under-
stood subject of to come is “John”, with “ordered” – it is “Tom”. In Polish, this distinction does not only matter for semantic interpretation, but is also correlated with certain agreement facts, i.e., it is useful even for purely syntactic parsers.

Some obligatory arguments look like typical adjuncts, for example those expressing location, time or manner, as in LIVE (somewhere), LAST (some time) or BEHAVE (in some way). Such arguments may be realised by a relatively large number of phrase types, including adverbial and prepositional, with a limited number of prepositions (cf. “John lives there / at Tom’s / in the suburb / under the bridge,” etc.). It would not be economical to list all possible realisations of, say, location in valence schemata of all verbs combining with such locative arguments, so – even in this morphosyntactic version – Walenty assumes a number of semantically-defined argument types, including locative, ablative, etc., temporal, durative and manner. A list of possible morphosyntactic realisations of such arguments accompanies the dictionary proper.

The final feature to be mentioned here is phraseology. Walenty includes special notation for various types of lexicalised arguments, from completely fixed (given as a string) to almost freely modifiable.

While the dictionary is still intensively developed, it is already – with over 11 400 lemmata and over 50 000 valence schemata – the largest and most detailed Polish valence dictionary.

References


Comparing tense and aspect forms across the world’s languages
Martin Haspelmath, Max Planck Institute for Evolutionary Anthropology, Leipzig

In this talk, I will give an overview of results from the cross-linguistic projects Atlas of Pidgin and Creole Language Structures (apics-online.info) and World Atlas of Language Structures (wals.info). I will also reflect on grammatical form-frequency correspondences in the area of tense and aspect, discuss semantic maps and tense-aspect categories, as well as the relationship between comparative concepts and descriptive categories.
The NOW and Consciousness: a Comparison
Dwight Holbrook, Adam Mickiewicz University in Poznan

It might be supposed that the NOW, or present time, features as nothing other than a linguistic extrapolation from clock time, or simply explains itself as an indexical signifier like “here” or “then”, or on the other hand that it falls into the same ballpark as consciousness, with its intractable problems of definability, the phenomenological vagueness of its “something it is like” or “nothing it is like”, the emergence conundrum, and so forth.

This paper will begin by showing in what ways consciousness and the NOW pose similar problems. These include the challenge of conceptualization, squeamishness about attaching any putative “material” existence to something that defies traditional notions of what qualifies as objective or thing-like, and the question of origins or emergence. Where do consciousness and the NOW come from?

From here we move to the subject of the NOW’s distinctiveness, and how it is that there can be more to it than clock time and indexicality. We begin by noting that the NOW is inter-relational (inter-relational rather than intersubjective) in a way that one’s privatized (brain-based) consciousness is not. What I observe and attend to is not what you observe and attend to – i.e., we each have our own individual perspectives. By contrast, we share NOWs, or rather the NOW (in the singular), in a communal way. Otherwise, consensus would be impossible. How would any knowledge be possible if the temporal occasion marking what I witness – an event of any kind – did not co-occur with the possibility of the temporal occasion of your witnessing as well? Or take the quandary that a notion of “knowledge” ends up with when the event I am witnessing is not a present-time event. Here before me stands a yesterday you or a tomorrow Einstein. This leads to a further point. If what I come to know is other than, or not reducible to, my own private mental constructions, if in other words there is an independent or objective reality that science and our own mutual consensus confirms, it must necessarily be predicated on the NOW’s co-occurrence, not on what one thinks or perceives as an individual. Finally, it will be explored how the NOW, unlike consciousness as such, secures a starting point for what comes to be known and what comes to be empirically investigated. In this respect, it can be argued that present time precedes and is prerequisite for the past. Conceptually it may be possible, but the conceptual always presumes, as a starting point, a first person perspective, an experiential NOW presupposed in any conceptualization.

The paper concludes by posing the problem of ontological relativism. If what comes to be known – and remembered – is contingent on the NOW, how can the objective world, the world that science explores, have ontological independence? Here we draw a distinction between observer dependence or consciousness dependence, and NOW dependence. We conclude that it is misleading to consider the NOW as privatized in consciousness.
We situate the NOW in nature. It exposes itself through the channel of a first person perspective but, being situated in nature, it is common to every human perspective, and in that respect comparable to objects of scientific investigation approached from a third person perspective.
Towards a road map for exploring evaluative meaning
Susan Hunston, University of Birmingham

There are by now a plethora of approaches to investigating evaluative meaning in corpora, but no very clear set of instructions for someone beginning work in this area. This paper considers the available approaches and asks whether sufficient common ground can be identified to devise a ‘road map’ for research.
Evaluating threats in the press
Jane Helen Johnson, University of Bologna

Referring to Beck’s influential book (Risk Society - 1992), one sociologist commented that the risk society is actually made up of “a vast number of relatively unfamiliar threats, with new threats always lurking in the background” (Ungar 2001: 276). There are suggestions even that the media and the political, cultural and economic world are promoting messages of fear and danger to encourage our feeling of being threatened by ‘the other’, thus not only outlining a ‘risk society’ but possibly even a ‘threat society’ (Nohrstedt 2010: 11).

It may be argued that this is just a question of perception. After all, we actually live in the most low-risk society in history, a society which is less violent (Pinker 2011) and more prosperous (Ridley 2010) than ever before. However there is no doubt that the lemma threat – and its near synonyms ‘risk’ and ‘danger’ – is a key lexical item in news discourse. Given that the media plays a major role in shaping the views of the public (Kitzinger and Reilly 1997: 320) and setting agendas regarding the importance of certain topics (Baker et al 2013: 2), it is not just frequency alone but how these lemmas are evaluated that perhaps holds more interest in terms of heightening perception. This paper explores the nature of ‘threat’ in UK newspapers in two distinct periods twenty years apart, using corpus linguistics methods to focus on the evaluation of threat in the discourse (Hunston 2011; 2004). More specifically, an MD-CADS approach (Partington et al 2013; Partington ed. 2010) was used to investigate evaluative patterns around the lemma threat in the online versions of two UK newspapers - The Guardian and the Daily Mail – in both 1993 and 2013. Corpus investigation of concordances and collocates of threat* enabled us to focus on lexical and grammatical evaluation, as well as textual evaluation (e.g. wordforms of threat are frequent lexical choices in headlines - Baker et al 2013: 70; Bevitori 2010: 145; a writer’s evaluation of a particular issue may often be found in the concluding paragraphs of an editorial - Morley 2004). Reference is also made to evaluative (semantic) prosody (Sinclair 1987) as well as irony, with findings in newspaper discourse around certain specific topics (in articles about the family, for example - Johnson 2014a, 2014b) suggesting that there is a recent trend in the Daily Mail towards using the word threat ironically, perhaps as a distancing device. The paper thus aims to describe diachronic changes in the way threat is evaluated in the press as well as differences and similarities (Taylor 2013) in evaluation of threat in newspapers of different political orientation, format and style.

References

bridge University Press.
Heuristic Detection of Zero Subjects for Polish

Adam Kaczmarek, University of Wroclaw / Wroclaw University of Technology
Michał Marcińczuk, Wroclaw University of Technology

In this paper we present a heuristic approach for zero subject detection in Polish texts. The algorithm makes use of premises from different sources including morphological analysis, dependency and syntactic parsing, valence dictionary. Zero subject describes phenomenon of a verb referring to a discourse entity by omitting its subject. Detection of zero subject is important for the task of zero anaphora resolution as a preprocessing step. However this problem was not yet very widely studied for Polish language. The only known attempt was done by Kopeć [1] who presented a supervised approach using machine learning method RIPPER utilizing orthographic and morphological features. The importance of zero anaphora resolution is emphasized by the fact that zero coreference is the second most frequent coreference relation type in KPWr corpus [2], right after the coreference of proper names.

The task of zero subject detection was divided into two major parts: determination of verb ability to be anaphoric and determination if a verb has explicit subject. For purposes of zero subject detection we adapted definitions of verb and noun following the idea from [1]. Verbs can only belong to following part-of-speech classes: fin, praet, winien, and bedzie. Words belonging to other classes, usually assumed to describe verbs, are discarded as they cannot have subject. Nouns are defined wider and contains also numerals, gerunds and pronouns, because they often occur to be subjects and share most grammatical properties with "standard" nouns. We also treat more complex constructions as subject candidates like conjunction of explicitly written number with dative noun as subject for neutral verb. (eg. "...206 posłów głosowało...")

We explored several lexical and grammatical properties of verbs as well as utilized informations form Polish Valence Dictionary [3] in order to determine if it is possible for a verb to be zero anaphoric. Subject detection is the part where we connected results from Polish Dependency Parser [4], ChunkRel [5] and deterministic windowed subject search followed by verb-subject agreement checking. We also took an attempt to retag certain words originally tagged as unknown in the testing corpus. The idea was to use morphological guesser implemented in the WCRFT tagger [6] to fill the missing morphological tags. This brought some improvement in subject detection and agreement determination what had positive effect on the final zero subject detection. The experiments were performed on the Polish Coreference Corpus [7] in order to provide direct comparison to MentionDetector [1] which is a state-of-art method for zero verb detection for Polish. The corpus was divided into two parts (development and test). The development part was used to develop the algorithm and to improve the rules on the basis of error analysis. The test part was utilized to compare the performance of the algorithm with MentionDetector. Our algorithm outperformed MentionDetector in means of F-score by more than 8 percentage points, having much greater recall (increase by ca. 22 percentage points)
and slightly lower precision (decrease by ca. 5 percentage points). The recall is much more important than precision as the zero subject detection is a preprocessing step for zero anaphora resolution and some false positives might be discarded on succeeding processing steps. The also report results obtained on the KPWr corpora. The resulting algorithm was implemented as module of Liner2 framework [8] called Minos.

References


Corpus-based Analysis of Czech Units Expressing Mental States and Their Polish Equivalents. Identification of Meaning and Establishing Polish Equivalents Referring to Different Theories (Valency, Pattern Grammar, Cognitive Linguistics)

Elżbieta Kaczmarska, University of Warsaw

The analysis is focused on Czech polysemous units mít rád and toužit, expressing mental states. The goal of the analysis is to test which theory can lead us to the closest equivalents of these units in Polish.

In 2013 a pilot research concerning the ambiguous Czech verb toužit was conducted (Kaczmarska & Rosen 2013). The study was supposed to reveal if valency can influence the choice of an equivalent in Polish. It was assumed that for some senses the equivalent can be established based on the convergence of the valence requirements. The hypothesis proved to be true. However, the influence of valency was not observed in all the senses of the verb. A more extensive research should be conducted to establish equivalents (or cluster of equivalents) for given units (Levin 1993; Lewandowska-Tomaszczyk 1984, 2013). To identify all the meanings of the units and finally to find a proper equivalent for each of their senses, we decided to take into consideration the approaches of Pattern Grammar (Ebeling & Ebeling 2013; Francis & Hunston & Manning 1996; Hunston & Francis 2000) and Cognitive Linguistics (Langacker 1987, 1991, 2008; Taylor 2002; Mikołajczuk 1999).

The analysis proper is preceded by automatic extraction (Och & Ney 2003) of pairs of equivalents from a parallel corpus InterCorp (Čermák & Rosen 2012, see http://korpus.cz/intercorp). These pairs constitute a kind of bilingual dictionary (Jirásek 2011). The analysis includes automatic excerption of chosen words (with aligned segments) from InterCorp. The segments are analysed manually. We check (in each segment) how the key word was translated and what kinds of collocations and arguments it had. On the sidelines of the analysis proper we conduct additional studies: the analysis of mít rád requires also a confrontative reaserach with the verb milovat. The analysis of toužit is supplemented by a contrastive study of nouns: touha, stesk and tužba.

Subsequently we can start to operate with the tools specific for each linguistics approach (Valency, Pattern Grammar, Cognitive Linguistics).

The results of the triple analysis will let us establish the semantically and syntactically closest Polish equivalents of Czech words.

As an outcome of this research, we hope to prepare an equivalent-searching algorithm that will be based on a syntactico-semantic analysis. The algorithm will be applied to the analysis of different words expressing mental states. The method is likely to become an important tool for lexicographers.
CL for CDA: What can we do better?
Victoria Kamasa, Adam Mickiewicz University in Poznan

In this study we will present a critical review of usage of corpus linguistics techniques (CL) in Critical Discourse Analysis (CDA). Baker and associates (2008: 274-275) state: “the number of such studies [using CL techniques for CDA] in proportion to the number of studies in CL or CDA is extremely small. However, more recently, it seems that use of CL techniques is becoming increasingly popular in critical approaches to discourse analysis”. Both the methods used and the research problems are very diversified. Methods vary from bare analysis of frequencies (Alcaraz-Ariza, 2002; Mautner, 2007) through analysis of collocations (Don et al., 2010; Freake et al., 2010; Lischinsky, 2011; Mautner, 2007; Weninger, 2010) or keywords (Bachmann, 2011; Don et al., 2010; Lukac, 2011; Weninger, 2010) to the analysis of key semantic domains (Prentice, 2010). Whereas research subjects range from national identity issues (Don et al., 2010, Freake et al., 2010, Prentice, 2010), through different social problems (Kirejczyk, 1999, Lukac, 2011, Yasin, Hamid, Keong, Othman, & Jaludin, 2012) to social construction of businesswomen (Koller, 2004) or economic crisis (Lischinsky, 2011). Such variety of methods and approaches calls for a critical metaanalysis which will be undertaken in this study.

Our review is based on over 30 papers in which authors declared to use some form of CL for some form of Critical Discourse Analysis (CDA/CL). We will analyze the methods used by paper’s authors as well as the results provided by those methods in order to propose some points for improvement. The analysis will concentrate on two major issues:

- the relation between used methods, received results and postulated conclusions e.g. the degree to which the results support the conclusion;

- the relation of the research practice to the benefits of using CL for CDA. Some of those benefits have been pointed out in the literature: reduction of researchers bias (Mautner, 2009), explicitness and systematicity (Marko, 2008), rather exhaustive than selective description (Hardt-Mautner, 1995) or more focused approach to texts achieved by highlighting lexical and grammatical regularities (Lischinsky, 2011).

As a result we will describe eight main points in which such CDA/CL analysis might be improved. The first one concerns corpus design: we will show how decisions taken on this stage might limit the results. Secondly, we will refer to usage of statistics and demonstrate how some results might be improved by extending the number of issues statistics is used for. Moreover, we will point to some inconsequence which may take place during the research process concerning both sticking to rules declared by the author and paying attention to numbers such as word frequencies. Another problem we will discuss can be called “mind-reading problem”: while the results concern the proprieties of text, the conclusions regard cognitive states of
the users. Then some possibilities to improve the research by grounding it more deeply in theoretical background will be presented. We will also refer to the so-called cherry-picking problem (Breeze 2011), which is postulated to be solved by the usage of CL techniques (Degano, 2007, Lischinsky, 2011). Finally, we will discuss briefly the role of researcher’s intuition and show some stages of CDA/CL research in which the intuition continues to play crucial role. For every of these points we will present some examples from research practice and demonstrate its possible influence on the results and conclusions.
Corpus tools and resources in lexicography
Mariusz Kamiński.
Panstwowa Wyższa Szkoła Zawodowa w Nysie

This paper aims to show how a corpus and corpus processing tools can help
the lexicographer in compiling a monolingual and bilingual dictionary. The
paper begins with a brief outline of the use of a text-based approach to comp-
piling dictionaries, showing in broad outline the main goals and problems of
lexicography. It is followed by a presentation of advancements in computer
technology in the context of dictionary compilation. The paper emphasises
practical utility of annotated corpora, as well as of the software to process
it, which ranges from simple programs generating word-lists, word-clusters,
and concordances, to more sophisticated ones, such as Sketch Engine. The
paper underlines the role of the lexicographer’s familiarity with computer
technology.
Automatic keyphrase extraction for documents in Polish - comparison of supervised and unsupervised approaches

Paweł Kędzia, Henryk Maciejewski, Marcin Oleksy, Błażej Ryśnik, Wrocław University of Technology

The purpose of this work is to compare different approaches to keywords extraction from documents in Polish language. We adapted several supervised and unsupervised methods suggested in literature to operate on Polish language documents. We performed a comprehensive study in which we compared the methods in terms of standard measures such as precision and recall. We also introduced a novel measure defined in terms of the meaningfulness of the automatically generated keyphrases as quantified by human evaluators. Training and evaluation of the methods was done using the KPWR corpus.

In this work we focus only on methods of keyword extraction, i.e. we attempt to select single or multi-term phrases which are present in the document and which are related to the semantic content of the document.

Supervised methods of keyword extraction were based on the work of Turney (2000) and Hulth (2004). These methods realize the following steps:

1. Selection of candidate phrases from a document. Candidate phrases are extracted as noun phrase (NP), agreement phrase (AgP) of adjective phrase (AdjP) chunks, with chunking for Polish done with the Jlobber program. This procedure was inspired by Turney (2000); we also tested a modification of this based of frequent sequences of parts of speech (POS) taken as candidate phrases, as suggested by Hulth (2004).

2. Candidate phrases are represented as vectors of features. The features are based on the frequency of the phrase and its terms in the document and on position of the phrase in the document.

Supervised methods rely on prediction models trained using a collection of documents with annotated keywords. At the keyword extraction stage, the models are used to make the prediction whether a candidate phrase is likely to be a keyword. We trained several models, e.g. decision trees (C4.5 algorithm), random forests, naive Bayes, KNN and we also attempted metalearning such as bagging or boosting.

We also tested two unsupervised methods adapted for Polish language: RAKE algorithm and TextRank (Mihalcea, 2004). Since the RAKE method, based on simple relative frequency of terms in the document, realizes much worse performance than the remaining methods, we focus here only on the TextRank method. TextRank is based on the well known PageRank, adapted for natural language processing. The idea is to construct a graph with nodes representing candidate terms extracted from the document and edges representing the recommendation relationships between the terms. The PageRank algorithm is used to rank nodes of the graph in terms of importance in
the document. The ranked lists of terms are postprocessed to generate final lists of keywords or keyphrases.

Technically, we used single words as the graph nodes (nouns and adjectives only, in lemmatized form). The nodes can also include proper names (option of the method). The edges between nodes are added to the graph if the terms represented by the nodes are adjacent in the document (we tested directed and undirected graphs). Then we used several heuristics to postprocess lists of terms ranked by PageRank, such as generating multi-word keyphrases from important terms occurring in the document in adjacent positions, or selection of AgP chunks containing important terms.

To evaluate the methods, we calculated precision as perceived by human evaluators, with the winner TextRank reaching 55% precision. Technically, the measure was defined as percentage of automatically generated keyphrases which were recognized by a group of human evaluators as appropriate keywords.
A corpora-driven study of social class in Britain (a case-study of abstract nouns)
Natalia Khokhlova. MGIMO-University

Present-day corpora present ample opportunities for exploring various aspect of language – from lexis to grammar to discourse analysis. They contain millions of words and sentences, which allows linguists to conduct research of standard or substandard language as well as language varieties and variations. Sociolinguistics, on the other hand, aims at a particular social group or a situation and, therefore, bases its research on experiment and observation rather than a collection of words, no matter how significant and complete. Nevertheless, annotations in modern corpora enable researchers to retrieve demographic information and analyze language phenomena or vocabulary of any given social group. This has led to an increasing number of sociolinguistic studies employing corpora as their primary source (Rayson, Xiao, Baker and others).

This paper aims to consider the peculiarities and differences in the use of abstract nouns by various groups of people by age, gender with special focus on class. The debate on social class as a demographic and linguistic phenomenon at different times varied from acceptance in interest to denial of existence. Many political claims that social equality has been achieved and thus social class is an outdated concept non-applicable in present-day society, influenced social studies and many researchers have been cautious in their estimates of data. Nevertheless, social stratification still exists even though classes may have changed over time. This paper follows the classification of social strata by the National Readership Survey adopted by the British National Corpus (BNC) and primarily focuses on the polar opposites of the spectrum – the upper-middle and working class.

Abstract nouns have been chosen as the research object because they present an interesting lexical group; psycholinguistic and neurolinguistic studies (Schwanenflugel, Paivio, Wiemer-Hasings, Caramelli) show that the rate at which people are able to remember and process abstract nouns depends on their age and mental capabilities (in case of deep dyslexia and aphasia research). Moreover, previous linguistic research called for an introduction of a “degree of abstractedness” as it is impossible to draw clear distinction between groups of abstract nouns and a scale of abstractedness would be more helpful in classifying this group of nouns. A number of such scales have been developed based on different principles (both linguistic and extralinguistic); current research uses an anthropological scale elaborated by L. Cherneiko which focuses on the degree things and concepts are perceived by people.

To determine the differences in the use of abstract and concrete nouns as well as establish the groups of abstract nouns used by different social groups the spoken demographic part of the BNC was employed. This subcorpus allows obtaining sufficient information about the speakers (age, sex, class, profession, role in a dialogue) and is relatively balanced. In order
to retrieve required data nouns from the subcorpus were extracted, both clearly annotated (NN0, NN1) and those with double annotation (e.g., NN0-ADJ, NN1-WS), which came up to the total of 228063 instances (including 46616 for the upper classes and 29939 for lower classes). These were further distinguished as either abstract or concrete and classified along the scale of abstractedness from things and objects to metaconcepts.

Such approach provides information which, on the one hand, enables detailed analysis of speech of any particular social group and, on the other, may further assist in evaluation of speech and its abstractedness to determine demographic characteristics of the speaker.
CorpoGrabber: The Toolchain to Automatic Acquiring and Extraction of the Website Content
Jan Kocoń, Wroclaw University of Technology

In this article we present a set of tools to get the most relevant content of the website, including all subsites (up to the user-defined depth). The proposed toolchain can be used to build a big Web corpora of the text documents. It requires only the list of the root websites as the input. Tools composing CorpoGrabber are adapted to Polish, but most subtasks are language independent. The whole process can be run in parallel on a single machine and includes the following tasks: downloading of the HTML subpages of each input page URL [1], extracting of plain text from each subpage by removing boilerplate content (such as navigation links, headers, footers, advertisements from HTML pages) [2], deduplication of plain text [2], removing of bad quality documents utilizing Morphological Analysis Converter and Aggregator (MACA) [3], tagging of documents using Wroclaw CRF Tagger (WCRFT) [4]. Last two steps are available only for Polish. The result is a corpora as a set of tagged documents for each website.

References

Avenues for Corpus-based Research on Informal Spoken Czech

Marie Kopřivová, David Lukeš, Zuzana Komrsková, Petra Klimešová, Charles University

This paper focuses on several phenomena typical for spontaneous spoken Czech in informal conversation (prototypical spoken texts). Spontaneous spoken language is the means of communication typically used in the family circle, among friends and close people, in cases in which the speaker is minimally self-conscious about the formal attributes of his/her speech. It is often characterized by no preparation, improvisation, dialogue, and informality.

In spoken Czech, there is a relatively strong difference between spoken and written Czech on the one hand and between spoken varieties used in formal and informal situations on the other hand. In general, we can say that written Czech is associated with standard Czech, whereas in the case of spoken language we choose between standard Czech and other (non-standard) varieties of spoken Czech. One of the important differences between spoken and written Czech is based on their perception and production. Unlike written language, spoken language does not permit the speaker to go back and correct himself/herself without a trace; it is anchored in time and often allows little prior preparation. Its perception and interpretation is therefore much more constrained by the linear succession of individual elements in utterances.

In our study, we will focus on three examples typical for spontaneous informal spoken Czech (resulting from the character of spoken Czech compared with the written Czech and from the character of communicative situation):

1. Lexical fillers prostě/SIMPLY, JUST, vole/MAN, (YOU) IDIOT

   Fillers are often used to gain or play for time in the process of speech production, when the speaker is at a loss for how to continue or formulate his/her utterance, or cannot recall the appropriate words, interrupting the speaker with a new speaking turn. A characteristic feature of lexical fillers is semantic bleaching, i.e. the loss of their original meaning.

2. Formally reduced pronunciation variants of frequently occurring words (as exemplified on protože/BECAUSE and člověče/MAN, DUDE)

   They could show lower demands on speakers in terms of the form and precision of their utterances. The formal attrition itself can be an indicator of semantic bleaching in progress, which can ultimately lead to a word being used as lexical filler.

3. [v]-prothesis It affects certain words beginning in [o] and is originally a dialect feature of the west of the Czech lands. It is currently a stylistic marker signalling a lower register, and potentially stigmatizing,
depending on the context in which it is utilized.

In a formal context, all three of these items would be considered stigmatizing by peers.

The material used to demonstrate these features consists of publicly accessible corpus data on informal communication in Czech collected within the years 1988–2011, particularly the ORAL2013 corpus, which provides the most up-to-date material of this sort from all over the territory of the Czech Republic. The recordings and transcripts in this type of corpus must fulfil the following criteria: informal setting, dialogical nature, non-public and unofficial communication situation, and no prior preparation/unscriptedness. For this study, we selected the corpora PMK, BMK, ORAL 2006, 2008 and 2013. Individual utterances in all of these corpora are annotated for the following sociolinguistic factors describing the speakers: sex; age group (under vs. over 35 years old); and highest education achieved (non-tertiary vs. tertiary).

In general, this paper aims to show possibilities of using the corpora of spoken Czech and to contribute to broader description of Czech which has been primarily based on written language (regarding prostě, opinions vary concerning its function and part-of-speech classification – e.g. Slovník spisovné češtiny pro školu a veřejnost (The Dictionary of Standard Czech for Schools and the Public) considers it to be only an adverb, while according to Frekvenční slovník češtiny (Frequency Dictionary of Spoken Czech), it is only a particle. In our study, we confirm the loss of the original adverbial meaning and consider prostě either as a particle or a filler word). The results will be compared on the basis of the above-mentioned sociolinguistic categories.
Distance, knowledge and axiology in TV news discourse
Monika Kopytowska, Distance, knowledge and axiology in TV news discourse

The significance of evaluation for legal discourse and, especially, judicial discourse can hardly be overestimated. Indicating an attitude towards a legal entity, process or interactant is inherent in the acts of persuasion and argumentation, both being an integral part of judicial discourse. A substantial part of judicial opinions involves expressing agreement or disagreement with decisions given by lower courts, opinions expressed by counsel representing the parties, as well as the opinions arrived at by fellow judges from the same bench. Based on a substantial corpus of US Supreme Court opinions, this paper investigates linguistic strategies employed by judges to evaluate epistemic status (Hunston 2011) of propositions contained in the arguments of other legal interactants. Combining both the quantitative (corpus linguistics) and qualitative (discourse analysis) approaches (Partington 2013), the paper explores the relationship between status and attribution by examining instances of the averred alignments of propositions, which may include their labelling as hypotheses, possibilities or facts. This paper is structured around three major points. First, it argues that judges exercise their subjective judgement when assigning status to propositions contained in the argumentation of other legal interactants (e.g. lower court judges, fellow judges, prosecution) and thus realize the basic function of evaluation. Secondly, the assignment of status is a matter of not only legal but social significance as evidenced by the degree of (social) media attention paid to US Supreme Court rulings in socially important decisions (e.g. same-sex marriages). Finally, status is intrinsically linked with positive or negative evaluations and this seems particularly true for judicial opinions, which are both part of a knowledge-building agenda and they seek to influence human conduct using rationality as a means of persuasion.

References


Time as a line and as a clock in signed languages: a view from cognitive linguistics
Krzysztof Kosecki, University of Lodz

Signed languages are fully capable of expressing the concepts of temporal succession and cyclicity. The former is effected by means of signs employing time lines, which are “spatial constructs within signing space that represent distinct types of temporal orientation” (Emmorey 2002: 108); the latter is realized by means of clock-signs. The paper is an overview of time line and clock signs present in various signed languages. It accounts for fine distinctions among them and shows how metaphors and metonymies in their structure reflect the cultural conceptions of time.

**Keywords:** clock time, culture, time, time line, metaphor, metonymy, signed languages.

References

Feminine derivatives in National Corpus of Polish

Katarzyna Krasnowska, Polish Academy of Sciences
Witold Kieraś, University of Warsaw

The presented study is an experiment on systematic formation of feminine derivatives based on personal masculine (m1) Polish nouns, e.g., nauczyciel -> nauczycielska. Instead of searching for such forms in a corpus and then performing an analysis of the retrieved material, we adopt an alternative approach. We first automatically generate a large list of feminine derivatives and then examine its corpus coverage. The experiment is based on two resources: the database of Grammatical Dictionary of Polish (Słownik gramatyczny języka polskiego, SGJP) and National Corpus of Polish (Narodowy Korpus Języka Polskiego, NKJP).

SGJP is used as a source of an exhaustive list of Polish masculine nouns together with their inflectional pattern. An inflectional pattern specifies what suffix of the noun’s lemma should be removed to obtain its stem, and what suffix should be appended to the stem to form each inflectional form. To generate feminine derivatives, we design a number of simple, hand-written derivation rules. Each rule consists of (1) masculine inflectional pattern and optional ending of the lemma (determining the group of nouns to which the rule should be applied), (2) procedure of forming the feminine derivative’s lemma (this requires specifying two suffixes, the one to be removed from the m1 noun’s lemma and the one to be appended in its place; e.g., nouns inflecting like Bolończyk, Pekińczyk require removing -ńczyk and appending -nka) and (3) the feminine inflectional pattern of the resulting feminine noun.

With about 150 such rules, we manage to cover a vast majority of masculine personal nouns noted in SGJP, leaving out the ones which apparently have no feminine derivative. Some rules allow to form more than one derivative (e.g., urzędnik -> urzędniczka/urzędnica), some stylistically marked derivatives are intentionally not formed (e.g., poet -> poetka, but not poetessa).

We then examine the hypothetical and actual productivity of our rule system (it is clear that some of the generated derivatives are purely „hypothetical”, e.g. celnik -> celnica). The hypothetical productivity can be easily measured by taking the number of masculine nouns to which each rule was applied. In an attempt to verify the actual productivity, we use two approaches. The first one is to compare the list of feminine derivatives produced by our rules with the database of SGJP and see how many of them are actually noted in the dictionary. Another way of examining the actual productivity is to check how many among the generated feminine forms occur in NKJP. This can be achieved by generating, for each derivative yielded by our rules, a query finding its occurrences in a corpus. The queries are run on the balanced, 300 million word subcorpus of NKJP using the Poliqarp corpus query engine.
Language Change in Video Game Instructions? - Testing Frequency Effects in Corpus and Experimental Data

Anne Krause, University of Freiburg

The present paper investigates how frequency effects are involved in a morphological change-in-progress in German. This change concerns the formation of the imperative singular in strong verbs with vowel gradation. In this paradigm, the established i-stem-formation is being replaced by an analogically built e-stem form; for example, the respective forms of geben ‘take’ and befehlen ‘command’ can be found today as geb(e)i and befiehl(e)i alongside gib! and befiehl!.

The dataset for this study is based on a specially-compiled internet corpus of “walkthroughs” (approx. 7 million words), i.e. step-by-step instructions for video games, containing a high number of singular imperatives. All instances have been annotated for a variety of frequency measures and other potentially influential factors, including intraparadigmatic frequencies and the authors’ social metadata. A number of mixed-effects regressions were applied to these data, employing the R function ‘glmer’.

The analysis has shown that analogy and entrenchment are strong predictors for the trajectory of the change. In this respect, morphological change in German exhibits similar frequency effects as have been established for other languages: lower frequency forms are weakly entrenched in the mental lexicon and thus more prone to analogical change than higher frequency forms whose stronger mental representations induce a conservative effect. Verb paradigms of higher type and token frequency than the strong verbs affected by the change serve as a model for the analogy.

However, also context-related factors such as priming feature prominently in the explanation of the current state of affairs. While a few of the lower frequency verbs are found almost exclusively with the unsuffixed i-stem-imperative, the proportion of suffixed e-stem-formations in the imperative of geben (the most frequent verb in the paradigm) is higher than expected. In a large part, this distribution can be explained by taking into account a predefined window of left context of the imperatives contained in the dataset. The occurrence of suffixed singular imperatives within this window triggers suffixation and ensuing stem vowel change in the imperatives under investigation.

In a second step, the project investigates speakers’ awareness of the phenomenon in the form of an experimental testing of the perception and processing of the imperative forms. Not only will the stimuli for the experiment be based on the actual corpus texts, thus ensuring authenticity of the material presented - in keeping with usage-based methodology. Stimuli creation will also follow the predictions from the regression models, hereby enhancing the predictability of subjects’ reactions. When encountering a suffixed e-stem imperative in a stimulus text, subjects can be expected to react differently, depending on which of the three imperative options (unsuffixed
i-stem, unsuffixed or suffixed e-stem) is predicted by the model to occur in the particular slot.

On the one hand, the present analysis solves the question of whether the phenomenon under investigation is properly described as ‘linguistic variation’ or as a genuine ‘change-in-progress’. Thus, it evaluates the explanatory potential of the annotated variables with regard to their impact on language change or the actuation of change. Among numerous other applications, the analysis of such corpus data as the Walkthrough Corpus can serve to capture linguistic variation and change at a very early stage and provide a very powerful tool for compiling experimental stimuli, hereby clearly reducing the amount of pre-testing necessary in empirical studies.
Hyperspace representations and free associations: comparison of two methods in a study of loan-word and native-word synonyms

Adrianna Kućinska, University of Warsaw
Joanna Rączaszek-Leonardi, University of Warsaw
Miroslaw Bańko, University of Warsaw
Marcin Tatjewski, Institute of Computer Science, Polish Academy of Sciences

There are multiple factors influencing loanword reception and adaptation in a language. A particularly interesting situation arises when a loanword is stably used in a language despite the presence of close synonyms. Our project aimed at uncovering some of these factors using a triangulation of methods: 1. corpus-based human-performed linguistic analysis, 2. computational corpus analysis and 3. psycholinguistic experiments. The main goal was to assess the degree of the synonimity of the pairs, if there are systematic differences, which can be detected for particular pairs, and if there are differences that can be generalized for loan-word native-word pairs.

In the present study, we predominantly focus on the results of two methods, namely psycholinguistic data yielded by participants producing free associations to loan and native words synonyms and the analysis of the multidimensional representations of these words extracted from large corpora. We analysed 35 near-synonymic lexical pairs that consisted of a synchronically foreign word and its native synonym.

Free association data were collected using association questionnaires. Data were coded for item analysis: i.e., for each word, a list of 66 associations was compiled and frequency distributions were calculated on this basis. For each word pair, an index of semantic proximity was computed as the proportion of the associations shared by both members of this pair to the overall number of associations (shared or not) given for them.

A matrix-based semantic model of Polish language computed on a large-scale text corpus has been created, using the method of Correlated Occurrence Analogue to Lexical Semantics (COALS) implemented in S-Space library [3, 4]. COALS is one of the methods of distributional semantics. It belongs to the same family of tools like Latent Semantic Analysis or Hyperspace Analogue to Language. We decided to apply the COALS method on National Corpus of Polish (NKJP) as this is one of the biggest corpora for Polish language (1.8 bln segments) [1, 2]. Particularly, we decided to use NKJP’s balanced version (300 mln segments), as its balanced version is the only balanced corpus for Polish language that is big enough for distributional semantics requirements. Word relationships in the produced computational, matrix-based semantic model were explored using cosine similarity of word-vectors. The following results were analysed for the investigated synonyms:
• Similarity scores for each synonym pair.
• Nearest semantic neighbours for each word.

We present the synonymity indices obtained from these two methods, which provide interesting data that may aid linguistic analyses, but we also relate the two methods by correlating the results. Independence test based on Spearman’s correlation yielded that with p-value at 0.06 we can reject the hypothesis that the index of semantic proximity computed from free associations and the similarity scores from matrix-based model are independent.

Analysis of synonyms’ free associations and synonyms’ nearest neighbours from the matrix model clearly showed on a synonym-pair level that both methods agree on the usage differences between the synonyms. Further quantification of this result is planned.

The study brings information not only about the potential systematic differences between the loanwords and their native synonyms but also gives insight about the relation between the corpus-based matrix measures and psychological free-association measures of similarity.

References

Grammars or corpora? Who should we trust

Dario Lecic, University of Sheffield

Morphological doubletism (also called overabundance, morphological competition by different authors) occurs when there are two (or more) morphological variants for one grammatical cell. Slavonic languages, with their rich morphological systems, exhibit this phenomenon to a great extent, and Croatian is no different in that respect. Unfortunately, grammar books and other reference works of Croatian do not deal with this phenomenon to a significant degree. Most of them adduce lengthy lists of lexemes that allow two or more alternative forms as opposed to those that do not. However, these lists often differ between different reference works. For instance, if the speaker wishes to find out what the Instrumental singular of masculine nouns and personal names that end in -io (such as radio, Mario etc.) is, he would find the following possibilities: the grammars by Težak-Babić (2003) and Raguž (1997) allow only the forms in -ijem, whereas the grammar by Barić et al. (2005) and a dictionary by Anić (2003) allow only -iom. The most recent orthographical manual of Croatian (Jozić 2013) gives both options equal status. The problem with Croatian reference works is that they are not based on real-life data (or if they are, they only rely on the classical works of Croatian literature) but rather on the authority of the grammarian himself.

In this research we compare the recommendations found in grammars with corpus data. There are numerous examples of grammars saying one thing and corpus data painting a completely different picture. We will give a few examples. In the Instrumental singular, nouns of a similar phonological structure (ending in the same phonological segment) show different preferential patterns for the two possible endings (/em/ and /om/) ranging from 100

So where does that leave the native speaker? Whose rules and patterns should s/he abide by? We are especially interested in items which are not sanctioned in the grammars but do appear in the corpus with low-to-middling frequency. At which point do such items become a part of our mental grammars? Bybee (2010) claims that speakers can register a new item with only one exposure. Contrary to this, McEnery and Wilson (2001: 15) say that “human beings have only the vaguest notion of the frequency of a construct or a word”. This research attempts to determine to what extent are speakers in fact sensitive to frequency. The corpus frequencies were transformed into relative proportions, which were then used as a predictor variable in several acceptability studies. The results show that speakers are more sensitive to the information found in the corpus than to those found in the grammars. We hope for this research to show that Croatian is in dire need of a grammar that pays more attention to the frequency patterns found in large-scale corpora.
References

Temporal and aspectual properties of events: degrees of deverbalization in Polish and English modifiers
Barbara Lewandowska-Tomaszczyk, University of Lodz

The main issue to be discussed in the present paper is the asymmetry (Lewandowska-Tomaszczyk 2008, 2014b, Dziwirek & Lewandowska-Tomaszczyk 2010) between an Event-semantically conceptualized in terms of a finite clause and that expressed by a non-finite construction, particularly by deverbal modifiers. The asymmetry observed is a consequence of juxtaposing of two (or more) events as used in the main (finite clause) and the events-described in terms of de-verbalized categories, used as weakly-grounded or non-grounded forms with a decreasing assertive force (Cristofaro 2003). In such constructions the event represented by the modifier is either more fully attributivized and used pre-nominally or left in its more verbal shape, frequently following the relevant nominal. Such construal types can vary in one language and cross-linguistically and are evident in a number of language-specific morphosyntactic parameters and their interrelationship. Their semantics is considered to be the function of both conceptual content of an expression and its construal (Langacker 2000:45-46), in which it is assumed that alternate profiles of expressions can evoke the same conceptual content. The profiles involve a number of relevant parameters discussed such as the nominal categories, i.e., articles and gender, case, and number marking, or else the verbal ones manifested as person, tense, aspect and mood. The position of the modifier with respect to the modified noun and the presence of the complementary sentence elements such as the original direct object or adverbial(s) are also relevant indicators of the particular construal parameters.

The focus of the paper is the asymmetry in the semantic interpretation of the morphosyntactic structure of classes of deverbal modifiers and their sentence positions. The language data are generated from the British National Corpus and the National Corpus of Polish (www.nkj.pl). Their frequencies and sentence and discourse distribution are examined and discussed to provide background for distinct language profiles in this respect. In the case of classes deverbalized to different degrees, which present a variety of functional and distributional patterns cross-linguistically as in e.g., cook – cooking – cooked in English and gotować – gotujący/gotując/gotowanie – gotowany - ugotowany in Polish, the expressions’ profiles (participles, gerunds functioning as different categories of modifiers) are determined by the grammatical categories of the words. It is precisely the differences in their construal that underlie their temporal, aspectual and other semantic properties and discourse interpretation (see Lewandowska-Tomaszczyk 2014a for the relevant construal differences of apprehensive verb constructions in English and Polish).

Distributional and structural differences between Polish and English in this respect are demonstrated to pertain first of all to the distinct aspectual systems in the two languages and embrace in particular pre-/post-modification
behaviour of the relevant constructions.

Acknowledgement: Research carried out within COST Action TD0904 TIMELY, supported by National Science Centre (NCN) grant No 2011/01/M/HS2/03042, Perception of Time as a Linguistic Category.

References

The persistence of stereotypes. A corpus-based analysis of the representation of rheumatoid arthritis in press discourse

Maria Libura, Lazarski University
Pawel Bechler, Warsaw University

There is an increasing recognition that language and linguistic construal play an important role in shaping the experience of health and illness, eventually impacting on patients’ behaviour. This has led to steady growth in the number of studies dedicated to medical and healthcare discourses, conducted both within the health sciences and linguistics (cf. Bernhardt 2004; Libura 2007; Budgell, Miyazaki, O’Brien, Perkins, Tanaka 2007; Mil- lar, Budgell 2008; Adolphs, Atkins, Harvey in press). Studies in this field have both theoretical and practical dimension, as they may inform actions within the healthcare system that address disparities between the clinical reality of a disease and its social perceptions, thus improving patient’s compliance and treatment outcomes.

This paper critically examines the representation of rheumatoid arthritis (RA) in the print media in Poland. RA is an autoimmune disease that leads to a chronic inflammatory disorder, which may affect many tissues and organs. Its progress results in disability, yet its common perception is that of a minor health issue. We hypothesized that - as a chronic non-communicable disease - RA will occupy a non-central place in the category DISEASE and thus exhibit some peculiar features, such as deviations from a prototype scenario of a disease; as a result, its representation may be incomplete and indistinct.

To reconstruct the media image of RA we conducted a corpus based analysis of the coverage of this condition in three daily newspapers (Gazeta Wyborcza, Rzeczpospolita, Nasz Dziennik) and three weekly magazines (Polityka, Wprost, Gość Niedzielny) in 2013. All references to the disease were retrieved and the context of their occurrence classified and coded; the revealed themes were subsequently analyzed and compared to the RA image emerging from the National Corpus of Polish on the one hand, and the clinical picture of the disease on the other.

The analysis confirmed the initial hypotheses; it also revealed blind spots in the representation of RA in the press, as well as substantial incongruities between the media image of the disease and its medical descriptions. Surprisingly, the representation of RA in the mainstream press reflects folk theories rather than medical knowledge. Certain aspects central to the clinical picture of the disease, such as progressing disability, were virtually missing from the press representation. Interestingly, and contrary to clinical practice, the disease was predominantly portrayed as a minor ailment treated with herbal infusions and bath salts. We argue that the peripheral status of RA in the DISEASE category is mainly responsible for the observed phenomena.
References


Evidentiality in Mandarin narrative discourse: A corpus-driven approach

Lu Lu, SOAS

This study employs a text-driven approach to address evidentiality, which is concerned with the evidence that a speaker/writer has for making factual claims (Chafe 1986). Evidential expressions mainly focus on issues like ‘who is the source of the information’, and ‘what is the basis (e.g. general knowledge, perception and such) of the speaker’s/writer’s knowledge’. This concept has been largely applied to languages which have grammaticalized evidentiality in the morphological system, such as Albanian and Northern Iroquoian (Chafe and Nichols 1986; Johanson and Utas 2000). However, the research exploring morphologically less rich languages such as Mandarin Chinese has been somewhat neglected. A good source for such studies would be narrative discourse. The researchers thus used Mr Bean in the Swimming Pool, a five-minute video clip, to elicit data (Pavlenko 2008). Ten Chinese students were asked to narrate the story as it played. Having manually annotated the transcriptions from the obtained spoken data, the researcher went through the corpus to seek the evidential expressions without categorising them on a priori grounds.

This research then analysed the evidential expressions in terms of their sources, the bases of his/her knowledge and the interplay between the two aspects. The results show that most of the evidential expressions are attributed to others in narrative discourse. Among the propositions which are derived from the narrators themselves, a large proportion of them are inferred from the narrators’ sensory perception, with seeing ranked the highest, followed by the evidence from general knowledge and self-evident expressions. There is a tricky category of the evidential basis, where expressions like kan (lit. ‘look at’) integrate both sensory and mental perceptions in Chinese narrative discourse. Mandarin narrators also demonstrate a high degree of certainty in relation to their evidential expressions. The findings of this research demonstrate that morphologically less or not rich languages like Chinese have their own unique ways of expressing the assessments of knowledge. This study also opens up an arena for future projects regarding evidentiality across different languages and in different text types.
References


Standardization in safety data sheets? A corpus-based study into the problems of translating safety documents.
Alekisandra Makowska, University of Lodz

In industry, production is organized into standardized processes to ensure the satisfactory level of the output. Apart from the machine stock, substances used in the process are also subject to standardization. For this reason production companies are particularly dependent on the technical support from their suppliers due to the use of chemical substances in the process of production. In order to use, handle and store the chemicals safely producers and their representatives are legally obliged to provide material safety data sheets, where clear and precise information is indispensable, according to REACH and CLP regulations in EU and, GHS on the global scale. The European Standard for Translation Services EN-15038 provides quality requirements for specialised translators which should be reflected in the translation process of analyzed safety documents. However, the scrutiny of safety documents reveals that the performance does not reflect the ideal: language of the original version is not adjusted to the end user, there are different versions of the same safety data sheet in different languages available at the same time, as well as different and even contradictory information.

For the purpose of this study, the author gathered and analyzed a corpus of 93 material safety data sheets available online in the Polish, English and German language for the respective products in order to spot the potential areas of the most frequent problems and to provide possible solutions to problems. The author offers concrete solutions to reach a high level of standardization in the process of translation of material safety data sheets and to produce a set of guidelines for the successful translation that would serve their purpose and be useful for target readers.
A corpus perspective on the football played in Brazil and the U.K
Sabrina Matuda, University of São Paulo

Football is the most popular sport in the present century. It has assumed a role beyond that of a national sport by becoming a cultural manifestation. Football relations across countries have increased significantly over the past decades.

In order to regulate these relations, we need to express ourselves through language. However, each culture has its own way of playing and of supporting its teams, expressed differently according to each mother tongue. The problem arises when there is an urge to express these particularities in a foreign language.

To cope with this problem, bilingual up-to-date and comprehensive reference materials as far as football terminology is concerned would be very helpful. This is especially true in Brazil nowadays, where major sports events like the World Cup and the Olympic Games are being held. Consequently, there is an increasing demand for translations of football news. However, in spite of the great demand, the area lacks terminological publications, especially bilingual materials which would provide professionals with appropriate terminology.

This presentation sets out to show how different football styles, the history of football in Brazil and in England, the cultural appropriation of the laws of the game practiced by each culture and other historical social factors have influenced the creation of the football lexicon both in English and in Portuguese. To that aim, this study draws on the principles of three related disciplines: Corpus Linguistics, Textual Terminology and Technical Translation. Corpus Linguistics has been of inestimable value to terminological studies as it allows for the possibility of studying linguistic features across a large number of texts, and thus provides a basis for generalizations across languages. According to Textual Terminology principles (Krieger & Finatto 2004), specialized texts are the central object of study of Terminology and textual type recognition and the culture in which the text is embedded play an important role in term identification. Technical translation is viewed as a communicative act subject to cultural constraints (Azenha 1999). Finally the anthropological concept of form-representation (Toledo 2002) is called upon to elucidate such differences. According to Toledo, the laws of the game do not determine the ways of playing. It is the cultural interpretation and appropriation of the laws by each country that creates different “forms of playing”. When juxtaposed, the laws of the game and the “forms of playing” give rise to different “representations”, that is, the empirical observation of the “different forms of playing” at a symbolic level, created by us, which consolidates various styles of playing (e.g. the English school and the South American school).

Our corpus consists of approximately two million words - 1.002.897 in En-
English and 917.073 in Portuguese. Each corpus is divided into three subcorpora: newspaper reports on match results, live minute by minute commentaries and live commentaries by sports journalists and by football fans via Twitter and Facebook. The analysis is being carried out semi-automatically on tagged corpora. After compiling and balancing the corpus, it was tagged with the Tree-Tagger. Helmut Schmid’s Tree-Tagger was used to tag the corpus and the analysis was carried out in the Portuguese-English direction using the WordSmith Tools suite.

Our preliminary results show that terms usually occur in context rather than as isolated items living a life of their own and that the socio-cultural and historical context within which they are embedded affects their meaning.
Evaluating the Libya crisis: through the eyes of journalists writing in English and Italian

Amanda C. Murphy, Universita Cattolica del Sacro Cuore

This research compares evaluations of the protagonists of the Libya crisis that exploded in 2011, as expressed in the Italian and British press. Specifically, it concentrates on evaluations of the protagonists in the conflict as they emerge in news articles from The Times and Il Corriere della Sera, two comparable quality newspapers from Britain and Italy, characterized by a centrist political stance, and similar circulation figures. Theoretically, it considers the role of emotion in evaluation and, in tune with literature on emotion from a linguistic point of view (Bednarek 2008), observes that it is rare to find an emotion expressed without an evaluation. By investigating evaluation in news articles, the paper, in agreement with Conboy (2007), also implicitly criticizes the view that hard news articles are written in an objective, neutral style. Lexical choices, as well as the point of view from which an article is written, encode evaluations in this text type.

The study draws on two, small, comparable corpora: The Times corpus, consisting of 186,284 tokens, and the Corriere della Sera corpus, consisting of 153,615 tokens. The news articles in both corpora were collected via LexisNexis and refer to the period between February and November, 2011. They all contain the word Libya/libia and refer to the crisis.

Evaluation is known to be pervasive in text (Thompson and Hunston 2000, Hunston 2011), and the study illustrates one method of examining evaluative strategies across languages. This consists in choosing one topic and investigating the surrounding evaluative phraseology and lexical choices in the different languages. The image conveyed of the protagonists of the Libyan crisis – Ghedaffi, the rebels/ I ribelli, the loyalists/ I lealisti, Nato and the UN/Onu – is compared in both corpora, drawing methodologically both on transitivity analysis (looking at the material, relational and verbal processes as identified by Halliday 2001, as they occur around the protagonists investigated), and using the categories of Martin and White’s Appraisal Theory (in particular, the sub-system of attitude). The results show cultural differences in the newspapers’ attitude towards certain protagonists (e.g. Ghedaffi, the UN and NATO are evaluated less negatively in the Italian corpus than in the British one), as well as differing choices in the intensity of emotion expressed in the British corpus, particularly referring to the rebels. Overall, it would appear that the journalists in the British corpus show a tendency to express more emotion than their counterparts in the Italian mirror corpus, also through a greater use of attribution of emotion to third parties, through quotations.
References

Studying ellipsis alternation in English corpus data
Joanna Nykiel, University of Silesia

Current models of grammar recognize that there is gradience in grammar. What speakers
know is not only whether a given construction is well-formed, but also how
likely it is
to be used, given the context and alternative ways of conveying the same
message. Some
models of grammar already include both categorical and ‘soft’ constraints
(Bresnan et al.
2001, Bybee and Hopper 2001 and Manning 2003), with the latter well-
suited to capturing
gradient data in syntax that is regularly found through corpus studies and
psycholinguistic
experimentation (Bresnan 2007, Bresnan and Ford 2010, Nykiel 2013, Wolk
et al. 2013).

Many of these soft constraints are grammatical constraints that have been
reanalyzed as
constraints motivated by independent principles of language processing
(Gibson 2000, Wasow
recognition
that grammar is shaped by principles of language processing calls for a
better understanding
of which aspects of grammar are shaped this way and how this process
evolves though
time. One way of exploring this issue is by studying cases of syntactic varia-
tion (Hawkins
2004, 2009). In this paper, I discuss what I term ellipsis alternation, shown
in (1)–(3).

Ellipsis alternation refers to the possibility of using remnants with preposi-
tions, as in (1b),
(2b), and (3b), or without, as in (1a), (2a), and (3a), in elliptical construc-
tions.

1. A: Pat saw me riding on the bus. B:
   (a) What bus?
   (b) On what bus?
2. A: What bus did you see Pat riding on? B:
   (a) None.
   (b) On none.

3. Pat won’t ride on any bus, let alone
   (a) an old one.
   (b) on an old one.

Modern quantitative methods (especially regression modeling) permit an in-depth analysis
of instances of syntactic variation extracted from corpus data in terms of which variants
are likely to be used in which contexts. For this study, I extracted instances of ellipsis alternation
from three corpora of spoken American English and developed a logistic regression
model of the data predicting the appearance of remnants without prepositions (as opposed
to remnants with prepositions). The model demonstrates that the probability that remnants
without prepositions will appear increases if (1) their antecedents are highly accessible, (2)
they are semantically dependent on the lexical heads present in the antecedents, and (3)
the correlates are discontinuous rather than continuous (compare the continuous correlate
on the bus in example (1) and the discontinuous one What bus...on in (2)). These patterns
may be subsumed under three independent principles of language processing. First, the
semantic and morphosyntactic content is reduced in those anaphors that have accessible
antecedents (Ariel 1990). Second, constituents that the parser must access together as it
computes their meanings are not separated (Hawkins 2004). Third, structures seen in previous
discourse are replicated in later discourse through priming or structural persistence
I discuss the significance of these results in the context of existing cross-linguistic research on ellipsis alternation and conclude that the grammar of English has in this case been shaped by principles of language processing, and that ellipsis alternation is a clear instance of gradient data.
Weblog, often called an “electronic journal”, is a simple website containing posts displayed in the reversed chronological order, which can usually be commented by other users (cf. McNeil 2003). It is not a simple recreation of a journal in the digital environment but rather a complex remediation of traditional genres (e.g. journal, diary, letter) under the influence of electronic discourse, which entails such phenomena as secondary orality and interaction with the user (cf. Davis and Brewer 1997).

Although blogging is a relatively new phenomenon and the scholarly interest in this form of writing could be traced back to the first years of the 21st century, there is already a substantial literature dedicated to the genre analysis of blogs. We may distinguish two main threads of scholarship on this field: one focusing on the relationships between blogs and the other forms of Computer-Mediated Communication (e.g. Herring et al. 2005; Devitt 2009), and the other tracing the „ancestry” of the genre in such forms as journals (e.g. Serfaty 2004; Miller and Shepherd 2009).

The aim of this study is the linguistic analysis of weblog genres, which would support earlier interpretative studies of the collected material (e.g. Maryl and Niewiadomski 2014). In this study we analyse the corpus of 260 popular Polish blogs. Construction of the blog corpus is not straightforward as blogs can very differ in the way of using HTML structures for presenting the textual content. In order to avoid laborious manual editing, a system called BlogReader was constructed. It is dedicated to corpus acquisition from structured web sources. It was built as an expansion of a corpus gathering system for based on several open components, e.g. justText and Onnion (Pomikálek, 2011). BlogReader works in a semi-automatic way: users first specify elements of the HTML structure that include text fragments that are relevant for the users. Identified fragments are next automatically acquired and saved in the selected corpus format.

The analysis concerns three aspects:

- Linguistic comparison of blogs and other forms of discourse (e.g. literature, news)
- Linguistic analysis of weblog genres, i.e. linguistic differences between particular groups of blogs and authors
- Linguistic variation between blog posts within individual blogs

Blog corpus analysis will be based on an open stylometrics system dedicated for Polish which is built as a part of the CLARIN-PL project (www.clarin-pl.eu). The system will enable description of text documents with features referring to any level of the linguistic structure: from the level of word forms up to the level of the semantic-pragmatic structures. The system will
combine several existing components: language tools for pre-processing, Fextor (Broda et al., 2013) – a system for defining features in a flexible way, Stylo (http://cran.rstudio.org/packages/stylo/versions/34587) - a stylometric package for English, SuperMatrix (Broda & Piasecki, 2013) – a system for building and processing very large co-incidence matrices (e.g. documents vs features) with interface to clustering and Machine Learning packages. With the help of the system we want to look for different groups of blogs and also for features (e.g. lemmas or lexico-syntactic patterns) that characterize different text groups.

This paper is a part of the ongoing research project “Blog as a new form of electronic writing” (2012-2014) funded by the National Science Center in Poland and conducted in the Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences. This work is also partially financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

References
Clancy Ratliff and Jessica Reymann. Minneapolis: University of Minnesota Libraries.


**Temporal Expressions in Polish Corpus KPWr**

**Jan Kocoń, Michał Marcińczuk, Marcin Oleksy, Tomasz Bernaś, Michał Wolski**

Temporal expressions are important for comprehensive information extraction task. They provide information about time, duration and frequency of some actions described in a text. In order to create robust tools for temporal expression recognition a manually annotated corpora are required. According to our best knowledge there is only one corpus for Polish annotated with temporal expressions called NKJP (Przepiórkowski et al., 2012), however, the scope of the annotation is limited only to dates and times. In this article we present a temporal expression annotations in the KPWr corpus (Broda et al., 2012) based on The Markup Language for Temporal and Event Expressions (henceforth TimeML; Sauri et al., 2006). The guidelines define four major categories of temporal expressions, i.e. date, time, duration and set. For date and time we annotate both absolute (specific date or time) and relative (“after 2 days”, “1 hour ago”) expressions. The set category refers to action frequency (i.e. “every two day”). We have also added a fifth category called span, that is used to annotate date and time ranges (i.e. “from Monday to Friday”, “11:00-13:00”). In the paper we present the adapted TimeML guidelines to Polish language, the results of manual annotation and inter-annotator agreement using positive specific agreement measure (Hripcsak and Rothschild, 2005). We have annotated 1637 documents from the KPWr corpus, which contains 4395 date, 635 time, 934 duration, 143 set and 388 range annotations. The corpus will be used to build and evaluate tools for automatic recognition of temporal expressions in Polish texts. In the future work the temporal expression annotations will be linked with event annotations in order to describe the temporal aspect of actions described in the text.

**References**


Light Verb Constructions in TICLE and PICLE: How should we accelerate their lexical priming in EFL Context?

Katarzyna Papaja, University of Silesia
Cem Can, Abdurrahman Kilimci, Cukurova University

Although there have been many studies examining the light verb constructions (LVCs) in the literature, Jespersen (1965) is usually credited for the coinage in the field. His fundamental argument regarding the distinguishing feature of such constructions is that the main semantic content of the predicate is provided by the nominal complement instead of the verb. Adapting a similar definition, Miyamoto (2000, p. 12) defines the light verb as a category “whose semantic content is “light” (or has little lexical meaning), as opposed to “heavy” (or lexically more specified), and much of the semantic content is obtained from its arguments.

In second language acquisition, these constructions have always been found problematic because of the impoverished exposure to L2 input and to a narrower range of language (Montrul, 1997; Oh & Zubizarreta, 2003). Moreover, as the L2 learner has previously been primed for L1, this interferes the process in which s/he is primed in English (Whong-Barr & Schwartz, 2002).

This particular corpus-based research study aims at investigating the use of verb + noun collocations of Turkish and Polish EFL learners in their argumentative essays in Turkish and Polish International Corpora of Learner English (TICLE and PICLE) as a sub corpora of International Corpus of Learner English (ICLE) and compare them to the ones in the argumentative essays written by native speakers of English in the Louvain Corpus of Native English Essays (LOCNESS). Within this context, the study aims at answering to the following three research questions:

1. To what extent do learners’ choices of LVCs to form English expressions correspond with native speakers’ choices?

2. Do learners’ L1 Turkish and Polish influence their acquisition of English LVCs?

3. What kinds of LVCs cause most problems to Turkish and Polish as represented in TICLE and PICLE?

In the last section, the paper emphasizes the role of transfer in SLA and presents some pedagogical implications to accelerate EFL learners lexical priming and some suggestions for further study.

References


Chronological Corpus of Polish Press Texts (1945–1954)
Adam Pawłowski, Wroclaw University

The Chronological Corpus of Polish Press Texts (henceforth Corpus) is an innovative resource consisting of thoroughly selected, representative, samples of Polish daily press, structured according to their publication time. Introduction of time variable allows users to discover and explore the dynamics of events and phenomena represented in daily press in long periods of time. In order to facilitate knowledge extraction (both linguistic and extralinguistic) from the data, the Corpus is annotated morphosyntactically in accordance with the CLARIN-PL standards and it will be provided with a “hinting” user interface. The interface will be gradually provided with scientifically effective tools of sequential analysis in the frequency and time domain (now in construction) (Pawłowski 2001).

The Chronological Corpus covers the period from 1945 to 1954. Despite being submitted to preventive censorship, large amounts of daily press texts from this period bear important cognitive value for potential users. The basic time spans are months, while the minimum time spans, visible from the user level, are subsequent weeks (there will be a possibility to aggregate them into longer periods). The user will be provided with a tool selecting journals’ titles or facets combined with periods. Facets will include journal’s titles clustered according to the target group (e.g. press for young people, for workers etc.) In the final version of the Corpus every month will be represented by ca 120 000 text words, i.e. ca 1.44 million words per year.

The Corpus has a strong multidisciplinary orientation towards the needs of users representing all the humanities and social sciences, mainly contemporary history, cultural anthropology, political science, not only linguistics. It will also offer dedicated tools of quantitative analysis, such as basic statistics and coefficients of lexical richness (now in construction). Text data (now available almost exclusively in a printed form) combined with automatic language processing tools are likely to foster qualitative, but also quantitative, studies in the disciplines where traditional methodology is still overwhelming.

The novelty of the Corpus consists in introducing explorative tools of sequential (chronological) text analysis. So far only qualitative analyses of this kind were realised on the Polish language (cf. Bajerowa 1988) and many of them were associated with diachronic studies. The Chronological Corpus will allow on-line modelling of historical dynamic phenomena in culture, politics etc., reflected in Polish daily press, by means of quantitative parameters (functional characteristics of trend and hidden periodical oscillations), which have not been applied in linguistics on such a scale yet. The Corpus will let the user discover, compare, validate or falsify the existence of unknown or hidden serial phenomena, e.g. political campaigns, natural and/or cultural cycles, sudden (catastrophic) phenomena (Pawłowski 2006, cf. also Brunet 1981).
One of the important features of the Corpus is a possibility of its future extensions in time and space. The extension in time means here gradual covering the periods 1945–1990, 1939-1945 and 1918–1939. The extension in space is also possible, since parallel data from other ex-Popular Democracies (e.g. Czechoslovakia, East Germany etc.) could be aligned with the Chronological Corpus of Polish Press Texts. Cross-correlations, differences and interactions between time-series representing press in various countries could be then discovered, validated or falsified.

References

Exploring prosodic patterns of discourse devices
Piotr Pęzik, University of Lodz
Ewa Guz, John Paul II Catholic University of Lublin

Research into spoken discourse has provided substantial evidence that prosody makes a significant contribution to the construction of meaning and coherence of communication (Couper-Kuhlen and Selting 1996; Couper-Kuhlen 2001; Wennerstrom 2001; Barth-Weingarten, Reber, Selting 2010; O’Grady 2010). Going beyond the level of a single utterance and adopting a top-down, discourse-oriented research perspective has allowed scholars to identify a range of prosodic features which are utilized by interlocutors to encode and decode meanings, functions and intentions in order to achieve conversational goals. Earlier accounts of the prosodic make-up of discourse units relied on structurally-driven descriptions of limited amounts of introspective data and sought to establish the relationship between grammar, intonation and function (Bolinger 1986; Brazil, Coulthard and Johns 1980, Brazil, 1995, 1997; Halliday, 1967). Others focused on the relationship between tone groups/units and discourse structure (Selkirk 1984; Pierrehumbert and Hirschberg 1990). Most recent studies utilised prosodically annotated corpus data and yielded systematic quantitative and qualitative descriptions of the prosodic patterns of recurrent sequences in spoken discourse (Ajmer 2002; 2004; Adolphs, 2008; Warren and Cheng 2008).

Depending on research objectives, the data used and overall theoretical orientation, the studies have adopted a variety of approaches to coding and segmenting the flow of speech, ranging from traditional orthographic word-based procedures to frequency-based extraction. The approach we adopt in this study is influenced by recent linguistic models (Sinclair 1990; Wray 2002), which suggest discourse is made up of single words as well as multi-word, holistically processed and stored formulaic chunks. Here, we follow O’Grady’s (2010: 8) assertion that ‘an information unit realized phonologically as a tone unit is a preassembled chunk.’ In fact, research has shown that some formulaic sequences display distinct prosodic patterns in terms of intonation (Peters 1983; Hickey 1993; Lin and Adolphs 2009), pausing and hesitation patterns (Wray 2004; Erman 2006, 2007; Dalhmann 2009; Lin 2010), stress placement (Bybee and Scheibman 1999; Bybee 2000, 2001, 2002; Ashby, 2006) and the location of the nucleus (Lin 2013).

Drawing on these two research traditions, this paper attempts to explore the hypothesis that specific discourse functions of recurrent single- and multi-word discourse devices (i.e. pragmatic and discourse markers, modal particles and stance expressions) found in conversational data may be systematically associated with distinct prosodic patterns. For the purpose of this paper we adopt the broader view of prosody which goes back to Crystal (1969: 3-5, 1970: 78) and covers a range of suprasegmental features which result from the interplay of pitch, loudness and duration in syllable-sized or larger segments and ‘expound contrast in meaning of an attitudinal, grammatical and social kind.’(1970: 8).
Using a corpus of 2 million words of casual conversational Polish, we generate a list of recurrent, independent n-grams in order to identify potential discourse-functional lexical items. One of the idiomatic clauses found in the list is *daj spokój*, which is prototypically used as a pragmatic marker with a clear perlocutionary force directed at the immediate interlocutor. A less obvious function of *daj spokój* emerging from the analysis of the conversational data is that of an indirect stance expression. We look at the degree to which this distinction can be associated with three prosodic features: pitch contours, intensity (normalised loudness) and duration. As a more practical implication of this study, we consider the extension of the Spokes search engine (spokes.clarin-pl.eu) (Pęzik 2014) with new functionalities which could facilitate similar research of the prosody of recurrent discourse-functional lexical units found in conversational corpora of Polish and English.
CLARIN-PL – a Polish Language Technology Infrastructure for the Users
Maciej Piasecki, Wroclaw University of Technology

Language technology infrastructure (LTI) (Váradi, et al., 2008) is a complex system that enables combining language tools with language resources into processing chains (or pipelines) with the help of software framework. The processing chains are next applied to data sources (i.e. selected language resources) in order to provide results that should be interesting for the researchers working with text or speech recordings. CLARIN (Common Language Resources and Technology Infrastructure) is a distributed network LTI focused on applications in Humanities and Social Sciences. The enabled processing chains are meant to be a basis for a number of research applications (i.e. research tools) in different subdomains. CLARIN combines centres, which provide language tools and resources, connected via internet and distributed across different countries. The main goal of CLARIN-PL is to build the Polish part of the infrastructure.

Language technology (LT) has been developed for more than 10 years. Its development was initiated by a shift from small limited systems characteristic for early NLP, to robust text processing technology based on a sets of exchangeable and reusable components: dynamic – language tools and static – language resources. However, LT still faces several barriers blocking its widespread use, e.g., as a background technology used in the methods of HSS, namely (Wittenburg et al., 2010):

- physical – language tools and resources are not accessible in the network
- informational – descriptions are not available or there is not means for searching
- technological – lack of commonly accepted standards for LT, lack of a common platform, varieties of technological solutions, insufficient parameters of the users’ computers
- knowledge – the use of LT requires programming skills or knowledge from the area of Natural Language Engineering
- legal – licences for language resources and tools limit their applications

CLARIN aims at removing most of the barriers, e.g. by providing network environments for users that do not require downloading and installing language tools or meta-data-based common search mechanisms. In the case of the knowledge barrier the goal is to lower it as far as it is possible, e.g. by providing to the users complete research applications hiding the details of the applied LT.

Practical applications of LT are possible if the quality of the language tools and coverage of the language resources is enough good. This problem is a
part of the technological barrier and it was also clearly visible in the case of LT for Polish. Thus one of the main goals of CLARIN-PL is to build lacking basic language resources and tools for Polish, e.g. in relation to BLARK quasi-standard (Krauwer, 1998). We plan to construct a system of large scale lexicosemantic resources, syntactic-semantic resources, annotated training text and speech corpora, aligned bilingual corpora, lexicosemantic bilingual resources, morpho-syntactic tools, syntactic and semantic parsers, tools for Information Extraction and stylometrics, to mention some of them. A concise overview of the work plan will be presented during the talk.

CLARIN-PL plans to provide all language tools accessible via web-based user interface. Moreover, a web-based environment for combing the tools and resources into processing chains will be constructed. However, this is still not enough to overcome the knowledge barrier. Potential users must possess detailed knowledge about LT to know why and how to use it. Thus, CLARIN-PL has established contacts with selected key users from H&SS in order to identify possible applications of LT in research, collect user requirements, designed prototype applications and, very often, to assist in development of new research methods in H&SS effectively utilising LT in spite of their limited accuracy. Several examples of applications will be presented during the talk. However, as the discovery process has been only started so far, we will also discuss our plans for the future continuous development of the CLARIN-PL LTI driven by the process of discovering and development of LT for Polish focused on research applications in H&SS.

References


A Multilayer System of Lexical Resources for Language Technology Infrastructure
Maciej Piasecki, Marek Maziarz, Michał Marcińczuk, Paweł Kędzia, Adam Radziszewski, Wrocław University of Technology

Semantic text processing (e.g. semantic text searching or Information Extraction) must be well grounded on the analysis performed on the lexical semantic level. Individual text words (at least selected) must be mapped on the repository of lexical senses to provide a basis for the processing of the lexico-semantic structures. The repository can be automatically induced from text corpora, but far better can be achieved if it is manually constructed. Thus, for the needs of the CLARIN-PL – the Polish part of the CLARIN infrastructure – we decided to build a system of lexical resources organised around plWordNet 3.0 as its fundament. The system comprises:

- plWordNet 3.0 – a huge wordnet for Polish providing description of lexical meanings (more than 260 000 planned) and the system of lexico-semantic relations (Maziarz et al. 2013),
- NELexicon 2.0 – a very large dictionary of Polish Proper Names (about 2.5 million PNs) classified into 102 semantic categories (Marcińczuk et al., 2013),
- mapping of the plWordNet senses to extended SUMO Ontology (Kędzia and Piasecki 2014),
- a lexicon of lexico-syntactic descriptions of multiword lexical units (about 60 000 descriptions, cf. Kurc et al. 2012),

PlWordNet is a huge network of lexico-semantic relations that provide description for lexical senses that are represented as the network nodes.

As plWordNet is exclusively focused on the faithful description of the lexical meanings, the common knowledge description is kept in a separated layer in the form of a combined upper level and middle level ontology. The mapping between the wordnet nodes and the ontology concepts provides also formalised interpretation for the lexical meanings.

Proper Names are an open class, are not a proper part of a lexicon and represent objects in the reality. That is why they are described in a separate layer described by 102 semantic categories related to the ontology and partially mapped onto the wordnet.

Multiword lexical units are semantically described as wordnet graph nodes, but the wordnet do not include the description of their lexico-syntactic structure that is necessary for their recognition in text (Radziszewski 2013). Thus, the whole system is supplemented by a dictionary of structural descriptions (Kurc et al. 2012). The last element – the mapping to Princeton
WordNet – goes beyond the monolingual system, has been described in (e.g., Rudnicka et al. 2012) and will be omitted here.

As CLARIN infrastructure is meant to provide support for text processing in Humanities and Social Sciences, one can expect practically any texts delivered to the system by the future users. Thus, in order to make the system of the lexical resources a feasible basis, we have to provide extensive coverage for word types and meanings. plWordNet 3.0 is built as a major extension of the previous version 2.0 and it final size was provisionally set to the size approaching the size of the Polish dictionaries ever made. plWordNet 3.0 will cover at least 260,000 lexical meanings that should correspond to about 200,000 morphological base forms (entry forms) described. Moreover, plWordNet has been built on the basis of the corpus as the primary knowledge source and its development was supported by a number of tools for extraction of the lexico-semantic knowledge from text. The most important aspects of the plWordNet, its development process and the present state (more than 214,000 senses and 152,000 base forms) will be presented in the talk and the paper.

plWordNet is stored in a network database and is accessible via web service. It can be exported to the XML representation based on the Lexical Markup Framework representation that allows for easy integration with the existing resources of the Linked Open Data initiative.

The described system of lexical resources has been applied in, e.g., Word Sense Disambiguation (e.g., Kędzia et al. 2014), Named Entity Recognition (Marciničzuk et al., 2013) Information Extraction (Marciničzuk and Ptak, 2012) and Open Domain Question Answering (Marciničzuk et al., 2013). The resource system and constructed language tools are used in several applications for the CLARIN-PL users, e.g. for text semantic classification or identification of text-to-text relations, that will be discussed in the talk.

References


Literary Criticism of Cartography and Indexing of Spatial Data and Semantic Relations

Maciej Piasecki, Wroclaw University of Technology
Aleksandra Wójtowicz, Institute of Literary Research, Polish Academy of Sciences

Combination of topographical analysis and literary thought has interested researchers since 90’s of the XXth century. Several projects were proposed which resulted in creation of maps illustrating writer paths and places mentioned in the literature works. In contrary to this, an opposite idea in which a map is a tool for the literary research is a relatively new proposal.

We propose a literary criticism of cartography method that originates from J. B. Harley’s expansion of the map deconstruction theory and also a redefinition of Karl Schlögel’s mental maps. The proposed method takes into account achievements of literary geography, cartography criticism and literary cartography. According to literary geography a map is used to interpret texts. Cartography criticism emphasises map textuality according to which a map is a creative work that can be interpreted and used as a research tool. Finally, literary cartography is based on mapping different spaces described in texts on the geographical space. An important common aspect of theses techniques is perception of a map as a representation of the reality of both kinds: real and fictional, that encompasses mutual relations between them. Development of a thematic map is always associated with a transfer from the geographical space mapping to the other aspects of the spatial representation concerning, e.g. politics, economy, culture or the language. Harley points to the “rules of the social order”, that interfere with codes and spaces of the cartographic transcription, as one of the ways of expression. We are going to reformulate this rule. Distance or a social role of the object is not longer the focus point, but we want to put the importance of the literary-cultural relations in the focus. Map scaling and a precise system of distance measuring are becoming secondary tools that facilitate building a model of the cultural relations. We pose a question – what are the tasks of the literary researcher in the map building and the analysis of the topographical dependencies performed from the perspective of the existence of intellectual societies (formal and informal) and their influence on the shape of a literary work. Cartography of the literary thought is also a new way of approaching the literary research, according to which philological reading (i.e. aimed at reconstructing circumstances of the work creation) is perceived in the context of mapping particular places (e.g. editorial offices, publishing houses, cabarets, literary cafes, etc.). The literary criticism of cartography method can benefit from the automated text analysis methods. Language tools for recognition and semantic classification of Proper Names can be applied to recognise all their mentions in the text. The mentions of the local names can be combined together into clusters enabling geo-locations with the help of the map-based services like Google Maps. As the geo-information expressed is text is usually incomplete and distributed across sentences, and
PN mentions are often referentially ambiguous, the task is not trivial. In the next step, a language tool for the recognition of the semantic relations can be used to find in text relations linking non-geographical PN mentions to the identified geographical references. The automated text-to-map mapping can be used as tool in the quantitative text analysis, e.g. of the large literary works or literary corpora, as well as a tool supporting literary researchers in the text analysis. All language tools express substantial errors, but the error rate of the Proper Name recognition seem to be enough small for statistical analysis.

A prototype system for mapping Polish literary texts on a geographical map will be presented in the paper. Result of the initial experiments will be also presented.

References

Adapting the Louvain Error Tagset for the annotation of the written component of the PELCRA Learner English Corpus

Agata Pieszczek, University of Warsaw

Learner corpora have become an important tool in SLA research. (Granger 2003a; Tono 2000; Lozano 2013) Independently, or alongside elicitation type experiments, they have been used in hypothesis testing and hypothesis forming investigations. (Díaz-Negrillo & Lozano 2013; Gilquin 2007) Additionally, they are a source of information in the creation of language teaching materials (Nicholls 2003), curriculum design (O’Donnell, 2013) and automated assessment and tutoring systems for language learning. (Andersen, Yannakoudakis, Barker & Parish 2013; Thouësny & Blin 2011)

Learner corpus research draws from the methodology of corpus linguistics and the theoretical framework of SLA. (Lozano & Mendikoetxea 2012) Learner language is investigated as a system in keeping with the Interlanguage Hypothesis. (Selinker 1972) Corpus techniques are used to extract information about systematic features of learner language. The methodology behind the study of learner language is twofold. First, Contrastive Interlanguage Analysis (CIA) can be used to compare learner language with native speaker output from a comparable corpus. Alternatively CIA allows drawing comparisons between two subcorpora of learner language from two different L1 backgrounds. (Granger 2003a; Tono 2000) Secondly, Computer Aided Error Analysis can be applied to quantify and interpret learner errors which had been identified according to a chosen error taxonomy. (Granger 2003a; Granger 2003: 466) Both approaches require annotation in order to make explicit the linguistic features of a text. (Díaz-Negrillo, Meurers, Valera & Wunsch, 2009) Contrastive Interlanguage Analysis generally relies on POS annotation. Error analysis requires a special type of annotation which involves identifying a word or sequence of words that constitute an error and then proposing a reconstructed learner utterance (target hypothesis) in its place. (Reznicek et al., 2013) Finally, the error is categorized according to an error annotation scheme. The different categories of errors comprising the error annotation scheme are chosen based on linguistic categories and a specific theory of language. Error annotation schemes are generally individually designed for specific projects. There is no single accepted system that could be used as a benchmark to date. (Díaz-Negrillo 2006, Díaz-Negrillo, Meurers, Valera, Wunsch 2009) One of the leading error annotation schemes in use today has been developed by the Centre for English Corpus Linguistics at Louvain for the purpose of annotating the International Corpus of Learner English (ICLE). (Dagneaux, Dennes & Granger, 1998) This paper reports on the adaptation of the Louvain Error Tagset for the purpose of annotating a 50 000 word subcorpus of EFL learner texts comprising the written component of the PELCRA Learner Corpus. (Pęzik 2012) The texts selected for the error annotated subcorpus contain information about the learner and task type in the header and are balanced with respect to several variables. The original error tagset,
adapted for the purpose of this study from MacDonald Lightbound (2005) has been modified both in terms of error categories used and also with respect to the architecture of the markup. A hierarchical ordering of error tags was introduced to allow for the coding of several errors on the same span of text. (cf. embedding of errors in Nicholls 2003:575) The occurrence of multiple errors on the same span of text is not meant to reflect competing target hypotheses as in Reznicek. (2013) Rather, the target hypotheses (and corresponding errors) are ordered into a certain progression where only the final target sequence is error free. Alternatively to Lüdeling et al. (2005) where errors on the same span of text are coded as independent layers, here the coinciding error tags have been arranged in a hierarchy and coindexed to link the subordinate and the superordinate error tag.
PoS-grams: an extra tool for the corpus linguist’s kit?
Antonio Pinna, David Brett, Università degli Studi di Sassari

A number of techniques have been developed in the field of corpus linguistics for the identification of phraseologies: while the most widely used being certainly the n-gram, skip-grams and conc-grams have also been experimented with (Cheng et al, 2009). Considerable attention has been paid in recent years to variability within fixed sequences: Biber (2009) and Gray & Biber (2013) have discussed single slot variability in great depth.

One technique that was proposed some time ago (Stubbs, 2007: 91), but has yet to be tested extensively is that of the Part-of-Speech-gram (usually abbreviated to PoS-gram). While n-gram analysis involves quantifying identical strings of n tokens, in PoS-gram analysis the token is substituted by its part-of-speech tag. Hence, the results are composed of strings that are syntactically, but not lexically identical. For instance, the PoS-gram type VBD VVN PRP AT0 AJ0 NN1 could have as tokens such strings as

<table>
<thead>
<tr>
<th>VBD</th>
<th>VVN</th>
<th>PRP</th>
<th>AT0</th>
<th>AJ0</th>
<th>NN1</th>
</tr>
</thead>
<tbody>
<tr>
<td>was</td>
<td>made</td>
<td>by</td>
<td>a</td>
<td>senior</td>
<td>lawyer</td>
</tr>
<tr>
<td>were</td>
<td>entered</td>
<td>during</td>
<td>the</td>
<td>alleged</td>
<td>attacker</td>
</tr>
</tbody>
</table>

While the word forms in each PoS slot in the above example show no similarity whatsoever, when examining greater numbers of tokens, interesting patterns often emerge. Reading the results vertically one may encounter several instances of word forms that are identical, synonymous, or from the same semantic set. When this occurs in several slots, one is left with a number of strings of word forms showing considerable similarity, so much so that they may be considered phraseologies characterised by a certain amount of variability in some slots, and fixedness in others. For example, in the table below, slots 1, 4 and 5 display fixedness, while slots 2, 3 and 6 each feature words that are roughly synonymous (2 and 3), or from the same semantic set (6).

<table>
<thead>
<tr>
<th>VBD</th>
<th>VVN</th>
<th>PRP</th>
<th>AT0</th>
<th>AJ0</th>
<th>NN1</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>great</td>
<td>position</td>
<td>for</td>
<td>a</td>
<td>tour</td>
</tr>
<tr>
<td>a</td>
<td>good</td>
<td>base</td>
<td>for</td>
<td>a</td>
<td>tour</td>
</tr>
<tr>
<td>an</td>
<td>unsurpassable</td>
<td>site</td>
<td>for</td>
<td>a</td>
<td>picnic</td>
</tr>
<tr>
<td>a</td>
<td>great</td>
<td>place</td>
<td>for</td>
<td>a</td>
<td>resort</td>
</tr>
<tr>
<td>a</td>
<td>great</td>
<td>place</td>
<td>for</td>
<td>a</td>
<td>honeymoon</td>
</tr>
</tbody>
</table>

Such examples may be evidence of phraseology ‘templates’, in which certain slots allow choice from a set of related items. Such ‘loose’ phraseologies would fly below the statistical radar of techniques such as that of the n-gram. For instance, in the above case n=3 would yield for a tour, a great
place, great place for and place for a, each with a frequency of two. Nothing would link great place for with great position for, good base for and unsurpassable site for. The power of the PoS-gram lies precisely in the fact that such strings are presented together, given their identical syntactical structure.

This paper will explore the utility of PoS-grams by way of analysis of a 1M token corpus composed of texts from ten subsections of the British newspaper The Guardian. The PoS-grams extracted from the different sections are compared with a database of PoS-grams obtained from the 100M token BNC. Those that are statistically significant are then analysed from a lexical point of view: very often the same slot is found to be occupied by the same word form, a synonym, or one from the same semantic field. When this occurs over several adjacent slots, it may be suggestive of a phraseology typical of the genre.

The results reported will concern phraseologies from the Travel, Crime, Football and Obituaries sections.

References


The concept of a combinatorial dictionary and its multi-stage verification
Adam Przepiórkowski, Institute of Computer Science, Polish Academy of Sciences

The aim of this talk is to present a concept of a combinatorial dictionary describing lexical units of all types – not only single-word lexemes, but also multi-word expressions – sufficiently precisely and formally, so that such descriptions may lead to computationally tractable representations of meanings of whole sentences. The dictionary should contain precise information about lexical units’ morphosyntactic, syntactic, semantic and – to some extent – pragmatic features. Such information should be encoded in formal systems, so that it can be presented in a form readable by humans, but also used by parsers, i.e., syntactico-semantic natural language processing systems. The correctness of the resulting semantic representations of sentences – and, hence, the correctness of the underlying lexical analyses present in the combinatorial dictionary – may be verified with the use of a textual entailment system. While this vision is ambitious, the increasing availability of more basic linguistic resources on which such a combinatorial dictionary may be based – corpora, wordnets, valence dictionaries, etc. – makes this vision increasingly more realistic.
A morphosyntactic approximation of semantic roles in a parser of Polish
Adam Przepiórkowski, Wojciech Jaworski, Institute of Computer Science, Polish Academy of Sciences

There is a strong tradition in Slavic linguistics of relating morphosyntax to semantics, especially, of claiming that morphological cases have unified meanings. One of the most prominent proponents of this approach was Roman Jakobson (see, e.g., Jakobson 1971a,b), and it has been further developed by Anna Wierzbicka (e.g., Wierzbicka 1980, 1981, 1983, 1986), who claims that “cases have meanings and that this meaning can be stated in a precise and illuminating way” (Wierzbicka, 1986, p. 386).

While we do not fully subscribe to this tradition, we show that it turns out to be a useful approach in Natural Language Processing (NLP). In particular, we discuss the role of semantic roles in grammar engineering and argue that – in case of languages with rich morphosyntax but no manually created semantic role resources such as VerbNet or FrameNet – a relatively simple way of inferring an approximation of semantic roles from syntax and morphosyntax may be sufficient for some applications. In fact, it seems that even when a resource like VerbNet is available, this simpler approach to semantic-like roles may be beneficial.

The broad aim of the work partially reported here is to add a semantic component to the manually created LFG (Bresnan, 2001; Dalrymple, 2001) grammar of Polish (Patejuk and Przepiórkowski, 2012), implemented using the XLE platform (Crouch et al., 2011) and currently developed within CLARIN-PL. Regardless of this particular context, we believe that the proposed approach has a wider applicability.

In the mapping from morphosyntax to semantic roles, we capitalise on the fact that Polish has a relatively rich morphosyntactic system, with 7 morphological cases, and a large number of preposition / morphological case combinations, many of which are highly correlated with specific semantic roles. The algorithm for assigning ‘semantic roles’ (scare quotes remind us that these are just approximations of true semantic roles) to arguments is rather simple. With one exception, the ‘semantic role’ is assigned on the basis of the grammatical function of the argument (as well as the voice of the verb). The exception is the OBL(ique) argument – in the LFG grammar of Polish this is prototypically the grammatical function of various prepositional arguments. In this case, also the form of the preposition and the case of its object is taken into account. (The precise algorithm is presented in the full paper.)

We argue that this approach is sufficient for making various inferences immediate which would not be straightforward if arguments were marked only with grammatical functions, e.g., inferences of the b. sentences from the corresponding a. sentences below:

1. (a) Janek pobił Tomka. ‘Janek beat Tomek up.’ –>
We conclude that – given the fact that creating resources such as VerbNet or FrameNet takes a lot of time, money and expertise – the ersatz solution consisting in assigning approximations of semantic roles calculated on the basis of syntactic (grammatical functions) and morphosyntactic (case, preposition form) features of arguments is a viable approach, when the resulting semantic representations are to be used in the task of recognising textual entailment. The algorithm presented in this paper makes it possible to assign such ‘semantic roles’ to arguments almost uniquely, and the resulting neo-Davidsonian representations facilitate textual entailments well beyond what would be possible if arguments were marked with grammatical functions only.

References

A Lexicographic Tool for Recognising Polish Adjective Classes

Radosław Ramocki, Mieszko Służewski, Marek Maziarz, Wroclaw University of Technology

Distinction between relational and qualitative adjectives is a well-known opposition in linguistics (see Trask 1996, Miller 1998, Post 1986). Several tests have been proposed to capture the difference, also for Polish (Nagórko 1987, Szymanek 2010). It is said that qualitative adjectives are gradable (mały - mniejszy - najmniejszy ‘small - smaller - the smallest’), they occupy both attributive and predicative position (mały budynek ‘small building’ - budynek jest mały ‘building is small’) and pre-modify their nouns (czerwony samochód ‘red car’). Qualitative adjectives also form derivatives with suffix -ość (nomina essendi: czerwoność ‘redness’, małość ‘smallness’). Relational adjectives lack such grammatical behaviour.

To support lexicographers who build a wordnet for Polish we have constructed a tool for recognising the adjective classes. It makes use of tests mentioned above. For every given adjective we seek for gradable forms (comparative and superlative grades), predicative function, pre-modifying positions in a large Polish corpus (ca 2*106 tokens). The constraints are written in WCCL language (Radziszewski et al. 2011). Then such specific word occurrences are collected, counted and grouped using a clustering algorithm for WSD (Broda 2011). In the Web application linguists are presented excerpts from the corpus and useful statistics. The examples are chosen in order to represent various grammatical features of a given adjective, as well as, different senses and subsidiary senses of a word. It facilitates wordnet making process.

In the paper we present the tool, its evaluation based on human judgment and some interesting statistics concerning the adjective classes in a large Polish wordnet (Maziarz et al. 2013).

References

In Honour of Jacek Fisiak, v. 2nd: Descriptive, Contrastive and Applied Linguistics, Mounton de Gruyter.


Brazilian food names in English: a corpus-based study
Rozane Rebechi, University of São Paulo

This paper examines how names of typical Brazilian dishes have been rendered in English. A people’s eating habits are not randomly chosen. Rather, they are dependent on culture. Despite its centrality to mankind, cooking was for a long time neglected by scholars (Capatti and Montanari 1999). Nevertheless, the last few decades have witnessed a growing interest in the area by a number of disciplines, including translation. Globalization has contributed to make food items, as well as the words they are associated with, ‘wander’ around the world (Gerhardt 2013). Regarding the translation of Brazilian cooking terminology in English – the lingua franca of international communication –, studies have indicated problems such as mistranslations, lack of standardization, wrong definitions of terms, confusion between Brazilian and Spanish terms, and substitution of ingredients leading to mischaracterization of dishes (Rebechi 2010). Using a methodology based on Corpus Linguistics, we compiled a corpus of 1,696 Brazilian recipes from cookbooks written in English and cookbooks translated into this language. These recipes were annotated with discursive tags (titles in Portuguese and in English, list of ingredients, preparation etc.), which enabled a semiautomatic analysis by the software WordSmith Tools 6.0 (Scott 2012). The concordance lines for the titles revealed a preference for maintaining the recipe name in Portuguese and suggesting a counterpart in English. From a manual analysis of the titles in Portuguese, we were able to roughly group them into: African-origin names (acarajé); Indian-origin names (tapioca); Portuguese-origin names (alfenim); foreign names (musse) and idiomatic names (arroz). Preliminary results demonstrated that rendering Brazilian recipe titles into English is not a standardized activity. Nevertheless, it was possible to identify some tendencies. Indian and African origin names are usually maintained and followed by a definition (e.g. vatapá ‘fish purée with coconut milk’, acarajé ‘black eyed pea fritters’). The analysis also demonstrated a recurring choice for the use of toponyms in rendering recipe names. The use of the adjective ‘Brazilian’ is a common practice in translating recipe names: almôndegas ‘Brazilian meatballs’, café com leite ‘Brazilian coffee with milk’, feijoada completa ‘black beans Brazilian style’ are some examples. Geo-food names (Giani 2009) may provide exoticism to foods and ingredients. Nevertheless, the overuse of this practice may sound redundant in cookbooks dedicated exclusively to Brazilian cooking. Adaptation was also observed. An example of this strategy is the rendering of quindim as ‘coconut cupcake’. Quindim is prepared with egg yolks, grated coconut and sugar, and cannot be described as a ‘cake’. Even more interesting are the strategies used to translate recipe titles which are idiomatic in Portuguese. The origin of many of these names is uncertain but some of them can be related to old practices in Brazil. An example is arroz de carreteiro, a mixture of rice and dried beef. Prepared by lonely men who crossed the country in order to transport goods using a cart pushed by oxen -carreteiro - this dish was simple and not dependent on refrigeration.
tion. Choices such as ‘wagoner’s rice’, ‘mule-driver’s rice’, ‘drayman’s rice’, ‘carter rice’ and ‘cowboy’s rice with jerked beef’, besides not transmitting the source language idiomaticity, would not give the English speaker a clue about the dish. Rather, literal translations or adaptations could lead foreign readers to a compositional interpretation, evoking in them an image which is different from the one evoked in Brazilians. Results of the analysis carried out so far demonstrated a tendency for translation choices which not only point to a mischaracterization of Brazilian cultural markers but may also compromise the communicative process.
Verbs and argument structure constructions in specialized languages

Anna Riccio, University of Naples “L’Orientale”

The special status of the verb in terminological resources other than the noun is an issue widely discussed in literature over the last fifteen years. Verbs, as nouns, tend to have particular usages within situational communication between experts of specific domains.

The aim of this study is to analyze parallel lexicon fragments in order to evaluate the cross-linguistic equivalents among languages on a translational corpus, by employing the semantic frames as an analytic tool for aligning multilingual sentences. The parallel lexicon fragments investigated are ‘subtechnical’ verbs (Baker, 1988) and their argument structure constructions. Medical discourse, for example, is a source of many subtechnical verbs, such as affect, involve, enhance, etc. Unlike terminological verbs (e.g. keratinize or lyophilize), they have a lower level of technicism.

The data are extracted from the OPUS corpus, a specialized multilingual parallel corpus (French, Italian and English). Their description is based on the theoretical model of Frame Semantics (Fillmore, 1977a-b, 1982, 1985; Fillmore and Atkins, 1992) and on the FrameNet methodology (Ruppenhofer et al., 2010), since verbs are “frame-evoking” or “frame-bearing” words par excellence (Pimentel 2012: 5). Each verb evokes a semantic frame representing a sentence-level scenario grouping together verbal forms that share similar syntactic and semantic valence patterns, both within and across languages. For instance, the cross-linguistic correspondence interessare/in-criminer/in-volve in (1) evokes the Damaging frame in medical discourse, since they all mean ‘to have a strong effect on something or someone’, or ‘causing physical damage to something or someone’:

- [...] trattamento laser della [patient zona] interessataTarget, eseguito mediante luce laser rossa. [CauseCNI]
- [...] le traitement de la [patient zone] incriminéeTarget, par une lumière laser rouge. [CauseCNI]
- [...] laser light treatment of the involvedTarget [patient zona], with a red laser light. [CauseCNI]

The resultant analysis shows how taking the English frame as an universal cognitive structuring device (“interlingual representation”, Boas 2005) can be successful to create parallel lexicon fragments among languages (Boas, 2013), which could be useful for translators, the teaching of specialized translation, or technical writers (in an L2).
References


Introducing a corpus of non-native Czech with automatic annotation
Alexandr Rosen, Charles University

A learner corpus can be annotated by standard linguistic categories (morphosyntactic tags and lemmas) and by labels specifying the nature of the errors. For standard corpora including native language, reasonably reliable taggers and lemmatizers are available for many languages. The situation is different for non-standard language of non-native learners. Such texts are often annotated manually, but with their volumes increasing, the need for methods and tools simplifying such tasks is increasing as well.

In this contribution we show that useful results can be achieved by applying methods and a sequence of tools developed for standard language. We focus on the CzeSL-SGTCorpus (Czech as a Second Language with Spelling, Grammar and Tags), which includes transcriptions of essays written by non-native speakers of Czech in 2009–2013, including 8,617 texts by 1,965 different authors with 54 different first languages; altogether 111 thousand sentences, 958 thousand words, 1,148 thousand tokens. Word forms are tagged by word class, morphological categories and base forms (lemmas). Forms detected as incorrect (including some ‘real word’ errors) are corrected by a stochastic spelling and grammar checker and the resulting texts are tagged again. Original and corrected forms are compared and error labels, based on criteria applicable in a formally specifiable way, are assigned. All the annotation is assigned automatically.

Most texts are equipped with metadata about the author and the text. Metadata are available for most texts: 15 items about the author of the text (such as sex, age, first language, CEFR level of proficiency in Czech, duration and method of study), 15 items about the text itself (such as date, time limit, topic, word count). The corpus is available either for on-line searching using the search interface of the Czech National Corpus (http://korpus.cz), or for download as a whole from the LINDAT data repository http://www.lindat.cz).

Automatic annotation in general, and automatic annotation of non-standard language in particular, should be evaluated in order to show its usability potential. Some aspects of the automatic annotation will be evaluated using an existing manually annotated subset of the corpus (the manual annotation includes a target hypothesis about the form and one or more error labels).

Besides error rate, the usability of a corpus annotated in this way should be assessed also by what types of added information are included in the annotation. To address this concern, we show possible uses of the corpus for a few practical tasks in language teaching.

The absence of automatic methods and tools targeting non-native language is not caused only by the computational complexity of the task and the absence of data resources, e.g. for machine learning applications. There is a more fundamental issue of largely missing concepts and schemes to de-
scribe non-standard linguistic phenomena. In conclusion, we introduce a proposal of categories and a tagset for annotating non-standard word forms, which can replace tagging schemes used for standard language.

We are aware that some aspects of manual annotation of non-standard language cannot be substituted by an algorithm or even by a stochastic model. The results still suggest that automatic annotation can be a useful help.
Stanisław Goźdź-Roszkowski, University of Lodz

The significance of evaluation for legal discourse and, especially, judicial discourse can hardly be overestimated. Indicating an attitude towards a legal entity, process or interactant is inherent in the acts of persuasion and argumentation, both being an integral part of judicial discourse. A substantial part of judicial opinions involves expressing agreement or disagreement with decisions given by lower courts, opinions expressed by counsel representing the parties, as well as the opinions arrived at by fellow judges from the same bench. Based on a substantial corpus of US Supreme Court opinions, this paper investigates linguistic strategies employed by judges to evaluate epistemic status (Hunston 2011) of propositions contained in the arguments of other legal interactants. Combining both the quantitative (corpus linguistics) and qualitative (discourse analysis) approaches (Partington 2013), the paper explores the relationship between status and attribution by examining instances of the averred alignments of propositions, which may include their labelling as hypotheses, possibilities or facts. This paper is structured around three major points. First, it argues that judges exercise their subjective judgement when assigning status to propositions contained in the argumentation of other legal interactants (e.g. lower court judges, fellow judges, prosecution) and thus realize the basic function of evaluation. Secondly, the assignment of status is a matter of not only legal but social significance as evidenced by the degree of (social) media attention paid to US Supreme Court rulings in socially important decisions (e.g. same-sex marriages). Finally, status is intrinsically linked with positive or negative evaluations and this seems particularly true for judicial opinions, which are both part of a knowledge-building agenda and they seek to influence human conduct using rationality as a means of persuasion.
Large Polish-English Lexico-Semantic Resource Based on plWordNet - Princeton WordNet Mapping
Ewa Rudnicka, Maciej Piasecki, Wroclaw University of Technology
Wojciech Witkowski, University of Wroclaw

CLARIN is a Language Technology Infrastructure (LTI) which is aimed at supporting research applications in Humanities and Social Sciences (H&SS) and covering all European applications. Many possible applications have cross-cultural and also cross-lingual character. The range of thematic domains and text genres being in focus of H&SS is very broad. Thus, language tools and resources offered by CLARIN must provide large, exhaustive coverage, both in monolingual and multilingual perspective. Lexico-semantic resources providing formalised descriptions of lexical meanings are among the most important ones, especially from the perspective of bridging the meanings of texts in different languages. As the number of different possible language pairs is very large, CLARIN-PL (the Polish part of the CLARIN ERIC scientific consortium) concentrated on the development of Polish-English resources, as a kind of a gateway to the CLARIN system of bilingual resources. The Polish lexico-semantic description is based on plWordNet - the huge wordnet for Polish, i.e. a huge network of lexico-semantic relations describing lexical senses represented by the network nodes. The target size of plWordNet 3.0 in CLARIN-PL is more than 260 000 senses and about 200 000 words (morphological base forms) described. All the senses are to be described by inter-lingual lexico-semantic relations linking them to Princeton WordNet. plWordNet and Princeton WordNet are currently the (two) world’s biggest resources of a wordnet type (see Piasecki et al. 2014). Unlike the majority of ‘national’ wordnets constructed by means of (semi)-automatic, translation-based transfer and merge method (cf. Vossen 2007, Pushpak 2010), plWordNet has been built ‘from scratch’ by a unique corpus-based method (see Piasecki et al. 2009). Thus, it can boast to be the actual reflection of a lexico-semantic system of Polish. Therefore, its linking to Princeton Word constituted a real challenge and required developing advanced mapping procedures and sets of inter-lingual relations (see Rudnicka et al. 2012). During the talk we are going to present the current stage of construction, development process and possible applications of a bilingual, Polish-English wordnet, built in result of mapping plWordNet (Słowsieć) onto Princeton WordNet, and the subsequent expansion of the latter. We discuss three major aspects of the development we made.

The first is the adaptation of tests for interlingual relations that were used to map noun synsets and the introduction of new relations that allow us to capture the derivational characteristics of adjectives in Polish. The second are two sets of algorithms generating automatic prompts for adjective synsets. The first of them used the existing interlingual synonymy relations between noun synsets. The second of them were based on derivational relations holding between units in plWordNet and the interlingual synonymy relations with Princeton WordNet synsets. Next, we move on to the strategy of expanding the resources of Princeton WordNet. The aim of expanding
Princeton WordNet is to reduce the amount of hyponymic links by adding synsets, which can serve as the interlingual synonyms for synsets in plWordNet. The addition of new synsets into the Princeton WordNet is currently based on a strategy, which allows for the expansion of resources with the minimal modification of the already existing networks of intra- and interlingual relations. plWordNet 3.0 and the expanded Princeton WordNet will provide descriptions for lexical meanings from diverse domains. In combination with the Word Sense Disambiguation wordnet-based methods developed for Polish, the resource will be a basis for cross-lingual applications in semantic searching, semantic indexing of texts, text classification, statistical semantic analysis of corpora in different languages, Information Extraction, as well as Machine Translation.

References

A corpus-based comparative study of lexical choices in Middle English dialects
Paulina Rybińska, University of Lodz

The idea of research in the area of Middle English word geography was largely motivated by the fact that it has been perceived by many scholars as a neglected field of historical linguistics (Kaiser 1937; McIntosh 1973, 1978; Benskin and Laing 1981; Hudson 1983; Hoad 1994; Fisiak 2000).

The aim of the study was to first investigate whether the choice of words in the selected Late Middle English texts from different regions is text-dependent or region-dependent, and then to examine if foreign influences have any impact on lexical preferences (Middle English Dictionary, Etymoline). The initial stage of the analysis involved the comparison of the first three chapters of each Mandeville’s Travels version. Therefore, a small corpus comprising c. 7200 words was created and analysed as regards possible discrepancies between the lexicon used. The comparison revealed visible preferences for certain words depending on the origin of the text. Then, other chosen texts were searched though using the Corpus of Middle English Prose and Verse to compare the distribution of the word pairs.

The results of the comparative study indicate that in some cases the choice of the words in the selected texts is certainly region-dependent. The words with higher frequency of occurrence than other analysed equivalents can be identified as northern or southern lexicon. However, foreign influences cannot be decisive as regards words’ popularity in a given region.

To conclude, this small corpus-based study of lexical choices based on the comparison of the two versions of Mandeville’s Travels is hoped to partially contribute both to the field of Middle English word geography and to the study of the aforementioned text in general because, as Hanna (1984: 125) states, “The most basic philological tasks have yet to be performed upon the English Versions [of Mandeville’s Travels].”

References


Report on Teacher Trainees’ use of their own Graded Reader Corpus
James Thomas, Masaryk University

Trainees are MA students in a Faculty of Arts teacher training program which prepares them to become secondary school English language teachers. In the Spring of 2014, their courses in Linguistics for Language Teachers ran in parallel with “Syllabus, Lesson and Material Design for ELT”. These two courses led up to the joint assignment which is described here.

The pre-service teachers (22) are mostly familiar with corpora and concordancers, as their undergraduate program offers several courses that revolve around corpus theory and practice. The in-service teachers (11), on the other hand, needed more induction. This was available through worksheets taken from Discovering English with the Sketch Engine (Thomas forthcoming). The Sketch Engine is the software we use to create corpora and use for linguistic analysis, mainly through concordances and word sketches.

Working in small groups, each chose a graded reader, scanned it and added it to an already existing corpus of graded readers. The graded readers already in this corpus are of the simplified classics type e.g. Jane Eyre (Bronte), The Picture of Dorian Gray (Wilde), whereas those the trainees added were contemporary works commissioned by ELT publishers, such as Jojo’s Story (Moses) Missing in Sydney (Hutchison). The trainees also provided the metadata for author, title, publisher, word count and level.

The trainees used our concordancing software to search their texts for examples of a wide range of language features that they will soon find themselves teaching. They made a subcorpus of books at the same GR Level. Using a “hierarchy of language” framework for linguistic searches, the trainees located examples at the levels of morphology, word usage, functional phrases, word templates, aspects of sentence structure. These include various MWUs including collocation and colligation. The trainees made some findings at the discourse level concerning pronoun usage, hypernyms and troponyms, articles, etc. Using CQL queries they were also able to search for syntagms that instantiate tense/aspect usage in their subcorpora and observe the influence that discourse has on such alternatives. The trainees also observed “topic trails” that appeared either across a whole graded reader or in selected sections. This informs their focus in vocabulary teaching and develops another aspect of working at the discourse level.

The next phase of this work involved presenting their findings in a dedicated Moodle course. Although the trainees have extensive experience of this LMS as students, they had not used it for creating teaching-learning materials. Most of the pre-service teachers are “digital natives” and were able to use the many avenues of Help that are available to teach themselves how to transform their corpus data into draft teaching material. The in-service teachers benefited from a training session using AdobeConnect. In
addition to the specific linguistic features that they derived from corpora, the graded reader corpus was also exploited for literary and cultural purposes. The draft material was submitted as their semester assignments in Moodle, which included activities that use wikis, quizzes and discussion forums.

The next stage of this work sees the trainees developing their work into lessons that they will pilot on students whom they will teach in our so-called Internal Teaching Practice (Autumn 2014). In groups of five, the trainees teach a one semester course, each class having twelve university students who major in any field but English.

Overall, the trainees are involved in identifying a wide range of linguistic, literary and cultural features of a text using a corpus that they contributed to. They conduct a great many searches using sophisticated search techniques. They have developed original teaching material that they will adapt to meet the needs of the students whom they are about to meet.
Evaluative folk metalanguage in the Polish corpus
Jerzy Tomaszczyk, University of Lodz

The subject matter of my contribution will be phrases like the italicized ones in the utterances below,

...and she would have careful fear ... I love that expression, careful fear;

The other thing that’s changed is, for want of a better word, and I hate using it, is the MUSIC INDUSTRY.

..better listening experience and it’s not all physics, some of it is psychoacoustics – it’s not a nice way to say it, but it is psychologically how you respond to a space.

A study of such ostensibly esthetic evaluation of - mostly – the speakers’ own usage in media conversational Polish has shown that most of it is negative, involves routine formulas (78.2% impersonal), and is self-presentational. In the language sample investigated almost all of the evaluation targeted lexical items. To the extent that the negative evaluation can be taken at face value, colloquial expressions turn out to be the most disliked (25.13% in a sample of close to 1000 evaluative utterances), followed by jargon (23.01%) and recent loans, almost exclusively from English, (13.7%). Male speech contains more evaluation than female speech, with journalists being given to such evaluation more than any other group. Genuine positive evaluation (less than 10% of the entire sample) relies on set phrases to a much smaller extent and is much less impersonal (49.43%). The purpose of the work to be presented is to see to what extent the findings summarized above find confirmation in the Polish corpus (nkjp), and what other relevant insight can be obtained from corpus analysis.
Time and space in distance representations: a corpus based study

Jacek Walinśki, University of Lodz

Enigmas of space, time, and their reciprocal relations in the human mind are among most intensely investigated topics in the contemporary cognitive science (Núnez & Cooperrider, 2013). Lakoff and Johnson (1999) assume that the abstract domain of time is conceptualized by metaphorical extension in relation to the physical domain of space (see also Boroditsky, 2011). However, Galton (2011) discusses limits to the spatialization of time. Engberg-Pedersen (1999) argues that the distinction between space and time should be attributed to the basic perceptual difference between static configurations of objects and dynamic developments of events (cf. Langacker, 1991), rather than space and time as such (see also Szwedek, 2009). She adds that at some cognitive levels it is possible to possible to talk both about time-to-space and space-to-time metaphors. Langacker (2012) notes that the relationship between space and time is somewhat circular. Space is more basic as the object of conception, which is indicated by the direction of metaphorical conceptions of time in terms of space. Yet, time serves as the medium of conception for spatial conceptualizations, which makes it more basic as the fundamental prerequisite for cognitive processing of space.

This paper discusses spatial-temporal complementarity of distance representations in the semantic context of motion events (Talmy, 2000). Empirical linguistic data found in the British National Corpus indicate that time is frequently used to express a distance from the speaker’s subjective point of view as a particularly short/long way to a destination, irrespective of the actual spatial separation. It suggest that within the motion-framed perspective (Kövecses, 2005), space and time can be viewed as components of a single conceptual frame, which dictates the relationship between space and time in a complementary fashion (Waliński, 2013).

References

Effective architecture of web services for Polish language processing
Tomasz Walkowiak, Marcin Pol, Wroclaw Univeristy of Technology

The paper presents an architecture of the service oriented system for processing texts in the Polish language. The architecture was developed within a framework of CLARIN-PL project. The development of the architecture follows the CLARIN project motto: "Make tools and resources available as webservises". The aim was to develop an open access, scalable and highly available web system with the various types of interfaces that allows to integrate and run different language resources and tools (LRT).

To meet these requirements, the natural language processing system is based on a scalable hardware and software architecture that can be easily optimized to deliver high performance. That’s why Cisco UCS B-Series Blade Servers based on Intel® Xeon® processors E5 product families, work with virtualized applications to increase performance, energy efficiency and flexibility. Every server is designed with 32 dual in-line memory module slots. Servers are connected by fast fiber channel connection with highly scalable midrange virtual storage designed to consolidate workloads into a single system for simplicity of management, the IBM Storwize V7000. Virtualization makes the virtual infrastructure management more convenient and efficient. Since operating systems are independent from the hardware they can be easily moved to another server as a reaction to any failure or resource shortage. Moreover, virtualization provides a disaster recovery mechanism ensuring that when virtualized system crash, it will be restored as quick as possible.

The core of the system consists of the LRT execution engine. It consists of the simple asynchronous REST service, task queues, data storage and a set of LRT workers. The data to be processed has to be stored in the data storage. Next, the REST service is called that creates the task to be processed (consisting of LRT name and input file name) and stores it in the queue. A client of the REST service can check if the task was finished and get the id of the results file. The data processing is done by LRTs working as workers. Each worker collects a task from the queue, loads data from the data store, processes them, returns the result to the data store and informs the execution engine that the task was finished. The workers and the queue system allows an effective scalability of the system. In case of a need a larger number of workers could be run. Due to a usage of virtualization it is achieved by running a new copy of virtual machine that hosts a given type of workers. The results of the system speed-up for a simple chain of LRTs is presented in the paper. The asynchronous mechanism prevents the clients from crashing due to time-outs caused by long time processing (for example due to system overload). The separation of loading/downloading from the task execution allows to call a sequence of tools without a need of loading/downloading large amount of data from and to a client. The paper present the comparison of processing time for the proposed architecture and a system without...
separate loading/downloading. The system core, for the security reasons, is working in a private network. The access from the Internet is granted by additional Web Service server. It works as a proxy for the core system delivering a large set of different APIs. They use different techniques for accessing the core services including synchronous as well as asynchronous services, SOAP and REST, as well as XML and JSON ones. Such approach allows an easy integration with any kind of applications (starting with mobile ones, through desktops, servers and JavaScript ones) from nearly any programming language.

Additionally for a usage in real time applications (like for example voice interface to a mobile system) the synchronous version of the core system is developed. The access to it is limited for short text and a given intensity of requests from a given IP. Moreover, a web application is planned that will allow to build tool chain using drag and drop technique. The application will be developed using HTML5 and JavaScript technology. Also, the technique for integration of external services is available, allowing to contribute external web services.
Corpus-based approach to finding targets of evaluations in Polish

Aleksander Wawer, Institute of Computer Science, Polish Academy of Sciences

This work is focused on one aspect of evaluative language, namely the detection of what is being evaluated. The specific problems under consideration are the analysis of what are the lexical and syntactic determinants of constituting opinion targets, creation of a manually-labeled corpus and an attempt at automated opinion target recognition.

In the frequent scenario of sentiment analysis, evaluations are transmitted by adjectives that modify nouns – targets of opinions, forming noun phrases. The problem of establishing opinion targets is alleviated, as the nouns are assumed to be targeted by evaluations. However, in some cases the noun phrase framework and adjective-noun pattern in general, is not sufficient. For example [Qiu et al., 2009] describe more complex dependency-based rules to capture relations between evaluative words and their targets. Our work does not impose part-of-speech constraints on opinion targets and evaluations, and goes beyond the rules proposed by [Qiu et al., 2009].

All the experiments in this paper were conducted on a corpus of reviews of two types of products, downloaded from one of the major Polish online opinion aggregation websites. For the detection of occurrences of evaluative language, we used the Polish sentiment lexicon [Wawer & Rogozińska 2012]. In order to identify opinion targets, we used semi-automatically compiled word lists described in [Wawer & Gołuchowski 2012].

Among all reviews, we identified 33014 sentences with the occurrences of both evaluative language and opinion targets. Using the Polish dependency parser [Wróblewska 2012] we performed dependency analysis of those sentences.

For further analysis we selected sentences with exactly one evaluative word (E) and one opinion target (T). We modified MaltEval software for dependency tree visualisation and manually annotated the dependency trees for the following information: (1) correctness of the dependency structure between E and T, (2) T being a correct opinion target, (3) E being a correct evaluative word, (4) the type of relationship between E and T (whether E is an evaluation of T or not). In the article we discuss a number of borderline and difficult cases.

In the inter-annotator agreement study, agreement between two linguists ranged from 91

In the last part of the article we describe preliminary results of machine learning recognition of the relationship between E and T. For each node in the dependency path between E and T, we generate a feature vector containing: its lemma, part-of-speech, and type of dependency relation. In order to classify the relation between E and T (we focus on task (4) - whether E is an
evaluation of T or not) we employ logistic regression and SVM algorithms and leave-one-out method. The results are promising: the accuracy reaches 0.88 for the logistic regression and 0.87 for the SVM.

The article ends with a discussion of future work on automated recognition of the links between E and T and findings formulated about the corpus. For instance, the annotation demonstrates the existence of sentences where relations between evaluative words (E) and opinion targets (T) go beyond noun phrase.
Effective processing of ever-growing amounts of social media textual data such as tweets, Facebook posts or blogs poses a great challenge to natural language processing. The problems arise due to vast amounts of available texts of multilingual, colloquial and noisy nature. Moreover, the data availability model is far different from the typical scenario of fixed, designed corpora – it is based on a constant stream of incoming messages, related to a changing context of extra-textual events. All these difficulties result in partial or even poor adequacy of existing natural language processing tools, especially in terms of accuracy and scalability.

To study problems of such difficult type of data we have chosen to analyse messages published via Twitter, a worldwide popular social messaging service. In Twitter, messages are limited to 140 characters and therefore are concise, yet published frequently and often seconds after an event, which is commented in them, happened. This type of data was rarely studied for Polish, with some notable exceptions (Piskorski & Ehrmann 2013), but is extremely important because of practical applications, in which constant monitoring of social communications plays a crucial role for many organizations.

To narrow the scope further, we decided to focus only on political communication on Twitter. Starting point of the analysis was to process the messages using state-of-the-art NLP tools for Polish, such as tagger Polita, shallow parser Spejd, rule-based sentiment analysis tool Sentipejd, named entity recognizer Nerf and Mention Detector. These tools were developed for more formal language and reveal multiple shortcomings when applied to Twitter data.

In our research, we present most common problems which occur when processing non-formal Twitter communication. At the most broad level, the problem comes down to fair amounts of noise, prevalent in gathered tweets. Specific issues include ungrammatical language with many misspellings, omitting Polish diacritical marks, usage of abbreviated word forms, mixing words with urls and other types of non-linguistic data such as hashtags or emoticons. We quantify these phenomena using a sample of hand-labeled tweets, then highlight some solutions to overcome these problems.

Another focus was put on establishing an architecture capable of monitoring Twitter for new messages (even few hundred users may post about a 1000 tweets/hour), constant NLP processing of new data and visualization of the analysis.

We describe some details of the system and highlight its design principles.
The infrastructure is integrated with Twitter API to obtain the data. The process of data collection is integrated with an ontology of Polish political scene, containing basic facts about its participants. Due to large amounts of data, the texts are then stored in a MongoDB database for further analysis. The system contains components for efficient frequency analysis of gathered tweets based on map-reduce paradigm. It includes also trend mining component, based on using unsupervised clustering algorithms and two sentiment analysis engines. The system provides visualization interface, integrated with both the sentiment and trend mining module.

The results of our work will be evaluated by making available the tools for monitoring Twitter streams of social data in Polish. A solution like this is highly anticipated by politicians, journalists and analysts of political life.
Morfeusz — the new prospects
Marcin Woliński, Michal Lenart, Witold Kieraś, Małgorzata Marciniak, Piotr Rychlik, Institute of Computer Science, Polish Academy of Sciences

Morfeusz, a morphological analyser for Polish, is well known in the linguistic environment for many years. It was used as a basic tool for morphological processing of Polish in many projects and language tools. Let us mention the most important practical applications:

- annotation of corpora: The IPI PAN Corpus of Polish (Przepiórkowski, 2004) and National Corpus of Polish, NKJP (Przepiórkowski et al., 2011);
- taggers: Dębowski (2004), TaKIPI (Piasecki, 2007), PANTERA (Acedański, 2010), WMBT (Radziszewski and Śniatowski, 2011), Concraft-pl (Waszczuk, 2012);
- parsing tools: Spejd (Przepiórkowski, 2008) and Światra (Woliński, 2004), SproUT (Piskorski et al., 2004);
- the multiword expression toolkit Multiflex (Savary, 2005);
- many other smaller projects, e.g., annotation of domain data used for terminology extraction.

These applications allowed us to collect a lot of experience on advantages and disadvantages of the previous version (Woliński, 2006) of Morfeusz and to identify drawbacks of the tool. Technical details of the current approach are given in Woliński (2014). In our presentation we concentrate on the new features of the tool from a user point of view. We describe the following improvements of the current Morfeusz:

- It includes not only an analyser but also a compatible generator. It is possible to generate all forms of a lemma or generate a desired form by giving its detailed description.
- It provides an infrastructure for including domain dictionaries or replacing the basic one completely. Two inflectional dictionaries are available with Morfeusz: data coming from the Grammatical dictionary of Polish (Saloni et al., 2007) and Polimorf (Woliński et al., 2012) data.
- It provides possibility to disallow movable inflections. So zrobiłeś ‘did you’ can be analysed standardly as two segments: zrobił ‘did’ and -eś ‘second person marker’, but it is possible to turn off this feature and analyse it as one segment.
- It adds two new pieces of information to its interpretations: classification of proper names and stylistic labels. So the labels like: ‘archaism’, ‘colloquialism’ or ‘medical’ (signalling domain terminology) are now available.
• A set of prefixes is established with rules for creating compound words. So the new Morfeusz analyses words like kilkunastostronicowe ‘several pages long’, wielkoformatowy ‘big format’, parapubliczne ‘para public’ or alterglobalista ‘alterglobalist’.

• A new GUI is implemented.

References


Verbalization patterns in learners' constructed responses to German listening comprehension items

Magdalena Wolska, Eberhard Karls Universität Tübingen

Constructed response is a type of test item format which elicits free text production, rather than specifying a closed set of answers from which a test-taker selects (as in selected-response items, such as multiple-choice or true-false items). Short constructed response items - so-called "short answers" - are frequently used to assess comprehension. The textual inputs may range from a single word or phrase through a few sentences.

Unlike selected-response items, in real-life assessments short responses need to be scored by human readers, due to their semantically-oriented and open-ended nature. Experimental automated short answer scoring systems exploit natural language processing to assign a grade. This is typically done based on shallow linguistic features and matching, alignment, and/or classification techniques. Scoring accuracy in this setup may be affected, among others, by the range of verbalizations and, in case of non-native production, also by form errors, such as misspellings and ungrammaticality.

In this work we analyze patterns of verbalizations in constructed responses to a listening comprehension test used in German as a Foreign Language course placement. We use data from over 350 learners (over 20 L1s) who participated in an actual placement test administered at a German university. Aside from a three-part listening comprehension section (24 questions of increasing difficulty) the placement test included a C-Test, a general language proficiency test in which half of every second word in short reading passages needs to be filled in by the test-taker. Learner responses to the listening comprehension items have been scored by teachers of German involved in teaching the courses. In this setting we ask the following questions: Are there patterns to verbalization diversity typical of correct, partially correct, and incorrect responses? Are there patterns to verbalizations typical of proficiency levels (as measured by C-Test) or/and L1s? Our larger goal is to formulate a language processing strategy, based on shallow linguistic features, for (semi-)automated scoring of non-native short answers to comprehension items in foreign language testing.

In the present work, we discuss initial findings based on a corpus of learner responses to one placement test, described above. The data have been automatically preprocessed to remove textual features irrelevant according to the scoring rubric: capitalization and punctuation have been normalized. We show that verbalization diversity is item-specific and related to item difficulty. We discuss diversity in patterns of misspellings, relate misspellings to L1-families for the largest L1 subsets (Romance, Germanic, and Slavic), and point out interesting tendencies. We formulate conclusions on prospects for consistent scoring accuracy across items.
Towards an Events Annotated Corpus of Polish
Michał Marcińczuk, Marcin Oleksy, Jan Kocoń, Tomasz Bernaś, Michał Wolski

Event recognition is one of the major information extraction tasks. It is used in many practical applications like text summarization, discourse analysis, events reporting and facts inference. Within the Clarin-PL project we plan to develop methods and tools to recognize event in Polish texts along with their description including event category, participants, time and location in order to enable deeper text understanding. In order to create and evaluate such tools we need a corpora annotated with events. The most popular specification of event annotation is the TimeML Annotation Guidelines Version 1.2.1 (Saurí et al., 2006) (henceforth, TimeML), which has been already used and adjusted to several languages, including Spanish (Saurí et al., 2010), Catalan (Saurí and Pustejovsky, 2010), French (Bittar, 2010) and Italian (Caselli et al., 2011). Applying an existing guidelines for another language requires a careful study of that language phenomena and might need some adjustments concerning language-specific issues. In this paper we present the results of such a study for Polish. The study was performed using the KPWr corpora (Broda et al., 2010). At this point we have focused on guidelines for annotation of event mentions only. In the future work we will also include annotation of event participants, time and location, and event linking. We have used 8 coarse-grained categories of events: action, state, reporting, perception, aspectual, i_action, i_state and light_predicate. Most of the categories have the same definition as in TimeML, but there are several differences. The two core categories are action and state — they represent the key opposition for semantic classification for situations (Vendler, 1967). The remaining categories might be treated as auxiliary categories, which introduce some additional information about the actions and states. Comparing to TimeML, the occurrence category was replaced with action. We decided to do so, because we found out that annotation of occurrences causes some problem because it requires the distinction between action which happen or occur in the world and action which refer to some general rules. We decided to annotate both types of actions as action. The distinction between them will be done as a separate step. Reporting, perception and aspectual have an argument that is an action or a state and they should indicate that the action which is the argument actually happened. I_action and i_state also have an argument of an action or a state respectively, but they do not indicate that the action or state was true. They place the events in an alternative worlds. Light_predicate is an grammatical category which refers to mentions which occur with another mention but does not introduce any additional semantic information about the event. They are synsemantic verbs that generally accompany the events expressed by a nominalization. We annotate events expressed by means of tensed or untensed verbs (including participles and gerundial forms), nominalizations (in the wider lexicalistic sense) and adjectives. In case of modal or phase-aspectual predicates and verbo-nominal analytisms (Jędrzejko, 2011) we tag both elements...
because they are relevant to different kind of event information — after that two tagged events will be linked as identical. In order to measure the inter-annotator agreement we have randomly selected 100 documents from the KPWr corpus. The documents were independently annotated by two linguists. We used the positive specific agreement (Hripcsak, 2005) as there are no negative decisions to count. The agreement on mentions annotations without categorisation is ca. 85% and with categories it drops to 68%. The agreement for each category separately varies from 36% for states to 86% for aspectual.

References

Force Dynamics and Time
Jacek Wozny, Wroclaw University

The concepts of motion and force are both extensively discussed in cognitive linguistics literature. But they are discussed separately. The first usually in the context of ‘motion situations’ (for example: Talmy 2000, Slobin 2003, Zlatev 2007, 2010, 2012), the other as part of the Force Dynamics framework, which was developed by by Talmy (1976, 1988, 2000) and adopted by, for example, Sweetser (1982, 1991), Johnson (1987), Boye (2001), Vandenberghe (2002) and Da Silva (2003). The aim of this paper is twofold: first, to argue that the modified Force Dynamics framework should include the parameter of time and second- to prove that the concepts of force and motion should not be isolated but considered as two inseparable parts of force-motion events (cf. Woźny 2013). To this end, a random corpus sample of 50 metaphors containing the verb ‘went’ is analyzed with respect to the linguistically coded parameters of force motion and time.
Poliqarp 2: a corpus lookup system for multiple annotation layers
Bartosz Zaborowski, Dominika Pawlik, Aleksander Zabłocki, Institute of Computer Science, Polish Academy of Sciences

One of the tasks of CLARIN-PL project is to build an efficient linguistic search engine which would be able to handle multiple annotation layers. The motivation for this task is the lack of a search engine which would support rich annotations spanning across multiple levels and would be efficient enough to perform search in corpora consisting of a billion tokens.

The available morphosyntactic search tools (e.g. CWB forks such as (No)Sketch, Poliqarp) operate only on a flat annotated text with no structure. On the other hand, existing treebank search engines (like TigerSearch, tgrep, etc.) do not support any complex annotations such as ambiguous morphosyntactic tags of words or segments. They also do not support deviations from tree structure which may occur in constituency graphs, e.g. sharing syntactic children between different syntactic groups.

The general concept of the Poliqarp 2 project is to borrow many ideas from the abovementioned tools and merge them in an efficient, powerful implementation. The implementation is planned to be suitable for various corpora based on significantly different data models, including the National Corpus of Polish, constituency and dependency treebanks (e.g. Składnica), LFG treebanks (e.g. POLFIE) and even the currently developed corpus of 17th century Polish language. The project involves also designing a possibly simple query language which should naturally generalize the old Poliqarp query language and at the same time provide the user with much more expressive power.

The available query constructs in the new version include in particular:

- simple term queries (with the terms specified literally or by regular expression)
- attribute-value queries (with special Poliqarp ambiguity operators available)
- Boolean combinations of the above
- accessing nested attribute structures (e.g. in LFG)
- horizontal regular expressions over segment/node specifications
- vertical relations (A is a child/descendant of B, with generalizations)
- unified access to morphosyntactic annotation, metadata and other annotation layers

and some general methods of expressing conditions of other kinds, e.g.

- all children of A are nominal groups
- A is a sentence with between 2 and 4 finitive verbal groups
• A is a right sister of B (even if it is not adjacent to B in the input text)
• A and B have the same (possible) values of lemma etc.

To the best of our knowledge, Poliqarp 2 is the first system with all the above features. In addition, we are developing a built-in SQL-like post-processing engine, which will allow e.g. creating configurable collocations lists.

As we have been working on this task for a while, we would like to present the current state of our work. For those who are not familiar with the Poliqarp 2 prototype we will talk a bit about the work done so far: general concepts, corpus data modeling and query language. We will also discuss some aspects of the project which have not yet been presented publicly. We will present the current (almost final) state of the query language. In particular, we will unveil some details about postprocessing abilities of queries.

We would also like to present some details concerning the prototype implementation, focusing on improvements regarding speed and the actual searching power of the implementation. Additionally, we would like to discuss some technical challenges and problems found during the work on this implementation, including some open questions. We will end the talk with some future plans/road-map of the work on Poliqarp 2.
Surviving across centuries: investigating the use of the enclitic -t' in the 19th century Czech
Anna Zitová, Charles University

The paper is concerned with a phenomenon which is, as far as we know, not well understood in Czech or world linguistics yet: a non-autonomous archaic enclitic element -t. A corpus-based inquiry is undertaken to describe the distribution of -t' in the texts of the period studied. In this light, an attempt is made to explain the functions of this enclitic in the 19th century. Restrictions of the corpora of historical Czech (Diakorp, comprising texts throughout the history, and DIA, corpus of the 19th century texts in preparation, both included in the Czech National Corpus) are discussed.

The 19th century represents a specific period in the development of Czech: the national movement in the first decades of the century was oriented philologically, i.e., on the restoration of the representative variety of vernacular language after the preceding 150-year period in which the communicative importance of Czech language in the public sphere was diminished. Grammarians and authors publishing in the 19th century took Old and especially Early Middle Czech as a model of proper and classy language usage (cf. HAVRÁNEK 1936, ADAM 2006) and, therefore, the language development was intentionally upset, influenced and formed.

The enclitic -t' (also in the form -tě, less often -ti) occurred in the oldest Czech monuments (e.g., Alexandreis) already and, as a particle with specific functions, it stayed in use until the beginning of the 20th century (cf. dictionary Příruční slovník jazyka českého). Old grammarians describe its function simply as to emphasize the word it is connected with (the same description is given also for another archaic enclitic particles: -(i)ž, -to; cf. DOBROVSKÝ 1809, NEJEDLÝ 1821), Zikmund speaks of marking words on which extra stress is put (ZIKMUND 1863). It remains unclear whether this stress brings out some pragmatic meanings – unlike the enclitic -ž, distribution of which is substantially limited to imperatives (and conjunctions or connective means in general but in these cases we assume a high degree of lexicalization, cf. STLUKA 2006), the enclitic -t' can be connected with various parts of speech in various inflectional forms (DOBROVSKÝ 1809, STLUKA 2006), so there are no obvious clues about what can be intensified or emphasized.

Modern linguists recognized another function of this particle: it regularly occupies the first or the second position in a clause and marks the explicative relationship as a discourse connective (TRÁVNÍČEK 1935, BAUER 1960). Other more possible functions are associated with the unclear origin of the enclitic: in some cases it could be a shortened form of a so-called ethical dative (ti, you.DAT+SG') and express pragmatic meanings in this way (cf. GEBAUER 1896). Eventually, the use of the enclitic as an expletive (dummy) pronoun could reflect its potential original relationship to demonstrative pronouns with the consonant t- (Ja’t’ jsem., It’s me.).
According to Stluka (2006), enclitic -t’ was presumably a frequent phenomenon in spoken Czech and it could occur multiple times in a sentence (cf. kněhyt’ jsú dobré a stojie’ za ty peniezě ,[the books],ENC are of good quality and [it is worth].ENC spending money on them’). How frequent could it be in the 19th century? To what extent did authors imitate the old language? Will there some restrictions in the distribution to various lexical items show? Which functions will be identified? The data from the diachronic corpus Diakorp as a background and the statistical analysis of the 19th century corpus help us to answer.