Proceedings

Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)

Verena Henrich & Erhard Hinrichs (eds.)

organized as part of the
ESSLLI European Summer School in Logic, Language and Information
Tübingen, Germany, August 2014

ESSLLI 2014
sfs Seminar für Sprachwissenschaft
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

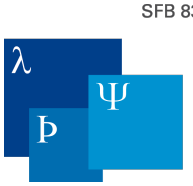
Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)

Workshop organized as part of the
ESSLLI European Summer School in Logic, Language and Information
August 11-15, 2014 (ESSLLI first week), Tübingen, Germany

Proceedings

Verena Henrich & Erhard Hinrichs (eds.)

Tübingen, Germany, August 2014



CCLCC website: http://www.sfs.uni-tuebingen.de/~vhenrich/cclcc_2014/

Publisher:

Department of Linguistics (SfS)
University of Tübingen
Wilhelmstr. 19
72074 Tübingen, Germany

and

Collaborative Research Center: Emergence of Meaning (SFB 833)
University of Tübingen
Nauklerstr. 35
72074 Tübingen, Germany

Contact:

acl-sekretariat@sfs.uni-tuebingen.de

http://www.sfs.uni-tuebingen.de/

This volume has been compiled from the pdf files supplied by the authors.

Table of Contents

List of Reviewers

- Réka Benczes (Eötvös Loránd University)
- Fabienne Cap (University of Stuttgart)
- Walter Daelemans (University of Antwerp)
- Corina Dima (University of Tübingen)
- Ulrich Heid (University of Hildesheim)
- Verena Henrich (University of Tübingen; co-chair)
- Erhard Hinrichs (University of Tübingen; co-chair)
- Christina Hoppermann (University of Tübingen)
- Jianqiang Ma (University of Tübingen)
- Preslav Nakov (Qatar Computing Research Institute)
- Diarmuid Ó Séaghdha (University of Cambridge)
- Stan Szpakowicz (University of Ottawa)

Workshop Program

Monday, August 11, 2014		
17:00 – 17:30	Erhard Hinrichs and Verena Henrich (University of Tübingen)	Introduction to the Workshop
17:30 – 18:30	Melanie Bell (Anglia Ruskin University)	Invited Talk: Compound stress, informativity and semantic transparency

Tuesday, August 12, 2014		
17:00 – 18:00	Eduard Hovy (Carnegie Mellon University)	Invited Talk: The Semantics of Word Collocations from a Distributional Point of View
18:00 – 18:25	Daniil Sorokin, Corina Dima, and Erhard Hinrichs (University of Tübingen)	Multi-label Classification of Semantic Relations in German Nominal Compounds
18:25 – 18:30	All	Discussion

Wednesday, August 13, 2014 – Computational Methods		
17:00 – 17:25	Pedro Bispo Santos (Technische Universität Darmstadt)	Using compound lists for German decompounding in a back-off scenario
17:25 – 17:50	Eleri Aedmaa (University of Tartu)	Statistical methods for Estonian particle verb extraction from text corpus
17:50 – 18:15	Daria Kormacheva, Lidia Pivovarova, and Mikhail Kopotev (University of Helsinki)	Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams
18:15 – 18:30	All	Discussion

Thursday, August 14, 2014 – Collocations		
17:00 – 17:25	Lothar Lemnitzer and Alexander Geyken (Berlin Brandenburgische Akademie der Wissenschaften)	Semantic modeling of collocations for lexicographic purposes
17:25 – 17:50	Zygmunt Vetulani and Grażyna Vetulani (Adam Mickiewicz University)	Verb-Noun Collocations in PolNet 2.0
17:50 – 18:15	Laura Giacomini (University of Heidelberg)	Variational models in collocation: taxonomic relations and collocates inheritance
18:15 – 18:30	All	Discussion

Friday, August 15, 2014 – Multi-Word Expressions		
17:00 – 17:25	Konrad Szczesniak (University of Silesia)	Too Colorful To Be Real. The meanings of multi word patterns
17:25 – 17:50	Francesca Quattri (The Hong Kong Polytechnic University)	Cross-language description of shape: Shape-related properties and Artifacts as retrieved from conventional and novel collocations across different languages
17:50 – 18:15	Petya Osenova and Kiril Simov (Bulgarian Academy of Sciences)	Treatment of Multiword Expressions and Compounds in Bulgarian
18:15 – 18:30	Erhard Hinrichs and Verena Henrich (University of Tübingen)	Final Discussion and Closing of the Workshop

Acknowledgments

We would like to thank the program committee of ESSLLI 2014 for selecting our workshop proposal as part of the ESSLLI 2014 program and the local organizing committee of ESSLLI 2014 for their kind assistance with all practical matters. Our special thanks go to all the authors, invited speakers, and reviewers who agreed to contribute their expertise to this workshop.

Support for this workshop was provided as part of the DFG grant to the Collaborative Research Center *Emergence of Meaning* (SFB 833). We are grateful to Cool Press Ltd for making the EasyChair conference system available to us for paper submission and paper reviewing.

Preface

The workshop on *Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations* (CCLCC 2014) was held at the Eberhard Karls University Tübingen, Germany, on August 11-15, 2014, as part of the 26th European Summer School in Logic, Language and Information (ESSLLI 2014).

Workshop Topic and Goals

The analysis of complex words, compounds, and collocations has received considerable attention in linguistics, cognitive science and computational linguistics. Research on these phenomena concerns theoretical, experimental, and applied aspects relevant to all three disciplines. This diverse and interdisciplinary perspective lends itself particularly well to an ESSLLI workshop on this topic.

The aim of the workshop is to stimulate a cross-disciplinary discussion that will be of mutual benefit to the three fields sketched above and that will provide a forum for junior and senior researchers alike.

Word formation processes such as cliticisation, compounding, and noun incorporation are highly significant for linguistic theory (since they concern the interface of morphology and syntax) and for linguistic typology (since languages differ considerably in the division of labour between morphology and syntax). The automatic analysis of complex words has also played an important role in computational linguistics. Here, the main tasks concern the parsing problem of assigning the correct bracketing of complex words and the semantic interpretation problem of automatically assigning the range of lexical-semantic relations among the constituent parts of a complex word. The automatic treatment of complex words and linguistic units "just above" the word level is also a hot topic from both an applied and a theoretical perspective in computational linguistics. N-gram models have played a major role in statistical approaches to a wide variety of natural language processing applications including machine translation, information retrieval, and text summarization. For computational semantics, complex words and collocations are a particularly interesting test bed for extending distributional approaches to word meaning (using vector space models) beyond the level of individual words and for investigating a synthesis between distributional models and model-theoretic approaches to compositional semantics.

From the perspective of cognitive psychology, the interpretation of novel compounds is an interesting domain of inquiry into human sentence processing since such compounds require access to the meaning of individual words as concepts in the mental lexicon as well as the selection of semantic relations that link these concepts.

*** *** ***

In this proceedings volume, we present the contribution by 14 authors from 7 countries and the abstracts of two invited keynote lectures delivered by Dr. Melanie Bell (Anglia Ruskin University, Cambridge, United Kingdom) and Prof. Dr. Eduard Hovy (Carnegie Mellon University, Pittsburgh, PA, U.S.A.).

Verena Henrich & Erhard Hinrichs
Editors

INVITED TALKS

Invited Talk:
Compound stress, informativity and semantic transparency

Melanie Bell, Anglia Ruskin University

In Present-day English, noun-noun compounds are attested both with stress on the first constituent (N1) and with stress on the second constituent (N2). The most reliable predictors of stress position are the identities of the two constituents (e.g. Bell & Plag 2013): particular nouns in first or second position are associated with particular stress patterns. Furthermore, the stress pattern associated with a constituent is largely determined by its informativity in that position (Bell & Plag 2012). For example, when N1 occurs rarely as a modifier, perhaps only in a single compound, the probability of N2 given N1 is very high; N2 therefore adds little information and is unlikely to be stressed. Another strong predictor of stress pattern is the semantic relation between the two nouns (e.g. Plag et al. 2008): for example, when N1 represents the material or location of N2, the compound is likely to be right-stressed. This raises two questions: firstly, what, if anything, do the relations associated with right stress have in common? Secondly, what is the relationship, if any, between constituent informativity and semantic relation? One hypothesis is that certain semantic relations are more transparent than others, and that right stress, being the phrasal pattern, is also associated with transparent, phrase-like semantics.

Bell & Schäfer (2013) modelled the transparency of both compound nouns and individual compound constituents and found that, while certain semantic relations are indeed associated with greater transparency, these are not only those associated with right stress. Furthermore, the best predictors of perceived compound transparency are the perceived transparency ratings of the individual constituents. Work on conceptual combination by Gagné and collaborators has also shown that relational information in compounds is accessed via the concepts associated with individual modifiers and heads, rather than independently of them (e.g. Spalding et al. 2010 for an overview). This leads to the hypothesis that it is not the semantic relation per se that makes a compound more or less transparent; rather, it is the degree of expectedness of the relation given the constituents. In this talk, I provide evidence in support of this hypothesis: the more expected the relation for a constituent, the more transparent that constituent is perceived to be.

Bell & Schäfer (in preparation) used the publicly available dataset described in Reddy et al (2011), which gives human transparency ratings for a set of compounds and their constituents. For each compound constituent in the Reddy et al. data, we extracted from the British National Corpus the family of compounds sharing that constituent. This larger set was then coded both for semantic relation (after Levi 1978) and for the particular sense of N1, using WordNet (Princeton 2010). This enabled us to calculate the proportion of compound types in each constituent family with each semantic relation and each WordNet sense. These variables were then used, along with other quantitative measures, as predictors in an ordinary least squares regression model of the transparency of N1. The model provides clear evidence for our hypothesis: N1 is perceived as most transparent when it occurs with its preferred semantic relation. Furthermore, transparency also increases with other measures of expectedness, namely when N1 is a frequent word, with a large positional family, occurring with its most

frequent sense, and with few other senses to compete. In so far as perceived transparency is a reflection of expectedness, it can therefore also be seen as the inverse of informativity.

These results suggest that, rather than reflecting overall compound transparency, right stress might be associated with transparency of N1, and that transparent modifiers tend to be associated with particular relations. Future work will test this hypothesis further, as well as investigating the relationship between semantic relation and N2.

References

Bell, Melanie J. & Ingo Plag. 2013. Informativity and analogy in English compound stress. Word Structure 6(2). 129-155.

Bell, Melanie J. & Martin Schäfer. 2013. Semantic transparency: challenges for distributional semantics. In Aurelie Herbelot, Roberto Zamparelli & Gemma Boleda (eds.), Proceedings of the IWCS 2013 workshop: Towards a formal distributional semantics, 1–10. Potsdam: Association for Computational Linguistics.

Bell, Melanie J. & Ingo Plag. 2012. Informativeness is a determinant of compound stress in English. Journal of Linguistics 48(3). 485-520.

Levi, Judith N. (1978). The syntax and semantics of complex nominals. New York: Academic Press.

Plag, Ingo, Gero Kunter, Sabine Lappe & Maria Braun. 2008. The role of semantics, argument structure, and lexicalization in compound stress assignment in English. Language 84(4). 760-794.

Princeton University. 2010. About WordNet. http://wordnet.princeton.edu

Reddy, Siva, Diana McCarthy & Suresh Manandhar. An empirical study on compositionality in compound nouns. In Proceedings of the 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011), Chiang Mai, Thailand.

Spalding, Thomas L., Christina L. Gagné, Allison C. Mullaly & Hongbo Ji. 2010. Relation-based interpretation of noun-noun phrases: A new theoretical approach. Linguistische Berichte Sonderheft 17, 283-315.

Invited Talk:
The Semantics of Word Collocations from a Distributional Point of View

Eduard Hovy, Carnegie Mellon University

The theme of the workshop has become increasingly popular in Computational Linguistics over the past years, as work focusing on Distributional Semantics is taking off. This complements previous research in lexical computational semantics, which draws from many variants within pure Linguistics, from Halliday and Firth to Mel'cuk and others.

The foundational assumption of Distributional Semantics is often attributed to Firth: "you shall know a word by the company it keeps" (1957:11). The idea is that the collocational context of a target word, represented as a frequency distribution of other words, characterizes the meaning of the target: "bank1" = {money 120, deposit 105, teller 98, ATM 94…} and "bank2" = {turn 38, veer 29, angle 21…} (numbers are just for illustration). Many Computational Linguistics applications, including word sense disambiguation, parsing attachment determination, and ontology construction, use some form of this approach, and it of course lies at the heart of Information Retrieval, in which instead of a word(sense) rather an entire document is characterized by the distribution of words it contains, organized within a vector space that arranges 'semantically similar' documents in close proximity for easy retrieval. While none of this work is true Semantics in its technical sense (for one thing, it lacks a well-founded notion of compositionality), it is still very useful.

This work usually ignores the fundamental question: what exactly are the boundaries of a collocation? Is "bank teller" a good/useful collocation, or is it better treated as the semantic composition of "bank" and "teller"? How could one decide? The methods developed in and around Distributional Semantics provide some useful diagnostics in this regard.

In this talk I provide a simple no-too-technical review of Distributional Semantics and highlight the principal open questions being studied. I relate these to the questions surrounding collocations, as posed in the workshop, and suggest some lines of interesting research.

Reference

Firth, J.R. 1957. Papers in Linguistics 1934–1951. London: Oxford University Press.

SUBMITTED PAPERS

Statistical methods for Estonian particle verb extraction from text corpus

Eleri Aedmaa
University of Tartu
Tartu, Estonia
eleriaed@ut.ee

Abstract

The current study compares lexical association measures for automatic extraction of Estonian particle verbs from the text corpus. The central focus lies on the impact of the corpus size on the performance of the compared symmetrical association measures. Additionally a piece of empirical evidence of the advantage of asymmetric association measure ΔP for the task of collocation extraction is given.

1 Introduction

Series of studies have been conducted on using association measures (AMs) to identify lexical association between pairs of words that potentially form a holistic unit. However, the question "what is the best AM?" is difficult to answer and the result depends on language, corpus and the type of collocations one wishes to extract (e.g. Evert 2008). Estonian lacks an extensive and systematic comparison of AMs for extracting collocations from a corpus. Hence, it is unknown how the AMs perform on Estonian data and which AMs are most successful for collocation extraction. In the present study, I provide an answer to that question and focus on a subtype of collocations or multi-word expressions, namely particle verbs – frequent and regular phenomena in Estonian and problematic subject in natural language processing.

2 Related research

Many comparative evaluations of the goodness of symmetrical AMs have been carried out. Most of them concentrate on English (e.g. Church & Hanks 1990), German (e.g. Krenn & Evert 2001) or French (e.g. Daille 1996), but collocation extraction work has also been performed for a number of other languages. For example, Pecina (2010) compared and evaluated 82 association measures for Czech collocation extraction. Kis et al. (2003) conducted an experiment on extracting Hungarian multi word lexemes from a corpus, applying statistical methods.

The research on asymmetric measures has been initiated by Michelbacher et al. (2007; 2011), who found that conditional probabilities identify well asymmetric association. Gries (2013) demonstrated that asymmetrical ΔP can identify asymmetric collocation and distinguish collocations with high and low association strengths well.

Studies on the automatic extraction of Estonian multi-word verbs from text corpora exist, but there are no systematic work on association measures. Kaalep and Muischnek (2002) described the extraction of Estonian multi word verbs from text corpora, using a language- and task-specific software tool SENVA. Their goal was to build a comprehensive list of Estonian multi-word verbs and the work resulted in a freely available database of Estonian MWVs, containing 16,000 entries. Uiboaed (2010) assessed few most widely applied AMs to extract phrasal verbs from the Corpus of Estonian Dialects.

3 Estonian particle verb

Estonian particle verb consists of a verb and a verbal particle. The particle can express e.g. direction (1), perfectivity (2) among many other functions.

(1) Ta kukkus trepist alla.
S/he fell stairs down.
'She fell down the stairs'

(2) Perekond suri välja 300 aastat tagasi.
Family died out 300 years ago.
'Family died out 300 years ago'

Most of the particles are homonymous with adpositions, which adds complexity to natural language processing tasks. That disambiguation problem is similar to the one in following English sentences (3) and (4).

(3) The editor looked through the new book.

(4) We looked through the window at the garden.

(5) Toimetaja vaatas uue raamatu läbi.
Editor looked new book through.
'The editor looked through the new book'

(6) Me vaatasime läbi akna aeda.
We looked through window garden.
'We looked through the window at the garden'

In English example (3) *through* is a particle constituent of particle verb *look through*, Estonian verb *vaatama* 'to look' and particle *läbi* 'through' form a particle verb in example (5). As in English example (4) word *through* is not part of the verb, word *läbi* in example (6) is not part of the verb and functions as adposition.

Muischnek et al. (2013) studied the disambiguation problem of Estonian particle verbs. They investigated the role of particle verbs in the Estonian computational syntax in the framework of Constraint Grammar. The two-fold approach, which they used for recognizing the particle verbs, turned out to be successful.

Estonian word order is relatively free, so the order of the components of the particle verb varies, and the verb and the particle do not necessarily appear adjacent to each other within a clause (Kaalep & Muischnek 2006: 60). In addition, the set of Estonian particle verbs has not been strictly specified and the topic is still the subject to debate in theoretical linguistics.

4 Evaluated measures

I have evaluated the following measures: five symmetrical association measures t-test measure (Church & Hanks 1990), log-likelihood measure (Dunning 1993), the X^2 measure (Manning & Schütze 1999), mutual information MI (Church & Hanks 1990), minimum sensitivity MS (Pedersen 1998) and the asymmetrical association measure ΔP (Ellis 2006).

Unlike the symmetrical AMs, ΔP distinguishes two perspectives and does not conflate two probabilities that are very different: $p(word_1|word_2)$ is not the same as $p(word_2|word_1)$ (Gries 2013).

ΔP has not been widely applied in multiword unit extraction tasks, but has been successfully tested in psycholinguistically oriented studies (Ellis & Ferreira-Junior 2009). As ΔP arose from associative learning theory it can be considered psycholinguistically more valid compared to symmetric AMs (Gries 2013: 6).

I compare the association measures against the co-occurrence frequency of verb and verbal particle.

5 The Data

I perform the evaluation of selected association measures over the list of particle verbs (total 1737) presented in the Explanatory Dictionary of Estonian (Langemets 2009). The latter is the gold standard in this work. The study is based on the newspaper part (70 million words) of Estonian Reference Corpus[1]. In order to investigate the impact of corpus size on the performance of tested measures, I divide the corpus into four parts and I aggregate the data stepwise. The first sample includes 5 million words and the next three steps 5, 10 and 50 million words correspondingly.

Corpus data are morphologically analyzed and disambiguated, also the clause boundaries are annotated. It is important to bear in mind that morphological analyzer does not distinguish between homonymous particle and adverb. Candidate pairs consisting of an adverb and a verb are automatically generated within the clause. There are multiple verbs and possible particles in one clause which results to a considerable amount of "noise" in the final list of candidate pairs. AMs are applied to distinguish true particle verbs from the "noise". The adverbs in the current study are constrained to the verbal particles listed in the gold standard.

clauses	CP	TPV	precision	recall
707,979	13,141	1,351	10.3%	77.8%
1,410,474	18,545	1,459	7.9%	84.0%
2,823,255	26,268	1,532	9.6%	88.2%
9,640,426	46,863	1,628	3.5%	93.7%

Table 1: The amount of true particle verbs in the candidate list with respect to the corpus size.

[1] http://www.cl.ut.ee/korpused/segakorpus/index.php

Table 1 presents in what extent the particle verbs (TPV) listed in the gold standard are extracted to the list of candidate pairs (CP) as the corpus size increases. For example, 18,545 candidate pairs have been automatically extracted from the 1,410,474 clause corpus. The total number of true particle verbs in this sample is 1,459, the recall is 84.0% and precision is 7.9%. The recall, in this case, measures what proportion of all the true particle verbs is identified. The precision measures what proportion of all candidate pairs are true particle verbs. Thus, 84.0% of 1737 true particle verbs are identified from the 1,410,474 clause corpus and 7.9% of 18,545 candidate pairs are true particle verbs.

As the corpus size increases the recall increases and precision decreases. For 707,979 clause corpus the recall is 77.8% and the precision is 10.3%, but for the largest dataset (9,640,426 clause corpus) the recall is 93.7% and the precision is 3.5%. Hence, the biggest dataset produces more true particle verbs, but also much "noise".

6 Results

First, I compare symmetrical AMs. I evaluate the set of the *n* highest ranked word combinations for each measure. Table 2 gives the precision values of the *n* highest ranked word combinations n=100, 1,000, with respect to the corpus size.

For the n=100 the precision of t-test is the highest and it does not change with the respect to the corpus size. For the smallest dataset the log-likelihood is the second best, but its precision decreases as the corpus size increases and for the 9,640,426 clause corpus the precision of log-likelihood is lower than the precision of MS and X^2. For n=100 the performance of (simple) frequency is worse than the above-mentioned AMs, but better than MI.

For the n=1000 the results are different than for n=100. Though the results of t-test are the best, irrespective of the corpus size, the performance of log-likelihood is the second best as the corpus size increases. The results of (simple) frequency are similar to the t-test and log-likelihood and it is better than the MS, X^2 and MI for the smallest dataset as well as for the biggest dataset. The MS as a whole produces better results than X^2, but as expected, the precision of MI is significantly lower than others.

The performance of t-test and (simple) frequency do not change significantly as the size of the corpus increases. The change of the performance of log-likelihood depends on the number of the candidate pairs. As the corpus size increases the precision of log-likelihood decreases for n=100 and increases for n=1000. The performance of MS and X^2 increase as the size of the corpus increases. The precision of MI decreases as the corpus size increases.

In addition, Figure 1 shows that for the larger number of candidate pairs (n=2500), the results are rather the same as for n=1000. The t-test performs better than others and its precision increases somewhat as the corpus size increases. The performance of MS, log-likelihood and X^2 increase as the size of the corpus increases. The precision of MI decreases as the corpus size increases. The expansion of the corpus size least affects the performance of the (simple) frequency. All in all, corpus size has an impact on the performance of AMs.

	707,979		1,410,474		2,823,255		9,640,426	
	n=100	n=1000	n=100	n=1000	n=100	n=1000	n=100	n=1000
t-test	95.0%	62.6%	97.0%	63.5%	96.0%	64.5%	95.0%	63.7%
X^2	71.0%	40.2%	71.0%	43.3%	73.0%	47.7%	78.0%	51.5%
log-likelihood	87.0%	58.2%	78.0%	59.1%	77.0%	60.2%	77.0%	59.4%
MI	10.0%	11.7%	7.0%	10.3%	8.0%	9.2%	4.0%	5.6%
MS	80.0%	51.0%	82.0%	51.8%	86.0%	56.6%	85.0%	55.9%
frequency	73.0%	56.9%	73.0%	58.3%	73.0%	59.0%	73.0%	57.8%

Table 2: The precision values of the n=100, 1000 highest ranked word combinations with respect to the corpus size.

Figure 1 presents that all precision curves are substantially above the baseline. The best AM is the t-test, followed by the (simple) frequency, log-likelihood function, MS and X^2. The precision of MI is the lowest, but still somewhat higher than baseline precision. Hence, all compared AMs are suitable for extracting Estonian particle verbs from the newspaper corpora.

Second, I compare bidirectional ΔP with symmetrical AMs. Table 3 gives the number of true particle verbs that two ΔP-values ΔP(verb|adverb) and ΔP(adverb|verb) extracted for n=100, with respect to the corpus size.

The ΔP(verb|adverb) extracted larger number of true particle verbs than ΔP(adverb|verb). This is the result of the fact that ΔP(adverb|verb) raises rare word pairs that contain infrequent or even grammatically incorrect verb. Thus, ΔP(adverb|verb) is suitable for extracting rare word pairs and less-common particle verbs.

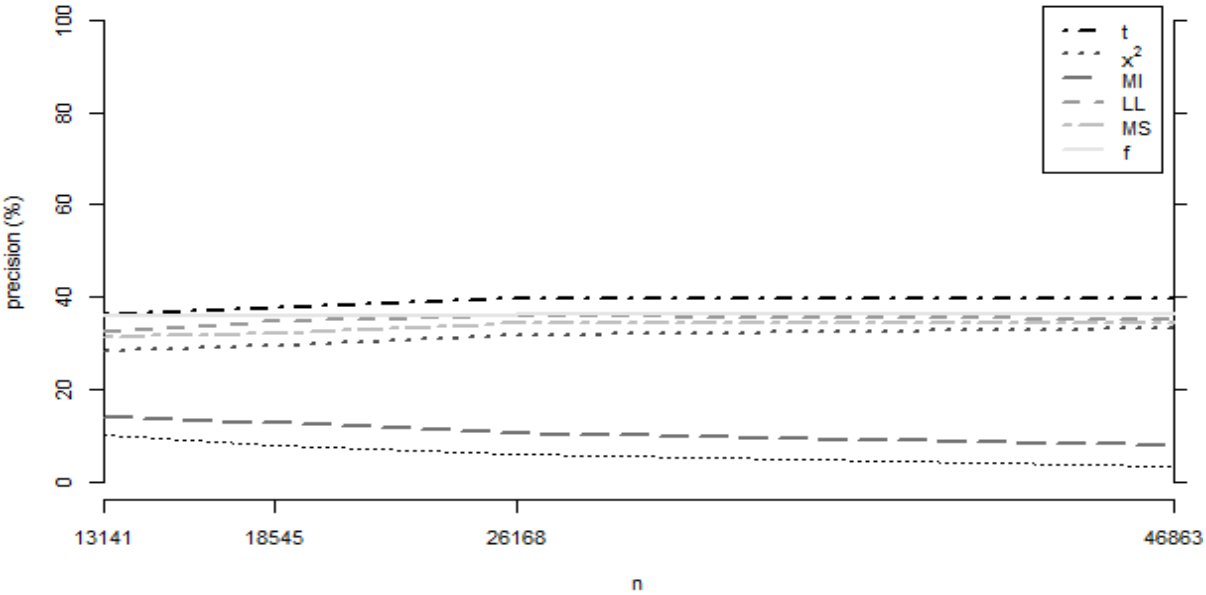


Figure 1. The precision curves for the n=2500 highest ranked word combinations with respect to the corpus size.

	ΔPverb|adverb	ΔPadverb|verb
707,979	67	1
1,410,474	74	1
2,823,255	72	0
9,640,426	72	1

Table 3: The number of true particle verbs of $\Delta P word_2|word_1$ and $\Delta P word_1|word_2$.

For example, the Estonian particle verb *sisse logima* 'to log in' and the verb *logima* 'to log' are both occasional and occur thrice in the dataset. The verbal particle *sisse* 'in' occurs 3008 times in the same dataset. Hence, the verb *logima* always occurs with the particle *in* in the same clause, but the particle *in* can also be a component of another particle verb. So, the value of ΔPsisse|logima is near to 1.0 and *logima* is better hint for *üle* than vice versa.

By contrast, the Estonian particle verb *pärale jõudma* 'to get across' occurs 47 times and the verb *jõudma* 'to get' occurs 9541 times in the dataset. The verbal article *pärale* 'across' occurs 50 times in the same dataset. Thus, the verbal particle *pärale* mostly occurs with the verb *jõudma* in the same clause, but the verb *jõudma* can occur also as a constituent of another particle verb or independently. So, the value of the ΔPpärale|jõudma is near to 0 and the presence of the verb *jõudma* does not increase the likelihood of the verbal particle *pärale*. Hence, ΔPadverb|verb prefers infrequent particle verbs that contain rare verb. However, ΔPadverb|verb successfully indicates the directionality of the association.

In order to compare asymmetrical ΔP with symmetrical AMs I investigated the difference

between two ΔP-values (ΔP(verb|adverb)–ΔP(adverb|verb)).

As there are 2,800 out of 46,863 candidate pairs with the difference between two ΔP-values, the set of the 2800 highest ranked combinations for each measure are included into the comparison. The lists of true particle verbs of symmetrical AMs are generated and unified. The latter is compared to the list of true particle verbs of ΔP (total 374). For n=2800, ΔP extracts 24 true particle verbs that symmetrical AMs do not. Hence, ΔP is successful for extracting Estonian particle verbs.

In addition, Figure 2 shows that there are more particle verbs in the data where verb is much more predictive of adverb than vice versa (negative ΔP(verb|adverb)–ΔP(adverb|verb) values). This is caused by the fact that there are 57 different verbal particles and 615 different verbs in the list of true particle verbs, thus, adverb occurs with numerous different verbs, but verb's distribution is more restricted. Hence, there are more particle verbs where verb selects adverb much more strongly than vice versa.

On the other hand, the particle verbs where adverb is much more predictive of verb than vice versa (positive ΔP(verb|adverb)–ΔP(adverb|verb) values) have higher t-test-values (the best symmetrical AM in current study) than the others with positive difference between the two ΔP-values.

The results of the study reveal that certain Estonian particle verbs can have asymmetric associations and ΔP provides us information about directionality and strength of this association.

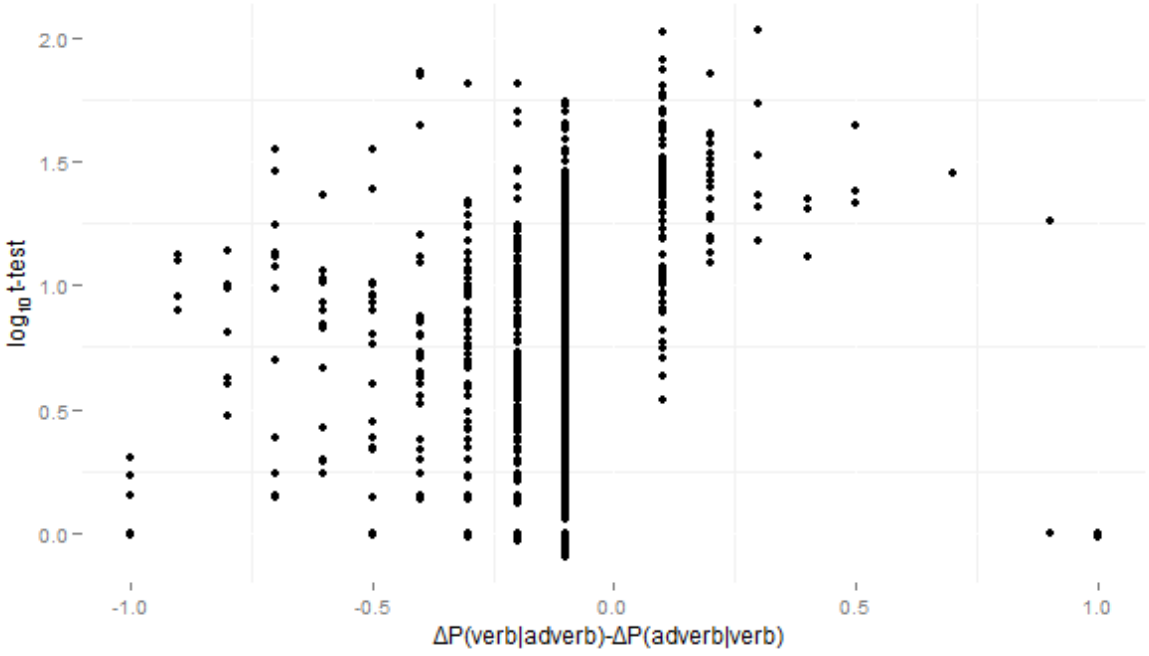


Figure 2: The distribution of true particle verbs according to t-test (on the y-axis, logged) against ΔP(verb|adverb)–ΔP(adverb|verb).

7 Conclusions

This paper focused on the comparison of association measures for extracting Estonian particle verbs from the newspaper part of Estonian Reference Corpus. I investigated the impact of corpus size on the performance of the symmetrical association measures and compared symmetrical association measures and asymmetrical ΔP. Overall, t-test achieved best precision values, but as the corpus size increased, the performances of X^2 and MS improved.

In addition, I have demonstrated that ΔP is successful for the task of particle verb extraction and provides us slightly different and more detailed information about the extracted particle verbs.

The results presented in this paper prove that further study of asymmetrical AMs is necessary and more experiments are needed.

Reference

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1). 22–29.

Daille, Béatrice. 1996. Study and implementation of combined techniques for automatic extraction of terminology. *The balancing act: Combining symbolic and statistical approaches to language* 1. 49–66.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1). 61–74.

Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.

Ellis, Nick C & Fernando Ferreira-Junior. 2009. Constructions and their acquisition Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7(1). 187–220.

Evert, Stefan. 2008. Corpora and collocations. *Corpus Linguistics*. *An International Handbook* 2.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next. *International Journal of Corpus Linguistics* 18(1). 137–166.

Kaalep, Heiki-Jaan & Kadri Muischnek. 2002. Using the Text Corpus to Create a Comprehensive List of Phrasal Verbs. *LREC*.

Kaalep, Heiki-Jaan & Kadri Muischnek. 2006. Multi-word verbs in a flective language: the case of Estonian. *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts*, 57–64.

Kis, Balázs, Begoña Villada, Gosse Bouma, Gábor Ugray, Tamás Bíró, Gábor Pohl & John Nerbonne. 2003. Methods for the Extraction of Hungarian Multi-Word Lexemes. *CLIN*.

Krenn, Brigitte & Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. *Proceedings of the ACL Workshop on Collocations*, 39–46.

Langemets, Margit. 2009. *Eesti keele seletav sõnaraamat*. . Vol. 6. Eesti Keele Sihtasutus.

Manning, Christopher D & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

Michelbacher, Lukas, Stefan Evert & Hinrich Schütze. 2007. Asymmetric association measures. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007)*.

Michelbacher, Lukas, Stefan Evert & Hinrich Schütze. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory* 7(2). 245–276.

Muischnek, Kadri, Kaili Müürisep & Tiina Puolakainen. 2013. Estonian Particle Verbs And Their Syntactic Analysis. *In: Human Language Technologies as a Challenge for Computer Science and Linguistics: 6Th Language & Technology Conference Proceedings*. 338–342.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation* 44(1-2). 137–158.

Pedersen, Ted. 1998. Dependent bigram identification. *AAAI/IAAI*, 1197.

Uiboaed, Kristel. 2010. Statistilised meetodid murdekorpuse ühendverbide tuvastamisel. *Eesti Rakenduslingvistika Ühingu aastaraamat*(6). 307–326.

Variational models in collocation: taxonomic relations and collocates inheritance

Laura Giacomini
Department of Translation and Interpreting
University of Heidelberg
Plöck 57a, D-69117 Heidelberg
laura.giacomini@iued.uni-heidelberg.de

Abstract

The paper presents part of the results obtained in the frame of investigations conducted at Heidelberg University on corpus methods in translation practice and, in particular, on the topic of paradigmatic collocates variation. It concentrates on collocates inheritance across emotion words by focusing on different syntactic frames and a multilingual perspective in order to highlight the potential benefits of this approach for automatic analysis of word combinations and its applications, e.g. in the fields of e-lexicography and machine translation.

1 Introduction: Purpose and Method

Paradigmatic variation in collocational structures, both on the base(s) and the collocate(s) level, always plays a key role in language production (cf. Hall 2010, Nuccorini 2001) and is far from being limited to the mutual substitutability of near-synonymic lexical elements. In particular, inheritance of collocates (cf. definition of *base/collocate* in Hausmann 1999) observed in the context of an ontology-based semantic analysis, turns out to be an interesting example of how languages tend to build collocational clusters and patterns that are poorly represented in existing lexicographic resources and still cannot be sufficiently grasped by available corpus query systems.

Initial observations made by Giacomini (2012) on collocates inheritance across emotion words in Italian can be summarised as follows:

a- meaning relations inside a semantic field
given a semantic field, a number of semantic (here taxonomic) relations can be identified between its lexical items;

b- semantically-based collocates inheritance
a corpus-based study of the collocational behaviour of these items points out that collocates of hypernymic bases are frequently inherited by the hyponymic bases according to semantic contiguity patterns acknowledged by language use.

This paper enlarges upon the topic of collocates inheritance by focusing on different syntactic frames and a multilingual perspective in order to highlight the potential benefits of this approach for automatic analysis of word combinations and its applications, e.g. in the fields of e-lexicography and machine translation. The paper presents part of the results obtained in the frame of investigations conducted at the Department of Translation and Interpreting of Heidelberg University on corpus methods in translation practice.

Lexical information on word combinations such as collocations (Burger 2007) was automatically retrieved from large multilingual web corpora, syntactically and semantically evaluated and compared with lexicographic data from collocation dictionaries. The focus on relatively small semantic fields, such as some subfields of emotions, and an ontology-based approach to the lexicon had the advantage of highlighting fine-grained semantic clustering of collocational elements and allowed for possible generalisations on this type of paradigmatic variation.

2 Observing Collocates Inheritance in Multilingual Corpora

2.1 Data and analysis

The excerpts from the extracted data contain equivalent collocations in four languages (Italian, French, German and English). Data refer to general language nouns denoting emotions and to the collocations they build in some of their usual syntagmatic constellations. For each collocational pattern, the hypernymic base is emphasized in bold letters and is followed by a list of relevant hy-

ponymic bases that share the same collocate. Taxonomic relations were assessed by using existing language-specific lexical ontologies such as the Princeton WordNet and by introducing the necessary adjustments on the basis of multilingual studies on emotion concepts and words (cf. Niedenthal 2004).

Lexical information was extracted with the help of the corpus-query system Sketch Engine (https://the.sketchengine.co.uk) from large web corpora in the four reference languages, namely itWac, frWac1.1, deTenTen10, and ukWac, that include around 1,5-2,8 billion tokens, are PoS-tagged and lemmatised. This level of annotation was required to identify also co-occurrent but non-adjacent bases and collocates. In particular, collocation candidates were retrieved by means of the Word Sketch function, which groups collocates of a given lexeme along predetermined syntactic patterns.

Relevance and arrangement of equivalent bases were determined through frequency criteria and statistical association measures (MI and logDice). Table 1 and 2 show a selection of collocation candidates obtained from data analysis and display the absolute frequency of each candidate in the corpus. The excerpts include only direct co-hyponyms of a specific base (the base is written in bold characters), but deeper and/or multiple taxonomic levels should also be taken into account in a large-scale analysis. The cross-linguistic comparison has demonstration purposes and is restricted to the most frequent equivalents of the same concept in the displayed languages, but, not least due to its context-free nature, it is not meant to exclude other lexical combinations.

The first data set (Table 1) covers binary combinations with a few syntactic variations on the multilingual level (signaled by =, e.g. nominal compounds like *Angstschrei* besides n-grams). Despite limited semantic specificity of the collocates, their inheritance is governed by selection preferences which do not seem to substantially differ across the four languages.

Table 2 shows collocations following more stringent selection rules. These rules regard, for instance, the polarity of emotion concepts: ancestral modifies names of negative emotions, whereas the word fleeting usually accompanies positive feelings). Another example are emotion nouns which, especially in their role as subjects, require

N(base)+PP	N+PP(base)	V+N(base)
paura (1073), terrore (80), orrore (66), angoscia (79) *della morte*	*grido di* **paura** (27), spavento (24), terrore (61), orrore (17)	*suscitare* **emozioni** (942), paura (154), odio (71), rabbia (59)
peur (155), terreur (47), horreur (21), angoisse (40) *de la mort*	*cri de* **peur** (19), terreur (53), horreur (7), panique (7)	*susciter* **emotions** (669), crainte (153), colère (211), haine (51)
Angst/ Furcht (1020/113), Schrecken (4), Panik (2) *vor dem Tod*	=**Angst***schrei* (94), =vor **Angst** *schreien* (76)	**Emotionen** (45), Gefühl (127), Angst (45), Hass (7) *hervorrufen*
fear/ =afraid (546/58), terror (28), horror (37), *of death*	*=to scream in* **fear/fright** (12/7), horror (12), terror (47)	*to arouse* **emotions** (123), fear (84), hatred (16), anger (67)

Table 1: Generic selection rules.

verbal collocates with specific aspect and Aktionsart (e.g. to creep, denoting a non-stative, continuous action performed by emotions that can manifest themselves gradually and almost unnoticed).

2.2 Results interpretation

The following observations and hypotheses can now be made in relation to the presented data:

- collocates inheritance seems to be particularly recurrent in the case of abstract (or, better, second entity) words, which often feature fuzzy semantic boundaries and overlapping traits;
- due to the overall tendency towards terminological univocity, collocates inheritance is likely to affect the general language more

N(base)+V	A+N(base)	A+N(base)
la **paura** (12), terrore (6), panico (6), angoscia (6) *si insinua*	emozione (13), odio (6), **paura** (189) *ancestrale*	**emozione** (8), gioia (38), piacere (48) *effimero/a*
la **peur** (11), panique (5), angoisse (4) *s'insinue*	**émotion** (6), haine (31), peur (87)*ancestrale*	**sentiment** (5), joie (23), plaisir (42)*éphémère*
die **Angst** (5), Panik (2)*schleicht sich ein*	*ursprüngliche* **Emotion** (3), Hass (4), Angst (4), =Urangst (595)	*vergängliche* **Gefühle** (2), Freude (18)
fear (10), panic (2)*creeps in*	*ancestral* **emotion** (9), hatred (7), fear (9)	*fleeting* **feeling** (5), **emotion** (8), happiness (7), joy (9)

Table 2: Specific selection rules.

than specialised languages (cf. analysis of *ansia*, *angoscia*, *panico* and *fobia* both in general language and in the domains of psychology, psychiatry and philosophy, Giacomini 2012; the study highlighted interesting differences in the way in which the same lexical items behaved in general language and in specialised language from a collocational perspective, with the exception of the subclass of their compounds);

- all selected co-occurrences are compositional, whereas non-compositionality (cf., for instance, semi-idioms such as *to frighten sb out of their wits*, *peur bleue*, *Heidenangst*) possibly inhibits taxonomically contiguous bases from sharing their collocates;

- generally speaking, in a monolingual context, collocations can be semantically grouped together along evident taxonomic patterns across a number of syntactic structures; however,

- the identification of inherited collocates can also highlight differences and similarities in the way in which distinct languages form collocates clusters along their own reality categorization and encoding models.

The findings from the study, which this paper introduces, are based on data extracted from web corpora, which largely match the results obtained with the help of newspapers corpora in Giacomini (2012 and 2013). Testing the validity of the original hypotheses in other semantic fields and specialised domains, also by using alternative corpus types and text genres, could contribute towards a better understanding of the phenomenon. A comparison between corpus data and lexicographic data included in collocation dictionaries (*Macmillan Collocations Dictionary*, Macmillan 2010; *Dizionario Combinatorio Italiano*, Benjamins 2013, *Dictionnaire des combinaisons de mots*, Le Robert 2007; *Wörterbuch der Kollokationen im Deutschen*, de Gruyter 2010) reveals the lack, at least in printed lexicographic resources, of an overall cross-referencing system which enables the user to recognize shared collocates. Undoubtedly, the electronic medium has the potential to offer this type of information and the representation of collocations in e-lexicography would derive significant benefits from further studies on this topic.

3 Conclusions

The practice of translation as well as linguistic applications such as e-lexicography could derive concrete tangible benefits from an in-depth investigation of paradigmatic collocates variation, both from a language-specific and a cross-linguistic point of view.

For NLP purposes, in general, this investigation could possibly lead to the specification of suitable statistical methods for the identification of inheritance patterns in corpora (cf. Roark/Sproat 2007 and work done by Alonso Ramos et al. 2010). The development of collocation-based interlinguistic models would be particularly useful in the field of Machine Translation and in enhancing functionality of Translation Memories. Finally, it is crucial to stress the importance of lexical ontologies for avoiding a fragmentary approach to collocation investigation, allowing for a better descriptive representation of the lexicon.

References

Margarita Alonso Ramos et al. 2010. Tagging collocations for learners. *eLexicography in the 21st century: new challenges, new applications*. Proceedings of eLex: 675-380.

Harald Burger. 2007. *Phraseologie: Eine Einfhrung am Beispiel des Deutschen*. Erich Schmidt Verlag, Berlin.

Laura Giacomini. 2013. Languages in Comparison(s): Using Corpora to Translate Culture-Specific Similes. *SILTA. Studi Italiani di Linguistica Teorica e Applicata* 2013/2: 247-270.

Laura Giacomini. 2012. *Un dizionario elettronico delle collocazioni come rete di relazioni lessicali*. Peter Lang, Frankfurt/Main.

Timothy Hall. 2010. L2 Learner-Made Formulaic Expressions and Constructions. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*. Vol. 10, No. 2: 1-18.

Franz Josef Hausmann. 1999. Praktische Einführung in den Gebrauch des Student's Dictionary of Collocations. *Student's Dictionary of Collocations*. Cornelsen, Berlin: iv-xiii.

Paula M. Niedenthal. 2004. A prototype analysis of the French category 'motion'. *Cognition and Emotion*. 18 (3): 289-312.

Stefania Nuccorini. 2001. Introduction. When a torch becomes a candle: variation in phraseology. *SILTA. Studi Italiani di Linguistica Teorica e Applicata* 2001/2: 193-198.

Brian Roark and Richard Sproat. 2007. *Computational approaches to morphology and syntax*. Oxford University Press, Oxford, UK.

Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams

Daria Kormacheva[1] **Lidia Pivovarova**[2] **Mikhail Kopotev**[1]
University of Helsinki, Finland
[1]Department of Modern Languages
[2]Department of Computer Science

Abstract

This paper focuses on automatic determination of the distributional preferences of words in Russian. We present the comparison of six different measures for collocation extraction, part of which are widely known, while others are less prominent or new. For these metrics we evaluate the semantic stability of automatically obtained bigrams beginning with single-token prepositions. Manual annotation of the first 100 bigrams and comparison with the dictionary of multi-word expressions are used as evaluation measures. Finally, in order to present error analysis, two prepositions are investigated in some details.

1 Introduction

In this paper we present our ongoing research on the distributional preferences of words and their co-occurrences in Russian.

Our research follows the tradition of distributional analysis, which takes its roots in the work of Harris (1951). The core idea of this approach is that the semantic similarity/dissimilarity between words correlates with the distributional properties of their context. The most known line of this research is *distributional semantics*, which is based on the assumption that "at least certain aspects of the meaning of lexical expressions depend on the distributional properties of such expressions, i.e. on the linguistic contexts in which they are observed" (Lenci, 2008). In theory, the distributional properties should be studied on all language levels, including phonetics, prosody, morphology and syntax, semantics, discourse, and pragmatics (Gries, 2010). In practice, however, some properties are more difficult to obtain than others; as a consequence, researchers usually focus on a limited amount of linguistic phenomena.

In particular, multi-word expressions (MWEs), in which a given word participates, form the immediate context of this word; the distributional properties of such context can be used for word categorization and description. However, this immediate context is not homogeneous; it is formed by MWEs of various semantic nature: idioms, multi-word lexemes, collocations, i.e. "co-occurrences of words", and colligations, i.e. "co-occurrence of word forms with grammatical phenomena" (Gries and Divjak, 2009).

Distinguishing all these types of MWEs is not a simple task, since there is no clear boundary between them. For example, a word combination can be simultaneously a collocation and a colligation – in (Stefanowitsch and Gries, 2003) this type of MWE is called *collostruction*. Goldberg (2006) proposed that language as such is a *constructicon*, with fusion being its core nature. Thus, measuring the strength of grammatical and/or lexical relations between words is not a trivial task.

The situation becomes even more complicated for morphologically rich languages, because each word may have several morphological categories that are not independent and interact with each other.

In our project we aim to implement the model able to process MWEs of various nature on an equal basis. It compares the strength of various possible relations between the tokens in a given n-gram and searches for the "underlying cause" that binds the words together: whether it is their morphological categories, or lexical compatibility, or both.

Our research is motivated by the recent studies on *grammatical profiling*, including those by Gries and Divjak (2009), Gries (2010), Janda and Lyashevskaya (2011), Divjak and Arppe (2013).

These works are focused on classification of the certain classes of words using *profiles*, i.e. distributions of grammatical and lexical features of the context. A profile does not necessary include all the context features, but only those for which the word has some distributional preferences. This selectivity, as Janda and Lyashevskaya (2011) fairly point out, is the crucial part of the methodology since "it is necessary to target precisely the level of granularity at which the interaction between linguistic category and morphology (or other formal structure) is most concentrated".

The main difference between these works and our study is that these researchers establish the proper level of granularity *before* the main phase of the analysis, while one of our main goals is to *extract* these profiles from the corpus. As has been mentioned before, we try to implement a unified model; the set of input queries for such a model is unrestricted and, as a consequence, the profiles cannot be set *a priori*.

For example, Janda and Lyashevskaya (2011) have shown that tense, aspect and mood form a sub-paradigm for Russian verbs, while person, number and gender are not relevant to this interaction. However, they have found that, for instance, a particular class of verbs – rude ones – has a significant preference of the singular number in imperfective imperative form. This demonstrates that no language property can be excluded from analysis beforehand.

In the previous stage of this project, (Kopotev et al., 2013), we mainly dealt with colligations. We have developed an algorithm that takes as an input an n-gram, in which one position is an unknown variable, and finds the most stable morphological categories of the words that can fill this gap. An in-depth evaluation focusing on a limited number of linguistic phenomena, namely bigrams beginning with single-token prepositions, has been conducted.

In this paper we continue to investigate the same material, i.e. Russian bigrams that match the [PREPOSITION + NOUN] pattern. Our particular task is to analyse MWEs, which are extracted with the help of our algorithm and can be free or stable to various extents. The n-gram corpus, extracted from a deeply annotated and carefully disambiguated sub-corpus of the Russian National Corpus is used as the data. The size of this corpus is 5 944 188 words of running text.

2 Method

In general, our system takes any n-gram of length 2-4 with one unknown variable as an input and tries to detect the most stable word categories that can stay for this variable. These categories include token, lemma and all morphological categories of the Russian language. The initial query pattern may contain various constraints, for example, number or tense can be specified for the unknown variable. Alternatively, the pattern can be unrestricted and formed only by the combination of the surrounding words.

The most stable lexical and grammatical features for a given query pattern are defined using normalized Kullback-Leibler divergence. The category with the highest value of normalized divergence is considered to be the most significant for the pattern. The detailed algorithm and evaluation of the first results can be found in (Kopotev et al., 2013).

Obviously, the most stable grammatical feature for the [PREPOSITION + NOUN] pattern is case that has maximal divergence for all prepositions. The next step is to determine the exact values of the category; i.e., continuing this example, the particular cases that can co-occur with the preposition. Note, that due to unavoidable noise in corpus annotation, the model cannot return all the values found in the data.

Dealing with grammar, we use simple frequency ratio that is able to find possible cases for each preposition with reasonably high quality: precision 95%, recall 89%, F_1-measure 92% (Kopotev et al., 2013). However, frequency ratio does not demonstrate such a performance on detecting stable *lexical* units.

In this paper we use various statistical measures to extract collocations from raw text data and analyse the obtained results. The following measures we applied:

frequency: $f(p, w)$, where p is the pattern, w is the wordform that can appear within this pattern, $f(p, w)$ is the absolute frequency of the wordform in the pattern.

refined frequency ratio:

$$FR(p, w) = \frac{f(p,w)}{f(w)}$$

, where $f(w)$ is the absolute frequency of the wordform in the general corpus. The grammatical categories of the wordform are taken into account, because its surface form can be ambiguous.

For example, many Russian nouns have the same form in nominative and accusative cases; if such a word occurs within the pattern in accusative case, we use only accusative case to count its corpus frequency.

weighted frequency ratio, which is a frequency ratio multiplied by logarithm of the word frequency in the general corpus:

$$wFR(p,w) = FR(p,w) \times \log f(w)$$

The idea behind this measure is as following. Let us consider two words, w_1 that appears in the corpus 2 times and w_2 that appears in the corpus 1000 times. Let $f(p,w_1) = 1, f(p,w_2) = 500$; hence, $FR(p,w_1) = FR(p,w_2) = 0.5$. It is obvious that the w_1 may appear within the pattern by accident, whereas the fact that w_2 occurs within the pattern 500 times out of 1000 is meaningful. We multiply the frequency ratio by logarithm of the word frequency to give more weight to frequent words.

Finally, we compare these three measures with the following widely used metrics:

mutual information, MI (Church and Hanks, 1990):

$$MI(p,w) = \log \frac{f(p,w)}{f(p) \times f(w)}$$

Dice score, (Daudaravicius, 2010):

$$dice(p,w) = \frac{2 \times f(p,w)}{f(p)+f(w)}$$

t-score, (Church et al., 1991):

$$t-score(p,w) = \frac{f(p,w)-f(w) \times f(p)}{\sqrt{f(p,w)}}$$

Thus, 6 different measures are used for the evaluation in this paper: part of them are widely known, while others are less prominent or new.

3 Experiments and Results

We evaluate the semantic stability of automatically obtained bigrams beginning with single-token prepositions. We investigate 25 prepositions, such as "без" (*without*), "в" (*in/to*), etc. For each preposition, algorithm collects all the bigrams that match the pattern [PREPOSITION + w], where w is a noun. In order to minimize noise in our data, bigrams containing infrequent nouns with $f(w) > 5$ are filtered out.

The remaining bigrams are sorted according to the aforementioned statistical measures, which means that for each preposition 6 different rankings are presented. We then compare these rankings to determine the most appropriate statistical measure. Such a comparison becomes itself a tricky task since no "gold standard", i.e. no complete list of collocations, is available. In this paper we perform two types of evaluation: comparison with the dictionary of multi-word expressions (Rogozhnikova, 2003), and manual annotation of the first 100 bigrams in each ranking.

3.1 Comparison with the dictionary

Explanatory dictionary of expressions equivalent to word (Rogozhnikova, 2003) contains approximately 1500 Russian MWEs. These expressions have various nature and can behave as either lexical or function words. They are not necessary idiomatic in terms of semantics, and their only common property is stability: they have the constant form that allows little or no variation.

In particular, the dictionary contains a vast amount of expressions with prepositions, including complex adverbs, prepositions and conjunctions, as well as idiomatic expressions. They constitute the most comprehensive list of Russian MWEs with prepositions, which is crucial for our current task.

For each ranking, we calculate the *uninterpolated average precision* (Moirón and Tiedemann, 2006; Manning and Schütze, 1999): at each point c of the ranking r where a dictionary entry S_c is found, the precision $P(S_1..S_c)$ is computed and all precision points are then averaged:

$$UAP(r) = \frac{\sum_{S_c} P(S_1..S_c)}{|S_c|}$$

The uninterpolated average precision (UAP) allows us to compare rankings and indirectly measures recall (Manning and Schütze, 1999). Results, showing the UAP for each ranking, are presented in Table 1; we report the results for 17 prepositions only, because the dictionary does not contains any entries for the rest.

It can be seen from the Table 1 that simple frequency is the most appropriate measure to determine fixed expressions and idioms; other frequency-based measures, namely weighted frequency ratio and t-score, demonstrate comparable performance, while the refined frequency ratio, Dice-score and MI are not appropriate for this task.

The possible explanation may be the fact that the dictionary contains many MWEs, equivalent to prepositions, conjunctions or adverbs. It has been shown before, (Yagunova and Pivovarova, 2010), that MI is more appropriate to extract *topical* units

of the corpus – such as complex nominations, terminology and noun groups that are significant for a particular document – while t-score tends to extract *pragmatic* units, which characterize the corpus in general.

3.2 Manual Annotation

The dictionary-based evaluation, presented in the previous section, cannot be considered a complete one. Although the high ranks of dictionary MWEs probably mean that for these expressions the ranking should be considered relevant, we cannot tell anything certain about other bigrams in the ranking. One obvious reason is that for many prepositions there are no entries in the dictionary. For example, although every native speaker of Russian knows the idiom "кроме шуток" (*joking apart*) (literally "all jokes aside"), the dictionary contains no fixed expressions for nouns with the preposition "кроме" (*beyond/except*). Moreover, as it is always the case with the dictionaries, some fixed expressions can be neglected in the list.

Furthermore, fixed expressions and idioms are not the only object of our study. Many MWEs do not fulfil the aforementioned requirement of stability; distributional preferences, which our model should be able to catch, do not necessary lead to the lexical rigidity of the expression.

Thus, in this section we present the second evaluation, based on the manual annotation of the extracted bigrams. The first 100 bigrams in each ranking were manually annotated and each bigram was categorized either as a fixed expression/idiom or as a free word combination. Then the uninterpolated average precision was calculated (see the formulae presented in the Section 3.1). Results, presenting the UAP for each ranking, are shown in the Table 2.

It can be seen from the table that the results we got for the manual annotation are quite similar to those obtained for the dictionary-based evaluation. As before, Dice score and MI proved to be not suitable for this task; frequency-based measures, namely frequency, weighted frequency ratio and t-score again demonstrated approximately the same performance. The refined frequency ratio performed slightly worse than these three measures, although in general the number of collocations obtained using this measure is higher than for the dictionary-based evaluation.

On the whole, these results can be considered negative: the first 100 bigrams extracted using the best statistical measure – weighted frequency ratio – in average contain less than 25% of fixed expressions and idioms. But despite the average low performance of the algorithm, it is worthy to note that there is a high variety among the prepositions. For the results based on manual evaluation and sorted according to the weighted frequency ratio, the UAP varies between 0 and 73.34. This can be partially accounted for by the fact that various Russian prepositions have different tendency to form fixed expressions. Below we will illustrate this on the example of two prepositions.

4 Error Analysis and Discussion

In order to perform error analysis, we investigate the following prepositions: "без" (*without*) and "у" (*near/at*). These prepositions were selected since for "без" (*without*) our method achieved the best result (73% of the bigrams extracted using *wFR* contain fixed expressions and idioms), while "у" (*near/at*) was among the prepositions, for which our method failed. Nevertheless, these two prepositions have a common feature that can be used to improve the performance of our algorithm in the future.

The bigrams restrained by both prepositions are often part of various constructions. Among the first 100 nouns extracted by wFR for preposition "без" (*without*), 11 are parts of the construction ["без"+piece of clothing]: "галстук" (*tie*), "перчатка" (*glove*), "погон" (*epaulette*), "шапка" (*cap*), etc.; 3 are included into construction related to the formalities at border checking points: "виза" (*visa*), "паспорт" (*passport*), "штамп" (*stamp*).

The same holds for the first 100 nouns extracted by wFR for preposition "у" (*near/at*). Nouns obtained for this pattern may be described in terms of the following constructions:
16 bigrams: ["у"+part of house]: "окно" (*window*), "крыльцо" (*porch*), "стена" (*wall*), etc.;
13: ["у"+animal] "кошка" (*cat*), "корова" (*cow*), "млекопитающее" (*mammal*), etc.;
10: ["у"+relative]: "ребенок" (*child*), "папа" (*dad*), "теща" (*mother in low*), etc.;
8: ["у"+part of interior]: "стойка" (*counter*), "телевизор" (*TV-set*), "камин" (*fireplace*), etc.;
6: ["у"+nationality] "немец" (*German*), "русский" (*Russian*), "цыган" (*Gypsy*), etc.

We may see that such constructions constitute

Preposition	**Meaning**	f	rFR	wFR	MI	$dice$	t
без	without	33.16	33.89	**35.58**	1.45	1.14	30.60
в	in/into	24.94	14.64	**29.55**	0.59	2.33	24.90
для	for	3.12	0.17	0.42	0.37	0.07	**4.41**
до	until	26.95	27.74	**38.67**	0.85	0.71	25.44
за	behind	22.62	25.56	**53.13**	0.17	0.16	23.06
из	from	1.28	0.86	**1.43**	0.18	0.10	1.27
из-за	from behind	33.33	29.17	**50.00**	0.42	0.37	33.33
к	to	34.62	3.19	24.84	0.25	0.23	**34.75**
между	between	**25.00**	0.27	0.56	0.38	0.16	**25.00**
на	on	**12.58**	8.32	7.83	0.72	0.47	11.85
от	from	**16.01**	1.98	5.15	0.25	0.16	15.60
перед	in front of	**50.00**	0.35	0.98	0.37	0.18	**50.00**
по	by/up to	**35.83**	16.72	34.36	1.44	0.98	35.22
под	under	**31.50**	20.04	21.73	1.01	0.86	31.13
при	at/by	**43.99**	8.77	43.08	0.75	0.34	**43.99**
про	about	**25.00**	7.69	20.00	0.18	0.18	20.00
с	with	13.20	7.63	**16.85**	0.59	0.58	13.22
Average		***25.48***	*12.18*	*22.60*	*0.59*	*0.53*	*24.93*

Table 1: The number of fixed expressions from the dictionary among Russian [PREPOSITION + NOUN] bigrams. For each preposition we present the uninterpolated average precision for *all* bigrams sorted according to the following measures: f – frequency, rFR – refined frequency ratio, wFR – weighted frequency ratio, MI – mutual information, $dice$ – Dice score, t – t-score.

Preposition	**Meaning**	f	rFR	wFR	MI	$dice$	t
без	without	72.86	68.38	**73.34**	7.17	5.83	72.60
в	in/into	47.93	35.14	**58.40**	7.87	4.33	49.37
для	for	7.28	12.32	**14.69**	0.13	0.42	7.26
до	until	44.03	52.38	**60.93**	0.00	0.00	44.37
за	behind	38.58	44.90	**51.58**	13.11	5.36	38.7
из	from	4.48	7.84	**12.29**	0.00	0.00	4.63
из-за	from behind	10.06	10.90	**11.47**	0.00	0.00	9.97
из-под	from under	6.60	**12.37**	8.92	6.72	8.19	5.99
к	to	11.99	0.97	22.28	2.43	3.19	**23.49**
кроме	beyond/except	**5.18**	3.68	**5.18**	0.00	0.00	**5.18**
между	between	**9.28**	5.18	5.18	2.00	1.88	9.25
на	on	23.95	**39.52**	25.32	10.10	10.16	34.
над	above	**0.48**	0.00	0.00	0.00	0.00	**0.48**
о	about	**0.98**	0.00	0.00	0.00	0.00	0.87
от	from	15.81	10.71	11.06	0.00	0.00	**15.94**
перед	in front of	**11.69**	0.00	0.33	0.00	0.00	**11.69**
по	by/up to	57.50	43.29	**60.18**	7.89	7.89	57.35
под	under	62.68	57.95	**62.69**	0.15	0.01	62.29
при	at/by	**32.49**	11.30	19.49	9.65	7.01	**32.49**
про	about	**3.08**	2.06	**3.08**	0.00	0.00	**3.08**
ради	for	**24.32**	20.11	22.14	3.58	4.03	23.22
с	with	36.34	30.97	**44.11**	0.57	0.75	36.69
у	near/at	3.97	1.92	**4.17**	0.00	0.00	2.92
через	through	4.93	3.23	5.06	**8.59**	5.82	4.94
Average		*22.35*	*19.80*	***24.25***	*3.33*	*2.70*	*23.24*

Table 2: The number of fixed expressions among Russian [PREPOSITION + NOUN] bigrams. For each preposition we present the uninterpolated average precision for *the first 100* bigrams sorted according to the following measures: f – frequency, rFR – refined frequency ratio, wFR – weighted frequency ratio, MI – mutual information, $dice$ – Dice score, t – t-score.

a considerable part of the extracted bigrams. Just to illustrate the point, counting these bigrams as relevant collocations would increase the UAP for "без" (*without*) from 73.34% to 85.47% and for "у" (*near/at*) from 4.17% to 73.82%. Similar observations can be done for other prepositions in our list.

Thus, we must re-think the initial problem statement and aim to not only extract fixed expressions and idioms for a given query pattern, but also deal with the kind of expressions described above. We should further define what is the status of such MWEs as *at the counter*, *at the TV-set*, *at the window*, etc. These are not fixed expressions in a sense. Their meaning can be inferred from the meanings of the parts, and the pattern is productive. Nevertheless, these expressions still have something in common and can be described in terms of constructions that predict some grammatical and semantic features of a word class. So we can suppose that in this case the choice of the collocate is not accidental either. This assumption returns us back to the initial point of this article. The model would be a more accurate representation of natural language, if it deals with collostructions rather than with two separate classes of collocations and colligations.

Practically, we assume that such constructional preferences can be found by similar algorithms if the corpus is semantically annotated. If we would have semantic annotation at our disposal, we would be able to group words according to their semantic tags (e.g., animal, relative or nationality) and extract different kinds of constructions in the same way as we do with other categories. Unfortunately, our data do not contain any semantic annotation and we do not have access to any Russian corpus suitable for this task. Still, in our future work, we will try to bootstrap semantic classes from the data on the grounds of the same procedure of distributional analysis.

Acknowledgements

We are very grateful to M. Pierce and R. Yangarber for their support and contribution to the preparation of this paper. Also we would like to thank E. Rakhilina, O. Lyashevskaya and S. Sharoff for providing us with the data for this research.

References

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Kenneth Church, William Gale, Patrick Hanks, and Donald Kindle. 1991. Using statistics in lexical analysis. *Lexical acquisition: exploiting on-line resources to build a lexicon*.

Vidas Daudaravicius. 2010. Automatic identification of lexical units. *Computational Linguistics and Intelligent text processing CICling-2009*.

Dagmar Divjak and Antti Arppe. 2013. Extracting prototypes from exemplars what can corpus data tell us about concept representation? *Cognitive Linguistics*, 24(2):221–274.

Adele Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press, USA.

Stefan Th. Gries and Dagmar Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. *New directions in cognitive linguistics*, pages 57–75.

Stefan Th Gries. 2010. Behavioral profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon*, 5(3):323–346.

Zellig S. Harris. 1951. *Methods in structural linguistics.* University of Chicago Press.

Laura A. Janda and Olga Lyashevskaya. 2011. Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian. *Cognitive linguistics*, 22(4):719–763.

Mikhail Kopotev, Lidia Pivovarova, Natalia Kochetkova, and Roman Yangarber. 2013. Automatic detection of stable grammatical features in n-grams. In *9th Workshop on Multiword Expressions (MWE 2013), NAACL HLT 2013*, pages 73–81.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Christopher D Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

Begoña Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in*.

Roza Rogozhnikova. 2003. Толковый словарь сочетаний, эквивалентных слову *[Explanatory dictionary of expressions equivalent to words]*. Astrel.

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Elena Yagunova and Lidia Pivovarova. 2010. The nature of collocations in the Russian language. The experience of automatic extraction and classification of the material of news texts. *Automatic Documentation and Mathematical Linguistics*, 44(3):164–175.

Semantic modeling of collocations for lexicographic purposes

Lothar Lemnitzer
Berlin Brandenburgische Akademie der Wissenschaften
Jägerstr. 22/23
D-10117 Berlin
lemnitzer@bbaw.de

Alexander Geyken
Berlin Brandenburgische Akademie der Wissenschaften
Jägerstr. 22/23
D-10117 Berlin
geyken@bbaw.de

Abstract

The study which we will present in our talk aims at investigating and modeling lexical-semantic properties of collocations. A collocation is considered as a relation R between a base X and a collocator Y. Pairs of co-occurring words will be extracted – for a set of bases – from a large German corpus with the help of a sketch-engine-like application ('Wortprofil'). From these sets of co-occurring words, collocations in a narrow sense are selected manually. With these sets of data, the following research questions will be tackled a) concerning the collocators: are we able to classify these into lexical-semantic classes and group them accordingly; b) concerning the bases: are we able to find significant numbers of shared collocates for lexical-semantically related bases and thus reach some form of generalization and regular patterns?

In our study we apply the Meaning-Text Theory of Mel'čuk, more precisely, the concept of lexical functions (LF), to guide our modeling efforts. The idea to employ LF for lexicographic work is not new (e.g. Atkins & Rundell 2008). However, the combination of LF with semantic wordnets for the abstraction over individual bases (aspect b above) has never been used for modeling a larger subset of the lexicon. One expected impact of the work will be guidelines for the encoding of lexical-semantic features of multi-word-lexemes in semasiological dictionaries such as the "Digitales Wörterbuch der Deutschen Sprache".

In our study, we have focused on some lexical items and their collocations in order to test the appropriateness of Lexical Functions to model the phenomena.

The paper proceeds as follows. After an introduction we will, in section 2, outline the preconditions of our work, in terms of the corpora and the language technological tools we have been using. In chapter 4 we will sketch the theoretical framework of our work, which draws mainly on the works of Igor Mel'čuk and his collaborators, i.e. Meaning-Text Theory in general and Lexical Functions (LF) in particular. Chapter 5 is devoted to three examples: two (German) nouns and one adjective. With these examples we will show the merits, but also the shortcomings of the theoretical approach taken.

We will close our paper with our view on the future of our investigations. From the analyses in section 5 it has been shown that the Mel'čukian framework is rich enough for the description and encoding of collocational bases in some part of the lexicon, but is less so in some other parts of the lexicon. An extension of the theoretical framework by including the "Generative Lexicon" approach by James Pustejovsky is therefore planned, since we consider both approaches to be complementary.

1 Einleitung

Kollokationen sind ein interessantes und zugleich mit den Mitteln der Lexikologie und Linguistik schwierig zu fassendes Phänomen. Sie erscheinen zunächst als syntaktisch transparente Kombinationen einfacher lexikalischer Zeichen, weisen aber als Ganzes den arbiträren und konventionalisierten Charakter (komplexer) lexikalischer Zeichen auf und müssen daher in der lexikalischen Semantik angemessen beschrieben und in Wörterbüchern gebucht werden.

Ein Ziel dieser Untersuchung ist es, eine linguistisch informierte Praxis für die lexikalisch-semantische Beschreibung und Gruppierung von Kollokationen bei der Ergänzung und Aktualisierung des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS, vgl. Klein/Geyken 2010) zu etablieren. So ist z.B. die Gruppe der Kollokanten zur Kollokationsbasis *Bau* (i.S.v. Gebäude) zu umfangreich, um einfach alphabetisch gruppiert zu werden. Es können Kollokanten unterschieden werden, die sich auf die Form beziehen (*langgestreckt, zweistöckig),* auf den Stil (*klassizistisch, gotisch)*, auf den Bauherrn (*öffentlich, staatlich)* u.s.w. Es wird ein Beschreibungsmittel gesucht für Gruppen von Kollokanten, die in ein- und derselben syntaktischen Relation zur Kollokationsbasis stehen, sich aber semantisch systematisch hinsichtlich ihres Beitrags zur Basis systematisch unterscheiden.

Wir hoffen, durch eine adäquate Modellierung von Kollokationen den Nutzern dieses Wörterbuchs eine bessere Orientierung in diesem schwierigen Teil des Lexikons zu bieten.

Kollokationen werden in Zusammenhang dieser Arbeit als zweistellige Relationen von einer Kollokationsbasis zu einer Menge von Kollokanten aufgefasst.. Das Hauptinteresse der hier beschriebenen Untersuchung gilt den (Mengen von) Kollokanten, die unter einer gemeinsamen Relation zur Basis stehend beschrieben werden können (weitere Details zur Modellierung finden sich in Abschnitt 4).

In Abschnitt 2 werden wir die Voraussetzungen für die Arbeiten vorstellen. In Abschnitt 3 gehen wir auf den theoretischen Rahmen ein. Im vierten Abschnitt stellen wir unsere (vorläufige) Modellierung semantischer Eigenschaften von Kollokationen vor und in Abschnitt 5 präsentieren wir einige bereits analysierte Beispiele. In Abschnitt 6 zeigen wir die Perspektiven für die weitere Arbeit auf.

2 Voraussetzungen

Die hier präsentierte Untersuchung hat explorativen Charakter (zu den Perspektiven der Arbeit s. Abschnitt 6). Die Datenerhebung stützt sich auf ein etwa 1,7 Milliarden Textwörter großes, linguistisch annotiertes Korpus der deutschen Gegenwartssprache; es umfasst das sog. Kernkorpus des 20. Jahrhunderts, eine ausgewogene Mischung von Texten der Belletristik, der Gebrauchsliteratur, wissenschaftlicher Literatur und von Zeitungstexten (vgl. Geyken 2007) sowie weitere Zeitungskorpora (zum Beispiel die 'Zeit' von 1946 bis heute). Die Daten wurden linguistisch annotiert und mit Hilfe des Dependenzparsers Syncop (vgl. Didakowski 2007) syntaktisch analysiert. Auf diese Analyseebene bezieht sich der ebenfalls an der BBAW entwickelte Kollokationsextraktor "Wortprofil" (vgl. Didakowski/ Geyken 2013). Dadurch, dass die Sätze im Korpus analysiert sind, können die typischen Wortverbindungen nach syntaktischen Relationen gruppiert angezeigt werden (s. Abb. 1). Für weitere Aspekte der semantischen Modellierung von Kollokationen (s. unten) verwenden wir das lexikalisch-semantische Wortnetz GermaNet (vgl. Henrich/Hinrichs 2010).

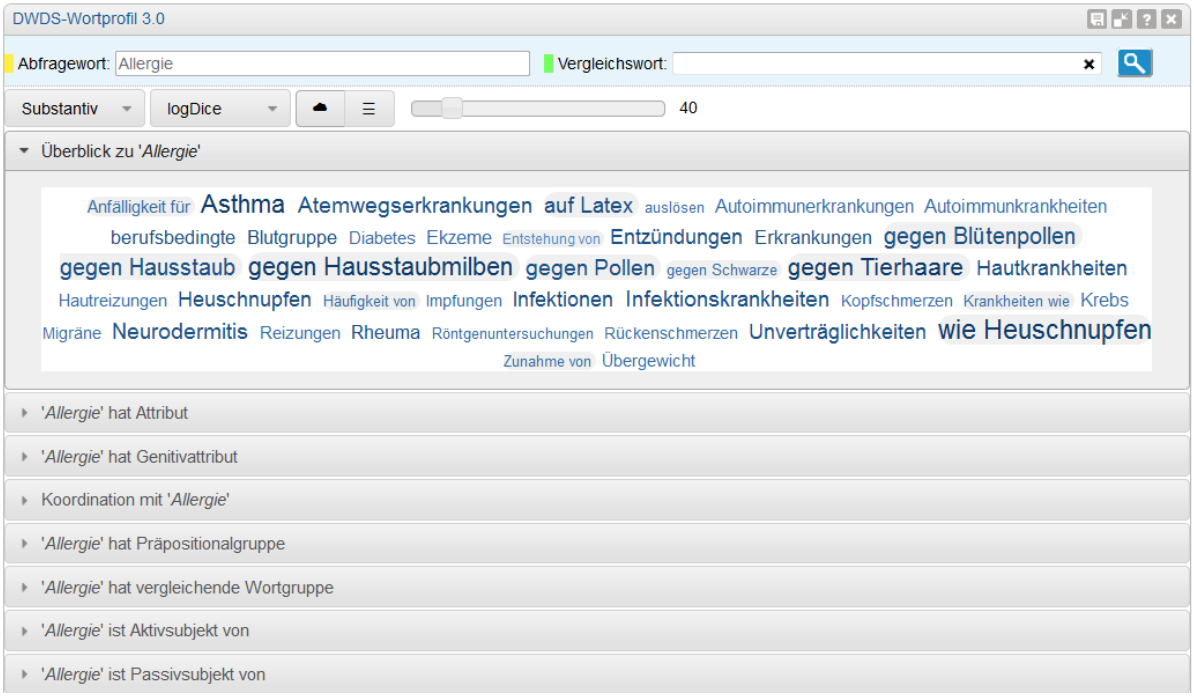


Abb 1: Das Wortprofil für das Stichwort "Allergie". Hinter jedem "Reiter" steht eine Liste von Kollokanten, die in der entsprechenden syntaktischen Relation, z.B. als Adjektivattribut, zur Kollokationsbasis stehen. Für jede Kollokation können die Korpusbelege, aus denen diese ermittelt wurde, betrachtet werden (unter www.dwds.de).

3 Theoretischer Rahmen

Theoretischer Bezugspunkt unserer Untersuchungen ist die von Igor Mel'čuk entwickelte "Meaning Text Theory" (kurz MTT, vgl. Mel'čuk 1998), für die die Konzepte der Kollokation und Lexikalischen Funktion(en) zentral sind.

Mel'čuk definiert Kollokationen als Paare lexikalischer Zeichen, deren Gesamtbedeutung sich zwar prinzipiell aus den Bedeutungen der Bestandteil ergibt, wobei aber ein Element, der Kollokant, in einer Abhängigkeitsbeziehung (engl. 'contingency') zum anderen Element, der Kollokationsbasis, steht. Diese Beziehung wird in der Theorie Mel'čuks als Lexikalische Funktion ("lexical function") modelliert. Mel'čuk entwickelt seine Theorie der lexikalischen Funktionen, insbesondere die der syntagmatischen Beziehungen (= Lexikalische Funktionen) mit dem Ziel, einzelsprachliche Beschränkungen der Kombinierbarkeit von Wörtern in allgemeiner Form darzustellen. Diese Darstellung ist besonders für die Sprachproduktion relevant, da viele dieser Kombinationen, besonders der Kollokationen, usualisiert sind (vgl. Steyer 2003) und sich nicht rein kompositional ergeben. Dies ist insbesondere für das Fremdsprachenlernen relevant (Beispiel: der *starke Raucher* formalisiert durch MAGN(Raucher)=*stark*, im Englischen aber wiedergegeben durch „heavy smoker" (MAGN(smoker)=*heavy*).

Die Menge der Lexikalischen Funktionen, die Mel'čuk und seine Mitarbeiter definiert und in verschiedenen lexikalischen Beschreibungen verwendet haben, besteht aus einer begrenzten Anzahl nzahl von atomaren Funktionen und Kombinationen dieser elementaren Funktionen. Das Inventar erscheint deshalb für die semantische Gruppierung von Kollokanten zu einer Kollokationsbasis gut geeignet, zumal es auch schon einige umfassende lexikographischen Referenzwerke gibt, in denen dieses Inventar angewendet wurde (z.B. Mel'čuk 1984-1999).

4 Das Modell

In der hier vorgestellten Untersuchung werden Kollokationen als zweistellige Relationen (Kollokationsbasis<-REL->{Kollokanten}) modelliert. Daraus ergeben sich die folgenden Aspekte der lexikalisch-semantischen Beschreibung von Kollokationen: a) Gruppierung von Kollokanten einer Kollokationsbasis nach lexikalisch-semantischen Kriterien und b) Exploration des bedeutungsdifferenzierenden Potenzials der Kollokanten für die Kollokationsbasis aufgrund

der Schnitt- und Differenzmengen von Kollokanten für semantisch verwandte Kollokationsbasen.

(zu a) Die Kollokanten zu einem Kollokator werden lexikalisch-semantisch klassifiziert und entsprechend dieser Klassifikation gruppiert. Hierzu wird das Inventar der Lexikalischen Funktionen verwendet.

(zu b) Die Kollokationsbasen werden daraufhin untersucht werden, inwieweit semantisch in Beziehung stehende lexikalische Einheiten, z.B. Ober- und Unterbegriffe (*Antrag → Erstantrag, Asylantrag* etc., s. Abb. 2), bis zu einer noch näher zu bestimmenden Abstraktionsebene, durch gemeinsam mit den meisten dieser lexikalischen Einheiten auftretende Kollokanten charakterisiert werden können. Für die Auswahl lexikalisch-semantisch verwandter Kollokationsbasen werden die lexikalisch-semantischen Hierarchien und lexikalischen Felder in GermaNet (vgl. Henrich/ Hinrichs 2010) herangezogen.

Für die Modellierung von Kollokationen in einem semasiologischen Wörterbuch wie dem DWDS-Wörterbuch bedeutet dies Folgendes:

(a) Kollokanten zu einer Basis, welchen in diesem Fall das Artikelstichwort ist, werden unter einer Lexikalischen Funktion gruppiert. Für die technische Bezeichnung dieser Funktion (z.B. MAGN) muss ein metasprachlicher Ausdruck gefunden werden, der für dem Benutzer die Art der Gruppierung in verständlicher Weise erläutert (z.B. 'verstärkend' vgl. hierzu Polguère 2000).

(b) Kollokanten, die von einer Kollokationsbasis und verwandten Kollokationsbasen geteilt werden (s. das Beispiel in Abbildung 2), können bei dem generellsten der Kollokationsbasen beschrieben werden ('Antrag' im Beispiel). Bei den spezielleren Kollokationsbasen ('Asylantrag' im Beispiel) wird auf diese gemeinsamen Kollokationen lediglich verwiesen.

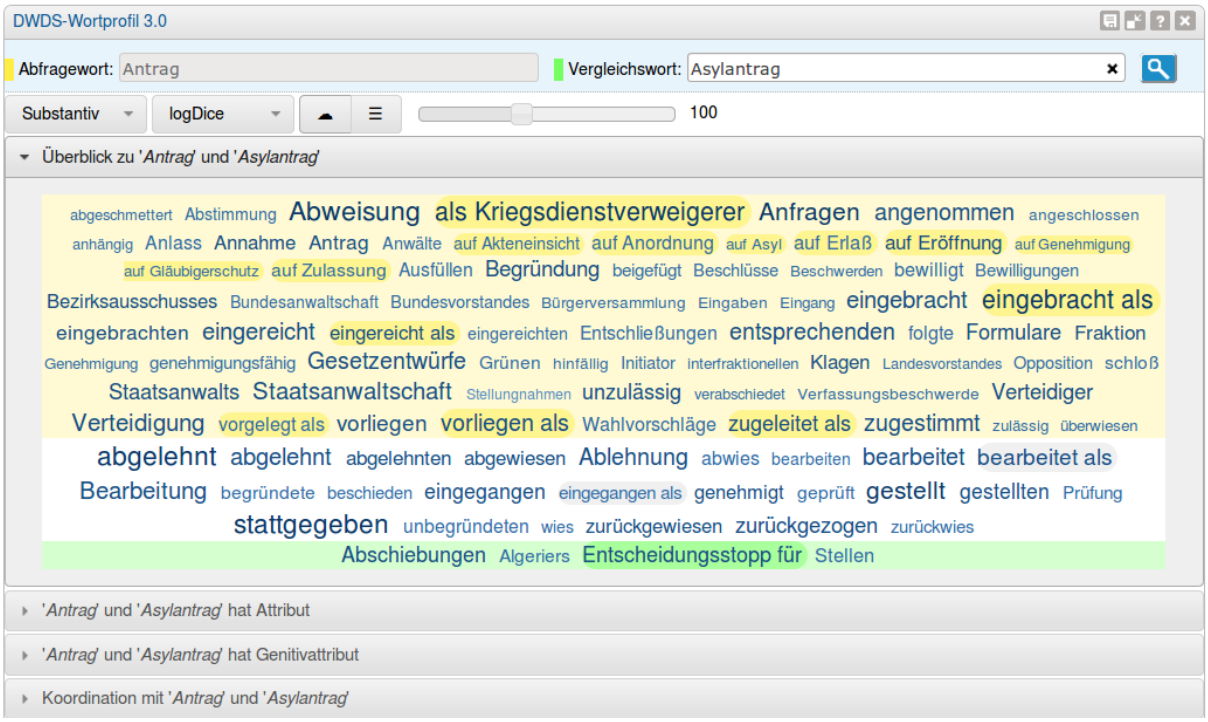


Abb. 2: Kontrastives Wortprofil für die Kollokationsbasen "Antrag" und "Asylantrag". Die gelb hinterlegten Kollokanten sind eher charakteristisch für "Antrag", die grün hinterlegten Kollokanten eher für "Asylantrag". Die weiß hinterlegten Kollokanten sind für beide Kollokationsbasen gleichermaßen typisch.

5 Beispielanalysen

In diesem Abschnitt werden einige Beispielanalysen vorgestellt. Insgesamt werden wir uns in diesem Projekt auf Substantive und Adjektive beschränken.

Zum Vorgehen: aus den Kookkurrenzen des Wortprofils wurden zunächst von zwei Personen die Kollokanten im engeren Sinn ausgewählt und als alphabetische Liste angeordnet. Eine weitere Person, die sich zuvor mit dem Inventar der Lexikalischen Funktionen vertraut gemacht hatte, hat diese Daten so gruppiert, dass die meisten Kollokanten im Wertebereich einer lexikalischen Funktion der jeweiligen Kollokationsbasis fallen.

5.1 Allergie

Allergie bezeichnet eine (zu) heftige Reaktion des Immunsystems gegen bestimmte im Normalfall für den Körper ungefährliche Stoffe. Sie kann zu einem krankheitsähnlichen Zustand bzw. Prozess führen, der deshalb in schweren Fällen medizinisch behandelt wird.

Dementsprechend gibt es eine Vielzahl von Kollokanten, die diese Aspekte der Kollokationsbasen realisieren:

MAGN(Allergie)=*schwer, heftig;*

S0INCEPFUNC0(Allergie)=Entstehung von, entstehen, ausbrechen;

LABOR(Allergie)=*leiden unter*;

INCEPLABOR(Allergie)=erkranken an;

PROPT(Allergie)=*gegen* (i.S.v.: 'eine Allergie gegen Hausstaub');

ANTIPROPT(Allergie)=*gegen* (i.S.v. 'ein Mittel gegen Allergie(n)').

Diese Beispiele mögen ausreichen, um das Prinzip zu verdeutlichen und einige Probleme mit diesem Formalismus: a) die Gruppen von Kollokanten bilden zum Teil unterschiedliche syntaktische Relationen und liegen damit quer zu der syntaktischen Gruppierung (*Entstehung von, entstehen, ausbrechen*); es werden viele komplexe Lexikalische Funktionen benötigt (z.B. INCEPLABOR), die in einem Wörterbuch für einen Benutzerkreis von Nicht-Spezialisten in eine halbwegs verständliche Menge von Deskriptoren übersetzt werden müssen (INCEPLABOR='beginnen von etw. betroffen zu sein').

5.2 Jeans

Mit *Jeans* wird ein Kleidungsstück aus Baumwollstoff bezeichnet, das die Hüften und die Beine umschließt. Dementsprechend sind Prädikationen, die sich auf die Form des Kleidungsstücks und die Farbe und Qualität des Stoffes beziehen, häufig.

A2(Jeans)=*ausgewaschen, verwaschen, ausgefranst, abgewetzt, zerschlissen;*
ANTIVERMINUS(Jeans)=*eng*;
FUNC0ANTIVER(Jeans)=*spannen, schlackern.*

5.3 alternativ

Dieses Wort trägt – in der heute gebrächlichsten Lesart- die Lexikalische Funktion ANTIVER bereits in sich, als Teil seiner Bedeutung, und so ist diese die häufigste LF:
ANTIVER(alternativ)=*vorschlagen,erwägen, produzieren*
ANTIVER(alternativ)=*Medizin,Energie,Wohnmodell, Lebensform*

5.4 Vorläufiges Fazit

Die genannten Beispiele zeigen, dass es Kollokationsbasen gibt, die eine Vielzahl unterschiedlicher Kollokanten im Bereich einer Vielzahl von Lexikalischen Funktionen gruppieren (*Allergie*), aber auch Basen, die viele Kollokanten unter wenigen (*alternativ*) oder wenig aussagekräftigen Lexikalischen Funktionen gruppieren (*Jeans*). Hier stellt sich die Frage nach dem Nutzen des Inventars für die Aufgabe, dem Wörterbuchbenutzer eine Hilfestellung beim Verständnis und der Anwendung der Kollokationen zu einer Basis zu geben.

Es ist zu vermuten, dass andere theoretische Rahmen wie die Theorie des generativen Lexikons von Pustejovsky (vgl. Pustejovsky 1991) für einige Bereiche des Lexikons, z.B. Artefakte, besser geeignet sind. Nichtsdestotrotz sind die Lexikalischen Funktionen für einen Teil des Lexikons, der hier durch das Beispiel *Allergie* repräsentiert wird, eine angemessene Beschreibungssprache, wenn sie für den Wörterbuchbenutzer in eine verständliche (Meta-)Sprache umgesetzt wird.

Interessant scheint uns auch der Vergleich dieser Wortprofile mit den Profilen semantischer verwandter Kollokationsbasen (z.B. Allergie – Krankheit, Unverträglichkeit; Jeans – Hose, Kleidungsstück), um auf diese Weise “reguläre” Kollokationen von spezifischen, irregulären Kollokationen unterscheiden zu können (vgl. hierzu Bosque 2011).

6 Ausblick

Die hier präsentierte Arbeit ist eine Pilotstudie und damit der Beginn eines umfangreicheren Vorhabens, das wir zusammen mit dem Seminar

für Sprachwissenschaft/ Computerlinguistik an der Universität Tübingen durchführen werden. Das Arbeitsprogramm in diesem Vorhaben wird dabei wie folgt erweitert:

a) der theoretische Rahmen wird neben der Meaning-Text-Theorie von Mel'čuk die Theorie des Generativen Lexikons (kurz: GL, Pustejovsky 1991) von James Pustejovsky umfassen, in Abschnitt 5 wird eine Begründung für diese Erweiterung gegeben. Wir gehen davon aus, dass die beiden Theorien sich hinsichtlich ihres Anspruches – der Ansatz von Mel'čuk ist beschreibend, der von Pustejovsky erklärend – wie auch hinsichtlich ihres Fokus – in der GL-Theorie die regulären Verbindungen, in der MTT irreguläre, aber typische Verwendungsmuster – sowie hinsichtlich der Aufgabe 'Semantische Modellierung von Kollokationen und ihrer Bestandteile' komplementär sind;

b) die praktische Anwendung der Modellierung wird sich nicht nur auf ein semasiologisches Wörterbuch (nämlich das DWDS) beziehen, sondern auch auf die semantischen Relationen in GermaNet und deren Erweiterung um syntagmatische Relationen.

Literatur

Atkins, Sue B.T. und Rundell, Michael. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: OUP

Bosque, Ignacio. 2011. Deducing collocations. In: *Proc. of the 5th International Conference on Meaning-Text-Theory, Barcelona, 9-11 Sept. 2011*, S. vi-xxiii.

Didakowski, Jörg. 2007. SynCoP - Combining syntactic tagging with chunking using WFSTs. Linguistik in Potsdam, In: *Proceedings of FSMNLP 2007*. Universitätsverlag Potsdam.

Didakowski, Jörg und Geyken, Alexander. 2013. From DWDS corpora to a German Word Profile – methodological problems and solutions". In: *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network "Internet Lexicography"*. Mannheim: Institut für Deutsche Sprache. (OPAL - Online publizierte Arbeiten zur Linguistik X/2012), S. 43-52.

Geyken, Alexander. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (Hg.. *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London, S. 23-41.

Henrich, Verena und Hinrichs, Erhard. 2010. GernEdiT - The GermaNet Editing Tool. In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010, 2228–2235.

Klein, Wolfgang und Geyken, Alexander. 2010. Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Heid, Ulrich/Schierholz, Stefan/Schweickard, Wolfgang/Wiegand, Herbert Ernst/Gouws, Rufus H./Wolski, Werner (Hg.. *Lexikographica.* Berlin/New York, S. 79-93.

Mel'čuk, Igor et. al. 1984-1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-semantiques*. Montreal: Les Presses de l'Université de Montreal. Vol. I: 1984; Vol. II: 1988; Vol. III: 1992;Vol. IV: 1999.

Mel'čuk, Igor. 1998. Collocations and Lexical Functions. In: A.P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*, Oxford: Clarendon Press, pp. 23-53.

Polguère, Alain. 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In: *Proc. Euralex 2000*, pp. 517-527.

Pustejovsky, James. 1991. The Generative Lexicon. In: *Journal of Computational Linguistics* 17(1991) 4; pp. 409-441

Steyer, Kathrin. 2003. Kookkurrenz, Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Kathrin Steyer (Hg.. *Wortverbindungen - mehr oder weniger fest*. Berlin/NewYork:de Gruyter, S. 87-116

Treatment of Multiword Expressions and Compounds in Bulgarian

Petya Osenova and Kiril Simov
Linguistic Modelling Deparment, IICT-BAS
Acad. G. Bonchev 25A, 1113 Sofia, Bulgaria
petya@bultreebank.org and kivs@bultreebank.org

Abstract

The paper shows that catena representation together with valence information can provide a good way of encoding Multiword Expressions (beyond idioms). It also discusses a strategy for mapping noun/verb compounds with their counterpart syntactic phrases. The data on Multiword Expression comes from BulTreeBank, while the data on compounds comes from a morphological dictionary of Bulgarian.

1 Introduction

Our work is based on the annotation of Multiword Expressions (MWE) in the Bulgarian treebank — BulTreeBank (Simov et al., 2004). We use this representation for parsing and analysis of compounds. BulTreeBank exists in two formats: HPSG-based (original - constituent-based with head annotation and grammatical relations) and Dependency-based (converted from the HPSG-based format). In both of them the representations of the various kinds of Multiword Expressions is an important problem. We need a mechanism for connecting the MWE in the lexicon with their actual usages within the sentences. As an interesting case of MWE at the interface of morphology and syntax we consider Compounds. They are usually derived from several lexical units, have an internal structure with a specified derivation model and semantics. In the paper we are especially interested in the mapping among deverbal compounds and their counterpart syntactic phrases.

Since there is no broadly accepted standard for Multiword Expressions (see about the various classifications in (Villavicencio and Kordoni, 2012)), we will adopt the Multiword Expressions classification, presented in (Sag et al., 2001). They divide them into two groups: lexicalized phrases and institutionalized phrases. The former are further subdivided into fixed-expressions, semi-fixed expressions and syntactically-flexible expressions. Fixed expressions are said to be fully lexicalized and undergoing neither morphosyntactic variation nor internal modification. Semi-fixed expressions have a fixed word order, but "undergo some degree of lexical variation, e.g. in the form of inflection, variation in reflexive form, and determiner selection" (non-decomposable idioms, proper names). Syntactically-flexible expressions show more variation in their word order (light verb constructions, decomposable idioms). We follow the understanding of (O'Grady, 1998) that MWEs have their internal syntactic structure which needs to be represented in the lexicon as well as in the sentence analysis. Such a mapping would provide a mechanism for accessing the literal meaning of MWE when necessary. The inclusion of the compounds into the MWE classification raises additional challenges. As it was mentioned, an important question is the prediction of the compound semantics formed on the basis of the related phrases containing verb + dependents.

In this paper we discuss the usage of the same formal instrument - catena - for their representation and analysis. Catena is a path in the syntactic or morphemic analysis that is continuous in the vertical dimension. Its potential is discussed further in the text. The paper is structured as follows: In the next section a brief review of previous works on catena is presented. In Section 3 a typology of the Miltiword Expressions in BulTreeBank is outlined. Section 4 considers possible approaches for

consistent analyses of MWE. Section 5 introduces the relation of syntax with compound morphology. Section 6 concludes the paper.

2 Related works on catena

The notion of catena (chain) was introduced in (O'Grady, 1998) as a mechanism for representing the syntactic structure of idioms. He showed that for this task there is a need for a definition of syntactic patterns that do not coincide with constituents. He defines the catena in the following way: The words A, B, and C (order irrelevant) form a chain if and only if A immediately dominates B and C, or if and only if A immediately dominates B and B immediately dominates C. In recent years the notion of catena revived again and it was applied also to dependency representations. Catena is used successfully for modelling of problematic language phenomena.

(Gross, 2010) presents the problems in syntax and morphology that have led to the introduction of the subconstituent catena level. Constituency-based analysis faces non-constituent structures in ellipsis, idioms, verb complexes. In morphology the constituent-oriented bracketing paradoxes have been also introduced ([moral] [philosoph - er] vs. [moral philosoph]-er). Catena is viewed as a dependency grammar unit. At the morphological level morphemes (affixes) receive their own nodes forming chains with the roots (such as tenses: has...-(be)en; be...(be)ing, etc.). In (Gross, 2011) the author again advocated his approach on providing a surface-based account of the non-constituent phenomena via the contribution of catena. Here the author introduces a notion at the morphological level — morph catena. Also, he presents the morphological analysis in the Meaning-Text Theory framework, where (due to its strata) there is no problem like the one present in constituency.

Apart from linguistic modeling of language phenomena, catena was used in a number of NLP applications. (Maxwell et al., 2013) presents an approach to Information retrieval based on catenae. The authors consider catena as a mechanism for semantic encoding which overcomes the problems of long-distance paths and elliptical sentences. The employment of catena in NLP applications is additional motivation for us to use it in the modeling of an interface between the valence lexicon, treebank and syntax.

In this paper we consider catena as a unit of syntax and morphology[1]. In a syntactic or morphological tree (constituent or dependency) catena is: Any element (word) or any combination of elements that are continuous in the vertical dimension (y-axis). In syntax it is applied to the idiosyncratic meaning of all sorts, to the syntax of ellipsis mechanisms (e.g. gapping, stripping, VP-ellipsis, pseudogapping, sluicing, answer ellipsis, comparative deletion), to the syntax of predicate-argument structures, and to the syntax of discontinuities (topicalization, wh-fronting, scrambling, extraposition, etc.). In morphology it is applied to the bracketing paradox problem. It provides a mechanism for a (partial) set of interconnected syntactic or morphological relations. The set is partial in cases when the elements of the catena can be extended with additional syntactic or morphological relations to elements outside of the catena. The relations within the catena cannot be changed.

These characteristics of catena make it a good candidate for representing the various types of Multiword Expressions in lexicons and treebanks. In the lexicons each MWE represented as a catena might specify the potential extension of each element of the catena. As part of the morphemic analysis of compounds, catena is also a good candidate for mapping the elements of the syntactic paraphrase of the compound to its morphemic analysis.

3 Multiword Expressions in BulTreeBank

In its inception and development phase, the HPSG-based Treebank adopted the following principles: When the MWE is fixed, which is inseparable, with fixed order and can be viewed as a part-of-speech, it receives lexical treatment. This group concerns the multiword closed class parts-of-speech: multiword prepositions, conjunctions, pronouns, adverbs. There are 1081 occurrences of such multiword closed class parts-of-speech in the treebank, which makes around 1.9% of the token occurrences in the text. Thus, this group is not problematic. Of course, there are also exceptions. For example, one of the multiword indefinite pronouns in Bulgarian shows variation in its ending part: какви-то и да е/са/било (whatever). The varying

[1] http://en.wikipedia.org/wiki/Catena_(linguistics)

part is a 3-person-singular-present-tense-auxiliary, 3-person-plural-present-tense-auxiliary or its 3-person-neuter-singular-past-participle. The semi-fixed expressions (mainly proper names) have been interpreted as Multiword Expressions. However, all the idioms, light verb constructions, etc. have been treated syntactically. This means that in the annotations there is no difference between the literal and idiomatic meaning of the expression: kick the bucket (= kick some object) and kick the bucket (= die). In both cases we indicated that the verb kick takes its nominal complement.

After some exploration of the treebank, such as the extraction of the valency frames and training of statistical parsers, we discovered that the present annotations of Multiword Expressions are not the most useful ones. In both applications the corresponding generalizations are overloaded with specific cases which are not easy to incorporate in more consistent classifications. The group of lexically treated POS remained stable. However, the other two groups were reconsidered. Proper names, as semi-fixed, are treated separately, i.e. as non-Multiword Expressions, since we need coreferencing the single occurrence of the name with the occurrence of two or more parts of the name. Light verb constructions have to be marked as such explicitly in order to differentiate its specific semantics from the semantics of the verbal phrases with semantically non-vacuous verbs. The same holds for the idioms.

4 Possible Approaches for Encoding Multiword Expressions in Treebanks

There are a number of possible approaches for handling idioms, light verb constructions and collocations. The approaches are not necessarily conflicting with each other. However, we also seek for an approach that would give us the mapping between compounds and their syntactic paraphrases.

The first approach is selection-based. This approach is appropriate for Multiword Expressions in which there is a word that can play the role of a head. For example, a verb subcategorizes for only one lexical item or a very constrained set of lexical items. When combined with nouns, such as време (time), форма (shape), надежда (hope), the verb forms idioms - губя време 'lose time' waste one's time; губя форма 'lose shape' to be unfit; губя надежда 'lose hope' lose one's hope). However, when combined with other nouns, such as портфейл (wallet) or роднина (relative), the verb takes canonical complements. In the latter cases, verbs like - обръщам 'to turn' pay - take only noun - внимание 'attention' attention - for making an idiom - обръщам внимание на някого 'to turn attention to somebody' pay attention to somebody. Another example is the verb - вземам 'to take', which combines in such cases with дума 'word' - вземам думата 'to take the word' take the floor. However, light expressions with desemantisized verbs, such as имам have or става happens (имам думата 'I have the word' to have the floor or става дума за нещо 'it happens word' something refers to something)can take numerous semantic classes as dependants. In this case we mark the information only on the head of these Multiword Expressions. In this approach the assumption is that the verb posits its requirements on its dependants. However, a very detailed valency lexicon is required. One problem with this approach is when the dependant elements allow for modifications.

The second approach is construction-based. In this case there is no head. Multiword Expressions are with fixed order and inseparable parts. They are annotated via brackets at the lexical level. One example is the idiom от игла до конец 'from needle to thread' from the beginning to the end. This approach is problematic for syntactically flexible Multiword Expressions.

The third approach marks all the parts of the Multiword Expressions. It is based on the notion of catena as introduced above. Here is an example of this annotation:

(VPS Той (VPC-C (V-C ритна) (N-C камбаната)))
(VPS He (VPC-C (V-C kicked) (N-C bell.DEF)))

He kicked the bucket

where the suffix "-C" marked the catena. This approach maybe adds some spurious compositionality to the idioms, but it would be indispensable for handling idiosyncratic cases, such as separable MWE. However, in order to model the various MWE and to ensure mappings among compounds and related syntactic phrases, the combination of catena with selection-based approach is needed. In case the MWE does not allow for any modifications, for each element of the catena it is specified that the element does not allow any modifications. Thus, catena plus selection-based approach is a powerful means for challenging analyses. Construction-based approach does not make any difference for the strict idioms, since there is

no lexical variation envisaged there.

In Fig. 1 and Fig. 2 we present two sentences from BulTreeBank in which the same verb затварям 'to close' is used in its literal meaning and as a part of idiomatic expression. The catena is highlighted.

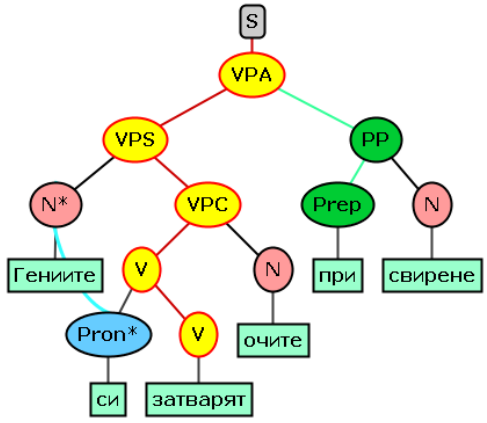


Figure 1: HPSG-based tree for the sentence "Гениите си затварят очите при свирене" ('Geniuses REFL.POSS.SHORT close eyes at playing' *Geniuses close their eyes when playing some instrument.*).

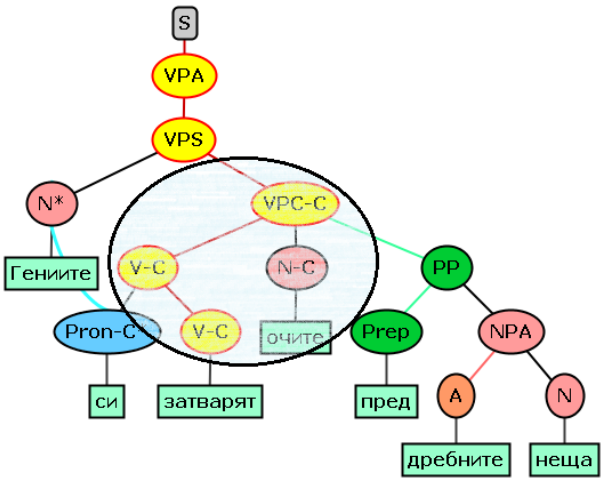


Figure 2: HPSG-based tree for the sentence "Гениите си затварят очите пред дребните неща" ('Geniuses REFL.POSS.SHORT close eyes before minor things' *Geniuses run away from the minor things.*).

In the lexicon each MWE is represented in its canonical (lemmatized) form. The catena is stored in the lexical unit. Additionally, the valency of MWE is expressed for the whole catena or for it parts. When the MWE allows for some modification of its elements - i.e. modifiers of a noun, the lexical unit in the lexicon needs to specify the role of these modifiers. For example, the canonical form of the MWE in Fig. 2 is затварям си очите. Its representation in the lexicon could be as follows:

[**form:** < затварям си очите >
catena:
(VPC-C
(V-C (V-C затварям) (Pron-C си))
(N-C очите)
)
semantics:
не-обръщам-внимание-на-фактите_rel(e,[1]факт)
valency:
< indobj; (PP (P x) (N [1]y)) : x ∈ { пред, за } >
]

The specification above shows that the catena includes the elements 'shut my eyes' in the sense of 'run away from facts', which is presented in the semantics part as a relation. In this part the noun 'fact' is indicated via a structure-sharing mechanism - [1]. This is necessary, because in the valency part of the lexical unit the noun within the subcategorized PP by the catena 'shut my eyes' reproduces some fact from the world. Also, if more than one preposition is possible, they are presented as a set of x-values.

5 From Syntax to Compound Morphology

The catena approach is also very appropriate for modeling the connection among compounds and their syntactic counterparts in Bulgarian. In (Gross, 2011) the notion of 'morph catena' has been explicitly introduced. By granting a node to each morpheme[2], the author makes the problematic morpheme a dominant element over the other depending morphemes. Thus, all these morphemes are under its scope. The catena set includes also the intended meaning.

Here we have in mind examples like the following: a) compound deverbal noun whose counterpart can be expressed only through a free syntactic phrase (билколечение ('herbcuring', curing by herbs), *билколекувам (*'herbcure.1PERS.SG', to cure with herbs) and лекувам с билки ('cure.1PERS.SG with herbs', to cure with herbs); and b) compound deverbal noun whose verbal counterpart can be either a compound too, but verbal, or a free syntactic phrase (ръкомахане ('handwaving', gesticulating), ръкомахам

[2]Such as, histor-ic-al novel-ist where the morpheme 'ist' dominates the rest of the morphemes, thus resolving the bracketing paradoxes of the type [historical [novel-ist]] and [[historical novel]-ist]

('handwave.1PERS.SG', gesticulate) and махам с ръка ('wave with hand', gesticulate).

A previously done survey in (Osenova, 2012) performed over an extracted data from a morphological dictionary (Popov et al., 2003) shows that in Bulgarian head-dependant compounds are more typical for the nominal domain (with a head-final structure), while the free syntactic phrasing is predominant in the verbal domain. Also, regarding the occurrence of dependants in the compounds, subject is rarely present in the verbal domain, while complements and adjuncts are frequent - гласоподавам 'votegive.1PERS.SG', vote - where 'vote' is a complement of 'give'. On the contrary, in the nominal domain also subjects are frequently present, since they are transformed into oblique arguments - снеговалеж 'snowrain', snowing.

Irrespectively of the blocking on some compound verbs, there is a need to establish a mapping between the nominal compound and its free syntactic phrase counterpart. Both expressions are governed by the selection-based rules. Thus, the realization of the dependants in the syntactic phrases relies on the valency information of the head verb only, while the realization of the dependants in the nominal or verb compounds respects also the compound-building constraints.

A mechanism is needed which relates the external syntactic representations with the internal syntax of the counterpart morphological compounds. Moreover, some external arguments which are missing in the compound structures may well appear in the free syntactic phrases, such as: ръкомахам с лявата ръка 'handwave.1PERS.SG with left.DEF hand', I am gesticulating with my left hand, where the complement ръка (hand) is further specified and for that reason is explicitly present. Thus, we can imagine that in the lexicon we have the deverbal noun compounds as well as verb compounds, presented via morphological catena. These words are then connected to the heads of the corresponding syntactic phrases (again in the lexicon), but this time the relations are presented via a syntactic catena tree. We can think of the morphological catena as a rather fixed one, while of the syntactic catena as a rather flexible one, since it would allow also additional arguments or modifiers in specific contexts.

Let us see in more detail how this mapping will be established. The first case is the one where the deverbal nominal compound connects directly to a syntactic phrase (with no grammatical verb compound counterpart). The morph catena will straightforwardly present the tree of: билк-о-леч-ен-ие. However, in the syntactic catena a preposition is inserted according to the valence frame of the verb лекувам (cure): лекувам с билки ('cure.1PERS.SG with herbs', to cure with herbs). Using catena, we can safely connect the non-constituent phrase лекувам с (cure with) with the root morpheme of the head in the compound - леч. Also, all the possible modifiers of билки (herbs) in the syntactic phrase would be connected to the head morpheme билк.

The second case is the one where the nominal compound has mappings to both - verb compound and syntactic phrase. The connection among the nominal and verb compounds is rather trivial, since only the inflections differ. (ръкомахане ('handwaving', gesticulating), ръкомахам ('handwave.1PERS.SG', gesticulate) and махам с ръка ('wave with hand', gesticulate): рък-о-мах-а-не vs. рък-о-мах-а-м. The connection with the syntactic phrase follows the same rules as in the previous case.

Here is the representation of the lexical unit for compound nouns: (билколечение ('herbcuring', curing by herbs):

[**form:** < билколечение >
catena:
(MorphVerbObj-C
(MorphVerb-C [1]билк-) (MorphoObj-C [2]леч-)
)
derivational catena:
(VPC-C
(V-C [2]лекувам (PP-C (P-C с) (N-C [1]билки)))
)
semantics:
лекувам_rel(e,x,y,[4]билки) & номинал_rel(e)
valency:
< mod; (PP (P с) [4](NP ModP* (N билки) ModS*)) :
ModP* or ModS* is not empty >
]

In this example we present two relations. First, the morph catena is presented with its roots (the role of affixes omitted for simplicity). Then, the catena reflecting the derivational syntactic phrase is shown. The correspondences are marked with tags [1] and [2]. The second relation is at the semantic level, where the semantics of the syntactic phrase (лекувам_rel(e,x,y,[3]билки)) is represented fully, and additionally the event is nominalized by the second predicate номинал_rel(e). In the valency list we might have a PP modifier (corresponding to the indirect object in the verb

phrase) of the compound only if the preposition is с (by), the head noun of the preposition complement is the same as the noun in the verbal phrase билки (herbs) and there is at least one modifier of the noun. Thus phrases like: билколечение с български билки ('herbcuring with Bulgarian herbs', curing with Bulgarian herbs) and билколечение с билки, които са събрани през нощта ('herbcuring with Bulgarian herbs that are collected during the night', curing with Bulgarian herbs that were collected at night) are allowed. But phrases with dublicate internal and external arguments like билколечение с билки ('herbcuring with herbs', curing with herbs) are not allowed. Many of the other details are left out here in order to put the focus on the important relations. Among the omitted phenomena are the representation of the subject and patient information as well as the inflection of the compounds.

As a result, we propose a richer valence lexicon, extended with information on mappings between compounds and their counterpart syntactic phrases. The morph catena remains steady, while the syntactic one is flexible in the sense that it encodes the predictive power of adding new material. When connectors (such as prepositions) are added, the prediction is easy due to the information in the valence lexicon. However, when some modifiers come into play, the prediction might become nontrivial and difficult for realization.

6 Conclusions and Future Work

The paper confirms the conclusions from previous works that catena is indispensable means for encoding idioms. Especially for cases where the literal meaning also remained a possible interpretation in addition to the figurative meaning in the respective contexts. We also extend this observation to other types of MWE.

Apart from that, we show that catena is a tool that together with the selection-based approach can ensure mappings between the expressions which have paraphrases on the level of morphology as well as syntax. While at the morphological level the catena is stable, in syntax domain it handles also additional material on prediction from valence lexicons and beyond them.

7 Acknowledgements

This research has received support by the EC's FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches" and by European COST Action IC1207: "PARSEME: PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing."

We are grateful to the two anonymous reviewers, whose remarks, comments, suggestions and encouragement helped us to improve the initial variant of the paper. All errors remain our own responsibility.

References

Thomas Gross. 2010. Chains in syntax and morphology. In Otoguro, Ishikawa, Umemoto, Yoshimoto, and Harada, editors, *PACLIC*, pages 143–152. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Thomas Gross. 2011. Transformational grammarians and other paradoxes. In Igor Boguslavsky and Leo Wanner, editors, *5th International Conference on Meaning-Text Theory*, pages 88–97.

K. Tamsin Maxwell, Jon Oberlander, and W. Bruce Croft. 2013. Feature-based selection of dependency paths in ad hoc information retrieval. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 507–516, Sofia, Bulgaria.

William O'Grady. 1998. The syntax of idioms. *Natural Language and Linguistic Theory*, 16:279–312.

Petya Osenova. 2012. The syntax of words (in Bulgarian). In Diana Blagoeva and Sia Kolkovska, editors, *"The Magic of Words", Linguistic Surveys in honour of prof. Lilia Krumova-Tsvetkova*, Sofia, Bulgaria.

Dimitar Popov, Kiril Simov, Svetlomira Vidinska, and Petya Osenova. 2003. *Spelling Dictionary of Bulgarian (in Bulgarian)*. Nauka i izkustvo, Sofia, Bulgaria.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for nlp. In *In Proc. of the CICLing-2002*, pages 1–15.

Kiril Simov, Petya Osenova, Alexander Simov, and Milen Kouylekov. 2004. Design and implementation of the Bulgarian HPSG-based treebank. In *Journal of Research on Language and Computation*, pages 495–522, Kluwer Academic Publishers.

Aline Villavicencio and Valia Kordoni. 2012. There's light at the end of the tunnel: Multiword Expressions in Theory and Practice, course materials. Technical report, Erasmus Mundus European Masters Program in Language and Communication Technologies (LCT).

Cross-language description of shape: Shape-related properties and Artifacts as retrieved from conventional and novel metaphors across different languages

Francesca Quattri
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
francesca.quattri@connect.polyu.hk

Abstract

In this paper we describe the setup of an ongoing study focused on collocations that deals with shape properties. By shape-properties it is meant properties that relate to shape extensions or physical default measurements, such as height, weight, length, width and volume. The research is currently conducted in Germanic (English, German), Romanic (Italian, French, Spanish) and Sino-Tibetan (Chinese) languages. The aim of this study is to test the validity of figurative expressions, such as collocations, idioms and novel metaphors, through a computable, i e. validated and reproducible model.

1 Introduction

The following paper presents an ongoing study on ontological default physical measurements and figurative language and the way to measure the literal validity of the latter through the defaults in a computable way.

The Suggested Upper Merged Ontology SUMO (Pease, 2011)(Niles and Pease, 2001) has been recently extended with 300+ default physical measurements (examples follow) as applied to the general ontology's Artifacts (which represent the SUMO classes) and most of their subclasses.[1] Artifacts in SUMO are defined as self-connected physical Objects, located in space-time ((Borgo and Vieu, 2009): 12). These measurements are hereby taken into account (a) to investigate how much collocational, metaphorical and idiomatic expressions make sense when considered in their literal meaning, and (b) if the defaults can actually mark a significant difference between, for instance, similes versus metaphors. The research is conducted in English and other languages.

The study could be summarized as follows:

- Default physical measurements are compiled for most of the classes and (partly) subclasses of the SUMO ontology.
- A large inventory of shape-related adjectives and their synonyms/antonyms, as taken from trustworthy dictionaries and thesauri[2], is collected in different languages.
- Figurative expressions (collocations, idioms, metaphors) containing these adjectives (synonyms and antonyms) are collected using corpora and dictionaries. Both expressions with high and low frequency are considered.
- The expressions obtained in the process are then semantically and ontologically analyzed. For the Artifacts in the expressions, defaults are derived from the sample of physical default measurements compiled for SUMO. The values of the respective Artifact are then compared to the intrinsic values of the adjective the Artifact collocates with. In order to do so, it is implied that the expression is analyzed in its *literal*, not conceptual meaning. The predominant default values for that Artifact to be selected are taken with respect to the shape-related adjective the Artifact collocates with in the expression. If for instance the Artifact owns features of length, height and width simultaneously, but the adjective it collocates with in the figura-

[1] The characters in monospace font indicate ontological Artifacs, Devices, Objects, their classes and other Instances as defined in SUMO

[2] Among the sources consulted are The Collins English Dictionary (www.collinsdictionary.com, Wörterbuch English-Deutsch dict.cc (www.dict.cc), Linguee Dictionary for German, French, Spanish and more (www.linguee.com), EUdict European Dictionary (www.eudict.com), IATE - The EU's multilingual term base (iate.europa.eu), Thesaurus.com (thesaurus.com), French dictionary - Larousse.fr (www.larousse.com/dictionaries/french), MDBG English to Chinese dictionary (www.mdbg.net), (ABC, 2010).

tive expression only reflects one property (for instance length), that property only is taken into account for the understanding of the expression.

- The entity Artifact+ adjective is eventually compared to the default ontological values set for Person (instance of Human). The two classes (Artifact and Person) are then compared to test the validity of the *standard* form. By comparing the expression to Person, a simile is created, where the two arguments are a standard metaphor on the one side and Person on the other. The comparison aims at understanding whether the metaphorical expression can possibly have any sense when related to a human being. This comparison of metaphors/idioms is generated in both the case the metaphor/idiom as one argument is very frequent in the selected corpora, or very low in frequency. The only selective criterium for these forms is that they exist in different languages bearing the same meaning.
- Eventually, in order to test the literal validity of not just canonical, but also novel forms, novel figurative expressions are also taken into account, meaning collocations of Artifacts and adjectives that cannot be found in corpora since they are created by a single mind or by random assembly of adjectives and Artifacts. For this part of the study, we rely on an automatic generator of similes from online sources, the simile-finder Sardonicus[3]
- These forms are also analyzed in a simile with Person as second argument, to test their validity as lexico-semantic units.

2 Physical default measurements and SUMO

The Suggested Upper Merged Ontology SUMO is an open-source ontology (Pease, 2011) (Niles and Pease, 2001) developed over the last fourteen years. It includes twenty thousands terms and eighty thousands axioms stated in higher-order logic. The terms can be either searched via Princeton WordNet ® (to which SUMO has been fully merged) or as KB (i. e. knowledge-based) terms. SUMO enables different formal languages, including TPTP and OWL. SUO-KIF has been selected for the development of the KB terms through the open environment SIGMA, integrated in the ontology.

One latest extension of SUMO includes the development of 300+ physical default measurements (notice that the term 'default' is used in SUMO and hereby as a synonym for 'approximation' or 'extension'). The appropriation of defaults to Artifact is usually backed-up by formal sizes and ISO standards. Only for the case these standards are not provided, the compilation of the defaults is left to the judgment of the one who compiles. For instance, in the case of CreditCard, the dimensions given are the same as defined under the international standard ISO/IEC 7810:2003, for which the object must carry a dimension of 3.3 by 2.1 inches. This is the case for many other Artifacts, such as Car, Truck, PaperSheet. Contrarily, as for Book or Painting, the defaults are given on a subjective basis, meaning that it is the compiler that decides what sizes a standard/prototypical[4] Book or Painting should have. In both the case that the sizes are picked from ISO standards or are given on personal judgement, all defaults have been double-checked by the compiler and the SUMO developer. SUMO appears to be one of the few databases where these properties are fully specified[5] and are based on *objective* properties (weight, height, volume, as afore mentioned). Further criteria for the compilation of the measurements are explained in recent work.[6]

The defaults contemplated in SUMO cover a range of maximum and minimum values. These values are not meant to be interpreted like the possible highest or lowest values a particular Artifact has in real world, but they rather aim at representing some scale or range of high and low values that an Artifact, always conceived in its

[3] http://afflatus.ucd.ie/article.do?action=view&articleId=26, developed by Tony Veale and his team at the Creative Language System Group, UCD Dublin.

[4] The concept of 'prototype' here refers to the definition given by Rosch (1973). It follows that the image for the Artifact selected by the compiler to attribute default measurements to follows the principle of graded categories as described by Rosch. For instance, of all the possible books one could possibly imagine, the compiler decides to describe the size of a classical printed Book of medium height and medium weight, thus excluding more modern versions of it, such as e-books or Kindles.

[5] contrarily for instance to DBPedia, whose developers have apparently just recently started the same integration, http://wiki.dbpedia.org/gsoc2013/ideas/CrowdsourceTestsAndRules?v=aq2

[6] Accepted co-authored papers to be presented at the upcoming CogALex workshop, COLING 2014, Dublin, Ireland.

prototypical form, can own.

Given that the study of these forms is still ongoing, the author is able to provide in this paper some examples showing the selection process that goes on with the different size-related expressions retrieved. Not all the collocations considered are eligible to be discussed in terms of novel or standard metaphors.

Let's consider for instance the collocation "high roller", which appears in (COCA, 2012) 89 times. In line with what stated above, `Roller` is firstly considered in its literal meaning as a `Device` or "cylinder that revolves" (SUMO-WordNet noun synset 104101497, enlisted under the SUMO Mappings: `Artifact`). Its default measurements in SUMO appear as it follows. The dimensions are set considering the ISO standards 355:2007:

```
;; Roller
(defaultMinimumWidth (MeasureFn 4 Inch))
(defaultMaximumWidth (MeasureFn 16 Inch))
(defaultMinimumHeight (MeasureFn 4 Inch))
(defaultMaximumLength (MeasureFn 9 Inch))
```

The metaphor with `Roller` ("she is a high roller") does not refer though to the material `Device`, but it refers to somebody who gambles high amounts of money with big losses and big wins. In order to express this figurative meaning, the `Artifact` collocates in English with a height-related adjective ('high'). "High roller" seems to exist in other languages as well, with the same meaning.

- Verschwender, High Roller (De)
- giocatore d'azzardo, scommettitore, high roller (It)

Despite the relatively high frequency of the collocation in the American corpus, it seems to be impossible to consider this expression along the list of potential data in the research, given that the English version seems to pervade in other languages. When the English version is not given, other words are used in other languages, which nevertheless neither do refer to shapes or sizes, nor need the support of size-related adjectives to outer their meaning. Eventually, for the case we consider the novel simile "she is as high as a roller", the comparison between `Person` (not an `Artifact` but still contemplated in its default measurements for this study) and `Roller` do not provide much ground for discussion, apart from letting one think that the person in the case is really short.

```
;;Person
(defaultMinimumHeight (MeasureFn 5 Foot))
(defaultMaximumHeight (MeasureFn 9 Foot))
(defaultMinimumWidth (MeasureFn 1.3 Foot))
(defaultMaximumWidth (MeasureFn 1.6 Foot))
(defaultMinimumMeasure (MeasureFn 119
PoundMass))
(defaultMaximumMeasure (MeasureFn 216
PoundMass))
```

Other conclusions could be drawn from an idiom (and hidden collocation) like "(light as a feather and) thin as a rail". The default measurements for `Rail` in SUMO are:

```
;; Rail
(defaultMinimumHeight Rail (MeasureFn 0.3
Inch))
(defaultMaximumHeight Rail (MeasureFn 0.6
Inch))
(defaultMinimumWidth Rail (MeasureFn 0.4
Inch))
(defaultMaximumWidth Rail (MeasureFn 0.6
Inch))
```

Of the two possible sizes a standard `Rail` can have, the one that needs to be considered in "thin as a rail" is obviously width. The idiom (and inner collocation) "thin as a rail" is chosen among other possible collocations for `Rail` (e. g. "leaned rail", "off rail", "standing rail") given that it exists in other languages and it bears the same meaning. The frequency of the expression is relatively low (10 entries in (COCA, 2012); 2 in the BNC (consulted via SketchEngine, (Kilgariff et al., 2004)), but it exists cross-linguistically:

- spindeldürr (Ge), lit. thin as a spindle
- 细如铁 xì rú tiě (Ch) (e. g. 细如铁线的枝条, xì rú tiě xiàn de zhītiáo), lit. thin branches
- squelettique, (être) très mince (Fr), lit. to be as thin as a skeleton
- delgada/o como un papel (Sp), lit. "thin as a sheet"
- magro come un chiodo (It), lit. "thin as a nail" sdic

Observations that can be drawn by analyzing the forms include:

- The adjective 'thin' remains in all the languages considered, but the `Artifact` changes
- The figurative meaning is not altered, but the literal meaning slightly changes

In order to understand how much the literal meaning for each of the enlisted forms changes, a comparison between the width of the respective `Artifact` is needed. Thus in the following, the respective defaults are extracted from SUMO:

```
;; Spindle
(defaultMinimumWidth 0.7 Inch))
(defaultMaximumWidth 0.8 Inch))
```

with defaults taken from ISO 30, 40, 50 and 60

```
;; Branch
(defaultMinimumWidth 0.07 Inch))
(defaultMaximumWidth 6 Inch))
;; Sheet
(defaultMinimumWidth 1.41 Inch))
(defaultMaximumWidth 33 Inch))
```

with defaults taken from ISO 216, series A, B and C

```
;; Bone
(defaultMinimumWidth 0.6 Inch))
(defaultMaximumWidth 6 Inch))
```

with defaults taken from ISO/TC 150/SC 4

```
;; Nail
(defaultMinimumWidth 0.2 Inch))
(defaultMaximumWidth 1.18 Inch))
```

with defaults taken from ISO XY common nail dimensions

The commonality shared by all kinds of `Artifact`'s widths is that all can be measured in inches. Also, all of them can be objectively thin and all `Artifacts` (`Rail, Spindle, Branch, Sheet, Bone and Nail`) extend in three dimensions.

When contemplated with other `Artifacts` though, the collocation $thin+$ can lead to further intuitions. In the case for instance of "thin soup" or "thin lat", the dimensionality changes, referring, as in the latter cases, to the volume, or density (and thus weight) of the soup or the narrowness and thus the width of the lat or latitude. Although the 3-D dimensionality is a common factor between all analyzed forms, we could nevertheless argue that the standard similes that can be drawn from them with respect to `Person` do not make entirely sense, given that not all `Artifacts` considered can have a width that reaches a width, or length or height measurable in feet. In other words, we can argue that a common expression such as "Jane is as thin as rail" as for instance instantiated in Italian ("magro come un chiodo") does not make literally any sense due to a mismatch in default measurements (remember that we are talking here about prototypical sizes).

Once the commonalities of the standard similes are found, novel collocations are generated to see whether they could make sense (always in their literal meaning) when randomly matched with other `Artifacts`. In the list of factual (and not ironic) matches between 'thin' (as taken from the adjective taxonomy "Domain" > "Physical" > "Physical Perceptibility") and other `Artifacts` it is also included: (1) "thin supermodel", (2) "thin fashion model", (3) "thin deer", (4) "thin potato chip", (5) "thin whippet" and (6) "thin greyhound". While (1) and (2) could make out a literally sound simile ("Jane [`Person`] is as thin as a supermodel [`Person`]), the mismatch reappears in (4) and partially in (5) and (6) (given the small frame of the dog size). Statements such as "Jane is as thin as a deer" or "as a whippet" could be nevertheless stated more credible in terms of defaults than comparing Jane to a paper towel.

3 Conclusion

In the following paper, the settings of an ongoing research on figurative language and default physical measurements are presented.[7] The intent of the research is to analyze the `Objects` or `Artifacts` in the collocation / metaphor / idiom by looking at them as ontological units with default dimensional properties as defined by SUMO. The approach to the "embodiment theory" (Lakoff and Johnson, 1980)(Kövecses, 2005) is therefore taken literally by considering size-related features. These expressions are also disputed in their literal, not figurative meaning.

References

ABC Dictionary. English-Chinese. Chinese-English. 2010. John de Francis and Zhang Yanyin. University of Hawai'i Press, Honolulu.

Corpus of Contemporary American English COCA 1990–2012. Electronic resource: `http://corpus.byu.edu/coca/`

Adam Kilgariff, Pavel Rychly, Pavel Smrz and David Tugwell. 2004. SketchEngine. *Proceedings of EURALEX 2004*. Lorient, France. Electronic resource: `http:\\www.sketchengine.co.uk`

Zoltán Kövecses. 2005. *Metaphor in Culture*. Cambridge University Press, US.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, US.

Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS 2001)*. Christopher A. Welty and Barry Smith (eds.)

Adam Pease. 2011. *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA.

Eleanor Rosch. 1973. On the internal structure of perceptual and semantic categories. *Cognitive Development and the Acquisition of Language*. T. Evan Moore (ed.) Academic Press, New York.

Stefano Borgo and Laure Vieu. 2009. Artefacts in formal ontology. Anthonie Meijers (ed.). *Handbook of Philosophy of Technology and Engineering Sciences*. 273–308, Elsevier.

[7] The author starts with the assumption that these forms are often embedded in and easily derivable from one another.

Using compound lists for German decompounding in a back-off scenario

Pedro Bispo Santos
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Dept. of Computer Science, Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de
santos@ukp.informatik.tu-darmstadt.de

Abstract

Lexical resources like GermaNet offer compound lists of reasonable size. These lists can be used as a prior step to existing decompounding algorithms, wherein decompounding algorithms would function as a back-off mechanism. We investigate whether the use of compound lists can enhance dictionary and corpus-based decompounding algorithms. We analyze the effect of using an initial decompounding step based on a compound list derived from GermaNet with a gold standard in German. The obtained results show that applying information from GermaNet can significantly improve all tested decompounding approaches across all metrics. Precision and recall increases statistically significant by .004-.018 and .011-.022 respectively.

1 Introduction

Compounds are words composed of at least two other lexemes and are a frequent linguistic phenomenon which can be found in several languages. English, Greek, Turkish, German, and Scandinavian languages are examples of languages which have compounds. In some languages, compounds can make part of a significant part of the corpus.[1]

Some compounds consist of two lexemes without any further modification, other require a linking element. *doorbell* and *toothbrush* are examples that do not require any change regarding their lexemes. However, this is not the case for every compound. *Verkehrszeichen*(*Verkehr+s+zeichen*, Engl = *traffic sign*) is a compound in German different from the ones presented before in English, as they require a linking element. The Greek word for cardboard box $\chi\alpha\rho\tau\acute{o}\kappa\upsilon\tau\sigma$ ($\chi\alpha\rho\tau\acute{\iota}$+$\kappa\upsilon\tau\acute{\iota}$) is a compound, for which both lexemes are modified as parts of the compound.

Although some compounds contain two other words, they may not be decompounded depending on the application. *Löwenzahn* consists of the terms *Löwe* and *Zahn*, however, this compound should not be split, since the compound itself has a different meaning from its constituents. This and the previous examples show why decompounding is not a straightforward problem to tackle.

Decompounding is of great importance for NLP tasks as its application as a preprocessing step improves results for several tasks. Monz and Rijke (2002) apply decompounding to information retrieval in German and Dutch and obtain an improvement of 25% for German and 70% for Dutch regarding average precision. Koehn and Knight (2003) obtain a performance gain of .039 BLEU in the German-English noun phrase translation task. Adda-Decker et al. (2000) apply decompounding to speech recognition and obtain a drop on the out of vocabulary word rate from 4.5% to 4.0%. These are just some examples of works in the literature that apply decompounding to other tasks. An improvement of decompounding methods might lead to further improvement of these tasks.

Lexical resources like GermaNet (Hamp and Feldweg, 1997) offer related German nouns, verbs, and adjectives semantically by grouping lexical units that express the same concept into synsets and by defining semantic relations between these synsets. Since version 8.0, GermaNet also offers a compound list indicating nouns that are compounds and how they should be split. In this work we tackle the question whether a prior decompounding step with a compound list improves results for existing decompounding algorithms. The existing algorithms are then used as a back-off solution.

[1] Schiller (2005) shows that for a large German newspaper corpus, 5.5% of 9,3 million tokens were identified as compounds.

2 Decompounding algorithms

Decompounding algorithms found in the literature can be divided in two categories: **lexicon-based** algorithms and **corpus-based** algorithms. Some of the **lexicon-based** algorithms base their lexicon on a corpus, although they do not use further information from the corpus. Additional information could be frequencies in monolingual corpora or words alignment in parallel corpora.

Among the **lexicon-based** algorithms there are works like the one from (Monz and Rijke, 2002), which used the CELEX lexical database for Dutch[2] and a tagger-based lexicon for German. The algorithm splits recursively a word from the right to left, as long as the remaining part of the word is also a word, so *Autobahnraststätte* would be split in *(Auto+(bahn+(rast+stätte)))*. They evaluated their results, and got reasonable results for Dutch and German when considering all nouns, more than 70% for micro/macro average precision/recall, but the results were not that good when evaluating only the complex nouns.

Corpus-based algorithms can then be divided in **monolingual** and **bilingual corpora** approaches. Among the **monolingual corpus** approaches there is the work from (Holz and Biemann, 2008) which filters splitting candidates by checking the minimal morpheme frequency in a corpus for each constituent. After this filtering process, it computes the geometrical mean of the constituent frequencies for each candidate and the one with the highest value is selected as the possible candidate. They use two corpora for evaluation, one from the CELEX lexical database for German and one manually constructed. The results were between 50%-70% of precision for both datasets, 1%-16% of recall for the CELEX database, and 36%-68% for the manually generated dataset.

Alfonseca et al. (2008) generates the candidates using a lexicon built from a corpus and then chooses the candidate by using a SVM classifier, wherein each training instance has different kinds of frequency-based features computed from a corpus. Weighted finite state transducers trained on a corpus are used by (Marek, 2006; Schiller, 2005) to split compound words.

Parallel corpora algorithms (Brown, 2002) are based on the idea that compounds in languages like German have their lexemes separated in their corresponding translation when translated to English. The work from (Koehn and Knight, 2003) uses both monolingual and parallel corpora in their work to learn morphological rules for compound splitting.

[2]http://wwwlands2.let.kun.nl/members/software/celex.html

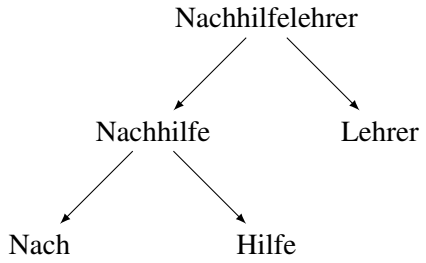


Figure 1: Decompounding of German term *Nachhilfelehrer* (Eng: Private tutor).

However, sometimes these methods might overlap. The work from (Monz and Rijke, 2002) relies on using lexical resources, but the German lexicon it uses for evaluation is based on a corpus. Brown (2002) uses a bilingual dictionary in its evaluation, which is derived from a parallel corpus.

Since some lexical resources offer compounds lists for languages like German. These compounds lists are specify how a compound must be split and the levels of decomposition, as Figure 1 shows. The hypothesis raised by this work is that these compound lists can be used as a prior decompounding step to improve the performance of **lexicon-based** and **corpus-based** algorithms.

3 Evaluation

The lexical resource GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2011) provides a list of compounds with their lexemes. This compound list was semi-automatically generated. A decompounding algorithm was run first, and then human annotators manually corrected the compounds which were wrongly split.

In this paper we present a system that uses this list as a primary source for decompounding and falls back to existing decompounding approaches if a word is not covered by this list. We analyze whether list-based decompounding improves existing decompounding algorithms.

Figure 2 illustrates our classification of the evaluated decompounding algorithms: **lexicon-based**, **corpus-based** and **compound list-based** algorithms. We use **lexicon** and **corpus** based algorithms as a back-off strategy for the GermaNet

Word	Split	Prefix String	Prefix Class	Suffix String	Suffix Class
Holzhaus	Holz-Haus	Holzhaus	4	suahzloH	4
Berggipfel	Berg-gipfel	Berggipfel	4	lefpiggreB	6
Hintergedanke	Hinter-gedanke	Hintergedanke	6	eknadegretniH	7

Table 1: Training set example for the prefix and suffix trie-based classifiers (Holz and Biemann, 2008)

compound list based algorithm.

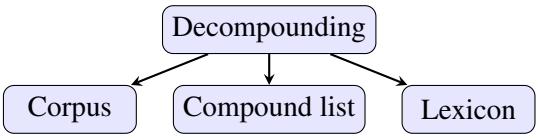


Figure 2: Decompounding algorithms used for evaluation

We use the lexicon-based decompounding API **JWord Splitter**[3]. It performs a dictionary look-up from left to right, and repeats this process if the remaining part of the word is not included in the dictionary. After JWordSplit finds words in both parts (left and right), it creates a split and stops.

This algorithm can generate several splitting candidates. A splitting candidate is a candidate to a possible decomposition. To judge which candidate will be the one selected, a ranking function is responsible for assigning scores to each candidate. We have ranked it by the geometric mean of the unigram frequencies from its constituents. This is based on the idea that the more frequent a candidate is, the more likely it is to be the correct decomposition

$$(\prod_{p_i \in C} count(p_i))^{\frac{1}{n}} \quad (1)$$

wherein C is a decomposition candidate, p_i is a constituent from the candidate and n is the number of constituents the candidate has. This frequency based metric is presented by Koehn and Knight (2003).

ASV Toolbox[4] is a modular collection of tools for the exploration of written language data. This toolbox offers solutions for language detection, POS-tagging, base form reduction, named entity recognition, terminology extraction and so on. It implements a decomposition algorithm which uses an information retrieval data structure called Compact Patrica Tree (CPT). It creates two CPTs (Holz and Biemann, 2008) from a specific corpus, one storing the suffixes for each word and another one storing the prefix, as Table 1 shows. More information about the construction of the CPTs can be found in (Witschel and Biemann, 2005).

A compound list-based decompounding algorithm is also implemented. This decompounding algorithm only splits a word if it is present in the compound list. If it is not there, then it supposes the word is not a compound. The GermaNet compound list[5] is chosen as the compound list for this list-based decompounder. This GermaNet list is also used as the prior step to JWordSplitter and ASV Toolbox in order to prove our hypothesis and check whether there is an improvement.

4 Results

The corpus created by (Marek, 2006) is used as gold standard to evaluate the performance of the decompounding methods. This corpus contains a list of 158,653 compounds, stating how each compound should be split. The compounds were obtained from the issues 01/2000 to 13/2004 of the German computer magazine c't[6], in a semi-automatic approach. Human annotators reviewed the list to identify and correct possible errors.

Koehn and Knight (2003) use a variation of precision and recall for evaluating decompounding performance:

$$\mathrm{P}_{comp} = \frac{\mathrm{cc}}{\mathrm{cc} + \mathrm{wfc}} \quad (2)$$

$$\mathrm{R}_{comp} = \frac{\mathrm{cc}}{\mathrm{cc} + \mathrm{wfc} + \mathrm{wnc}} \quad (3)$$

wherein **correct compound (cc)** is a compound which was correctly split, **wrong faulty compound (wfc)**, a compound which was wrongly split and **wrong non compound (wnc)**, a compound which was not split.

Table 2 shows that although GermaNet list approach's precision is very high. However, its recall is quite low, since it misses too many compounds

[3]https://github.com/danielnaber/jwordsplitter
[4]http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/
[5]http://www.sfs.uni-tuebingen.de/lsd/compounds.shtml
[6]http://www.heise.de/ct/

Algorithm	R_{comp}	P_{comp}
GermaNet list	.083	.917
ASV Toolbox	.755	.799
ASV Toolbox with GermaNet list	**.766**†	**.803**†
JWord	.766	.799
JWord with GermaNet list	**.780**†	**.808**†

Table 2: Evaluation results. † indicates a statistical significant difference according to McNemar's Test.

which are not in the list. It is very hard to obtain a list-based decompounder with a good recall when applied to such datasets since it is impossible to obtain a list with every possible compound from the German language. The results show an improvement of the decompounding methods by the usage of compound lists in recall and precision with a statistical significance according to McNemar's (McNemar, 1947) Test, proving our hypothesis.

Using a list as a prior step could improve cases like *Badezimmer* (*Bad+zimmer*, Engl = bathroom), which is not split by ASV Toolbox and JWord original implementations. The reason is that *Badezimmer* by itself is a very frequent word since both approaches rely on corpus frequency. *Nordwestdeutschland* (*Nord+west+deutschland*, Engl = Germany northwest) is another case which the dictionary-based extension correctly solves. ASVToolbox splits only in two parts the compound, nord-west+deutschland, and JWord Splitter splits as nord+west+deutsch+land.

However, some cases could not be solved for none of the approaches. Cases like *kartenaufbau* (*karte+auf+bau*) are split like *karten+aufbau* by ASV Toolbox and JWord Splitter with and without compound list. GermaNet list does not contain this compound in its compound list, so no method was able to deal with this case. That is the case also for *ausdrucken* (*aus+drucken*), which is considered as not being a compound for every approach. Most of the cases which have a preposition as modifier were the cases which could not be solved by any of the decompounding algorithms.

5 Conclusion and Future Work

This paper raised the hypothesis of whether compound lists improve the performance of decompounding algorithms. We evaluated three different types of decompounding algorithms. Each algorithm was implemented and tested with a German gold standard containing more than 150,000 compounds. The results show that the best outcome is achieved by using a compound list as a prior step to existing decompounding algorithms, and then relying on the original algorithm as a back-off solution if the word is not found in the compound list.

For future work we want to test the algorithms on a dataset containing compounds as well as non-compounds. The reason for that is that we cannot evaluate false positives, in other words, non-compounds that are should not be split, but are. These cases need also to be considered.

References

Martine Adda-Decker, Gilles Adda, and Lori Lamel. 2000. Investigating text normalization and pronunciation variants for german broadcast transcription. In *Sixth International Conference on Spoken Language Processing*, pages 266–269.

Enrique Alfonseca, Slaven Bilac, and Stefan Pharies. 2008. German Decompounding in a Difficult Corpus. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 128–139.

Ralf D. Brown. 2002. Corpus-driven splitting of compound words. In *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Birgit Hamp and Helmut Feldweg. 1997. Germanet - a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Verena Henrich and Erhard Hinrichs. 2011. Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 420–426, Hissar, Bulgaria.

Florian Holz and Chris Biemann. 2008. Unsupervised and knowledge-free learning of compound splits and periphrases. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 117–127.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 187–193.

Torsten Marek. 2006. Analysis of german compounds using weighted finite state transducers. *Bachelor thesis, University of Tübingen*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Christof Monz and Maarten Rijke. 2002. Shallow morphological analysis in monolingual information retrieval for dutch, german, and italian. In *Second Workshop of the Cross-Language Evaluation Forum*, pages 262–277.

Anne Schiller. 2005. German compound analysis with wfsc. In *5th International Workshop on Finite-State Methods and Natural Language Processing*, pages 239–246.

Hans Friedrich Witschel and Chris Biemann. 2005. Rigorous dimensionality reduction through linguistically motivated feature selection for text categorization. In *Proceedings of NODALIDA*.

Multi-label Classification of Semantic Relations in German Nominal Compounds using SVMs

Daniil Sorokin, Corina Dima and Erhard Hinrichs
Collaborative Research Center 833 and Department of Linguistics, University of Tübingen
Wilhelmstrasse 19, 72074 Tübingen, Germany
firstname.lastname@uni-tuebingen.de

Abstract

This paper reports on novel results for the automatic classification of semantic relations that hold between the constituents of nominal compounds in German. It utilizes a hybrid annotation scheme that models semantic relations using a combination of prepositional paraphrases and semantic properties. The machine learning (ML) experiments use the support vector machine (SVM) implementation in Weka for single-label prediction tasks and Weka SVMs in conjunction with the Mulan library for multi-label prediction.

1 Introduction

The fact that the interpretation and generation of compound nouns pose a major challenge for natural language processing has been known for quite some time (Spärck Jones, 1983; Sag et al., 2002). This challenge is particularly evident for languages like English and German, where compounding is a highly productive process of word formation. Apart from splitting the compound into its constituent parts, recognizing the semantic relation that holds between the constituent parts is key to the correct understanding of compounds for humans and machines alike.

The purpose of the present paper is twofold: (i) to present an annotation scheme for nominal compounds that builds on the insights to be gained from applications like machine translation in that it classifies nominal compounds in terms of prepositional paraphrases and semantic properties; (ii) to present a set of machine learning experiments that make use of this hybrid annotation scheme and that demonstrate the disambiguation potential of both prepositions and semantic properties on a dataset of German noun-noun compounds.

The remainder of this paper is structured as follows: Section 2 situates the present research with respect to the state of the art in compound annotation and in automatic classification of semantic relations for compounds. The annotation scheme is introduced in Section 3 and inter-annotator agreement (IAA) results are reported in Section 4; Section 5 describes the German dataset used in the ML experiments, motivates the choice of features used and describes the experimental setup; Section 6 presents the results of the single-label and multi-label experiments; the overall conclusions are presented in Section 7.

2 Related Work

2.1 Annotation Schemes

Several annotation schemes have been proposed for the semantics of compounds in theoretical and computational linguistics. Levi (1978) devises a predicate-based annotation scheme for compound-internal relations, according to which a compound can be formed either via predicate deletion using a fixed set of predicates (CAUSE, HAVE, MAKE, USE, BE, IN, FOR, FROM and ABOUT), or via predicate nominalisation. Warren (1978) proposes a larger taxonomy, where two-place category labels encode ontological distinctions about the constituents of the compound (e.g. SOURCE-RESULT, PART-WHOLE, ORIGIN-OBJ, COMPARANT-COMPARED, etc.), but also makes a survey of the prepositional paraphrases that can be considered typical for each of the categories (e.g. *of*, *with*, *from* and *like*, respectively, for the categories listed above). Downing (1977) and Finin (1980) postulate that there is an infinite number of possible relations. More recent work presents annotation schemes based on prepositional paraphrases (Lauer, 1995), verbal paraphrases (Ó Séaghdha, 2008), or semantic categories (Rosario and Hearst, 2001; Girju et al., 2005; Tratz and Hovy, 2010).

2.2 Automatic Classification

One of the earliest computational approaches to the classification of compound nouns is due to Lauer (1995), who reports an accuracy of 47% at predicting one of 8 possible prepositions using a set of 385 compounds. Rosario and Hearst (2001) obtain 60% accuracy at the task of predicting one of 18 relations using neural networks and a dataset of 1660 compounds. The domain-specific inventory they use was obtained through iterative refinement by considering a set of 2245 extracted compounds and looking for commonalities among them. Girju et al. (2005) use WordNet-based models and SVMs to classify nouns according to an inventory containing 35 semantic relations, and obtain accuracies ranging from 37% to 64%. Kim and Baldwin (2005) report 53% accuracy on the task of identifying one of 20 semantic relations using a WordNet-based similarity approach, given a dataset containing 2169 noun compounds. Ó Séaghdha and Copestake (2013) experiment with the dataset of 1443 compounds introduced in Ó Séaghdha (2008) and obtain 65.4% accuracy when predicting one of 6 possible classes using SVMs and a combination of various types of kernels. Tratz and Hovy (2010) classify English compounds using a new taxonomy with 43 semantic relations, and obtain 79.3% accuracy using a Maximum Entropy classifier on their dataset comprising 17509 compounds and 63.6% accuracy on the Ó Séaghdha (2008) data.

All these efforts have concentrated on English compounds, despite the fact that compounding is a pervasive linguistic phenomenon in many other languages. Recent work by Verhoeven et al. (2012) applied the guidelines proposed by Ó Séaghdha (2008) to annotate compounds in Dutch and Afrikaans with 6 category tags: BE, HAVE, IN, INST, ACTOR and ABOUT. The reported F-Scores are 47.8% on the 1447 compounds Dutch dataset and 51.1% on the 1439 compounds Afrikaans dataset.

3 Annotation Scheme

This section introduces a hybrid annotation scheme that attempts to combine the relative strengths of the property- and the paraphrase-based approaches. The annotation scheme allows the specification of compound-internal semantics via a combined label, typically one preposition and one semantic property. The set of possible prepositions is language-dependent and has to be instantiated each time the annotation scheme is applied to a new language. The semantic properties are, in contrast, language-independent, and can be used directly for annotating nominal compounds in new languages.

Apart from the hybrid nature of the annotation scheme that combines property and paraphrase-based labels, another novel aspect of the annotation scheme is that the annotation is performed on a per head basis rather than on a per compound basis. Thus, the annotation task is defined as follows: given a set of compounds with the same head, identify and group together *similar* compounds. Table 1 illustrates the annotation process for German compounds with the head *Haus* 'house'.

Prepositional paraphrases are one method of defining such similarity. In this case, all the compounds that can be paraphrased using the same preposition will belong to the same group. While this type of grouping seems to do very well in the case of the preposition *aus* 'of', where compounds with the meaning 'houses made of **material**' are clustered together, it is less useful for the other prepositions. In the case of the preposition *für* 'for', the clustered compounds can be further differentiated: *Konzerthaus* and *Auktionhause* are 'houses **used for** concerts or auctions', while *Autohaus* and *Möbelhaus* are 'buildings where certain **goods** are sold', like cars and furniture. In contrast, the compounds paraphrased with the prepositions *in*, *an* and *auf* all refer to 'a type of house specified by a **location**', and should be grouped together. This type of analysis justifies the complementary annotation with a semantic property, in addition to the intuitive but potentially more ambiguous annotation with prepositions.

4 Inter-annotator Agreement Results

An inter-annotator agreement (IAA) study was conducted using a sample of 500 nominal compounds headed by concrete nouns from GermaNet. Written guidelines were given to two student annotators, native speakers of German, who performed the annotation independently. They had previously been trained on the compound annotation task, but had never seen any of the compounds that were part of the study.

Separate IAA scores were computed for the property labeling task, for the preposition labeling task as well as for the task of assigning a com-

German compound	Preposition & Property	English translation
*Auto**haus***	[*für* 'for', goods]	'car dealership' lit. 'car house'
*Möbel**haus***	[*für* 'for', goods]	'furniture store' lit. 'furniture house'
*Mode**haus***	[*für* 'for', goods]	'fashion house'
*Konzert**haus***	[*für* 'for', usage]	'concert hall' lit. 'concert house'
*Auktions**haus***	[*für* 'for', usage]	'auction house'
*Geburts**haus***	[*für* 'for', usage]	'birth house'
*Gäste**haus***	[*für* 'for', user]	'guest house'
*Armen**haus***	[*für* 'for', user]	'poor house'
*Waisen**haus***	[*für* 'for', user]	'orphanage' lit. 'orphan house'
*Holz**haus***	[*aus* 'of', material]	'wooden house'
*Stein**haus***	[*aus* 'of', material]	'stone house'
*Schnee**haus***	[*aus* 'of', material]	'igloo' lit. 'snow house'
*Baum**haus***	[*in* 'in', location]	'tree house'
*Eck**haus***	[*an* 'on', location]	'corner house'
*Land**haus***	[*auf* 'in', location]	'country house'

Table 1: Annotating compounds headed by *Haus* 'house' with prepositions and semantic properties.

bined (property, preposition) label. The property annotation resulted in a percentage of agreement of 76.4% and a Kappa score (Cohen, 1960) of 0.74, while the preposition annotation resulted in a percentage of agreement of 79.5% and a Kappa score of 0.75. It is noteworthy that the amount of agreement is roughly the same for both property and preposition labeling. We conjecture that this similar agreement is due to the parallel annotation as the property labeling helped to disambiguate the preposition labeling and vice versa. Our findings regarding the agreement levels for the preposition and property labels are in stark contrast with the IAA results by Girju et al. (2005). In a similar two-label annotation experiment, they report a Kappa of 0.80 for annotation with the 8 prepositions proposed by Lauer (1995) and 0.58 for the annotation with their inventory of 35 semantic relations.

The agreement measured for combined property and preposition assignment resulted in a percentage of agreement of 68.6%. All of the IAA results reported in this section correspond to a *substantial* agreement according to the classification of Kappa coefficients proposed by Landis and Koch (1977). A more thorough discussion of the reported IAA results is provided in Dima et al. (2014).

5 Experiments

5.1 Dataset

The experiments described in this section use a dataset of German compounds that was obtained by extracting compounds headed by concrete nouns from the German wordnet *GermaNet* (Hamp and Feldweg, 1997). The dataset contains 5082 compounds but only a subset containing 4607 compounds was used for the classification experiments[1]. The 4607 compounds correspond to 2171 distinct modifiers (2.1 compounds per modifier on average) and 360 distinct heads (12.8 compounds per head on average). This dataset was labeled using the annotation scheme described in Section 3 which in its instantiation for German contains 17 prepositions and 38 semantic properties. In terms of size, the German dataset is comparable with English datasets surveyed by Tratz and Hovy (2010), and is, to the best of our knowledge, the largest German noun-noun compound dataset annotated with compound-internal relations.

The annotated dataset exhibits a highly skewed distribution with respect to the semantic property annotation, with 3 properties (*usage*, *part* and *part-1*) accounting for more than 40% of all data instances, and with the other properties forming a long tail distributed over the remaining approximately 60% of the data instances. It is important to note that the skewed distribution of semantic properties reflects the overall patterns of productivity exhibited by nominal compound usage and formation of novel compounds. *Usage* and

[1]The strongly lexicalized compounds such as *Eselsbrücke* ('mnemonic', lit. 'donkey bridge') that were assigned neither property nor preposition and those compounds that were annotated with multiple prepositions were removed because they require a special treatment.

part-whole relations like *part* and *part-1* are generally applicable properties and thus lead to large clusters of compounds. This observation has been corroborated by Moldovan et al. (2004), who report that the *part-whole* relation is the most frequent (19.68%) among all occurrences of their 35 distinct semantic relations in their corpus of annotated compound-internal relations for English. Similarly for German, 34% of all *part-whole* relations recorded in GermaNet release 6.0 involve nominal compounds (Hinrichs et al., 2013).

5.2 Feature Selection and Compound Modeling

The experiments are based on the assumption that the meaning of a compound can be predicted based on the semantic characteristics of its constituents. The models used in the experiments make use of two types of features: corpus-based features extracted from the German corpus *web-news* (Versley and Panchenko, 2012) and knowledge-based features extracted from the German wordnet GermaNet (Henrich and Hinrichs, 2010).

The meaning of the compound as a whole is modeled using distributional information extracted for compound constituents, following the dual setup described in Ó Séaghdha (2008): (i) model the constituents individually, by extracting co-occurrence information separately for the modifier and the head; (ii) model the constituents in conjunction, by considering only those sentences where they appear together. In each case, two lists of reference elements are used for collecting the co-occurrence information from a fixed-size context: a corpus-derived list of the 1000 most frequent German words[2] and the list of 17 prepositions defined by the annotation scheme. The motivation behind extracting the co-occurrences with the prepositions is that in many cases the choice of a correct preposition depends on the lexical associations between constituents and particular prepositions. Lemmas are used as a basis for extracting co-occurrence counts and the context size is fixed to three tokens on the right and on the left. The extracted raw observations are transformed by computing the pointwise mutual information (PMI) scores between the target word and the reference element.

[2] http://wortschatz.uni-leipzig.de/html/wliste.html

The knowledge-based features are extracted from GermaNet. A first set of binary indicators shows which ones from 1000 the most frequent German words occur in the GermaNet gloss of the head or the modifier. Another set of binary indicators encodes which of the 940 top concepts in GermaNet are hypernyms of the either of the constituents. Additional indicators for two-place GermaNet relations such as hypernymy, antonymy, meronymy etc. and the Hirst-St. Onge relatedness measure (Hirst and St-Onge, 1997) are used to model the connections between the head and the modifier. A final set of features encodes which one of the 17 unique beginner categories in GermaNet (*Place*, *Artifact*, *Person*, etc.) includes the head, the modifier and their least common subsumer.

In all the classification experiments described in the next sections a compound is represented by a 6943-dimensional feature vector which contains 3051 (43.9%) co-occurrence-based features and 3892 (56.1%) knowledge-based features. All features have the same weight in the vector.

5.3 Experimental Setup

The hybrid annotation scheme described in Section 3 provides the opportunity to conduct: (i) single-label experiments, in order to assess the predictive strength of the prepositional paraphrases and the semantic properties in isolation; (ii) a multi-label experiment which attempts to simultaneously predict the propositional paraphrases and the semantic properties in question.

The experiments were carried out using Support Vector Machines. SVMs have been successfully applied to a variety of natural language processing tasks including compound interpretation for English and Dutch (Ó Séaghdha, 2008; Verhoeven et al., 2012). We use the SVM implementation from Weka Data Mining Software with a simple linear kernel (Witten et al., 2011) for the single-label classification tasks. The Mulan library (Tsoumakas et al., 2011) was chosen to transform the multi-label classification task in a format that can be used directly by the Weka SVM implementation. All the experiments use a 10-fold cross-validation setup. For each fold the SVM C parameter was optimized through the 5-fold cross-validation on the training test.

Classifier	F-score preposition	F-score property
Baseline	0.182	0.084
SL preposition	0.616	–
SL property	–	0.601
Multi-label classifier	0.639	0.601

Table 2: Single-label (SL) experiment results

Classifier	Combined label accuracy
Baseline	22.66%
SL preposition + SL property	48.44%
Multi-label classifier	59.61%

Table 3: Multi-label experiment results

6 Results

Tables 2 and 3 summarize the results of the single-label and multi-label experiments respectively. In order to estimate the difficulty of the tasks they also includes the most frequent baseline.

For both single-label and multi-label setups, the prediction of prepositional paraphrases achieves a higher F-score compared to the prediction of semantic properties. However, it has to be kept in mind that the set of property labels used in the annotation of the German dataset is twice as large as the set of prepositions used in this dataset. Hence, the classification task involving semantic properties is considerably harder, and it is noteworthy that there is only a small difference in F-score between the two.

The results in Table 2 also clearly show that multi-label prediction outperforms single-label prediction. The F-score for preposition label prediction increases from 0.616 to 0.639 while the F-score for property label prediction remains unchanged. This suggests that the simultaneous prediction of both labels aids in the correct prediction of preposition labels, and has no negative impact on the property prediction.

The deeper reason for conducting the multi-label annotation and the corresponding multi-label experiments derives from the disambiguation requirements for natural language processing applications involving compounds. Applications such as machine translation, where a compound in the source language can correspond to a prepositional phrase in the target language or vice versa, require mutual disambiguation of prepositional paraphrases and semantic properties.

The most significant results of our experiments is that using a multi-label classifier we obtain more than 10% increase in combined label accuracy (see Table 3), i.e. the task of predicting both the semantic property and the prepositional paraphrase correctly. By using the hybrid annotation scheme we are able to give a more accurate specification of the compound-internal relation while improving over the results of automatic classification experiments that use single-label annotation schemes.

Notice also that the multi-label annotation setup can be seen as an instance of the more general scenario of multi-task learning (Caruana, 1997). The results obtained in the experiments reported here corroborate the claim of Caruana (1997) that using a shared representation for multiple learning tasks enables the fine-tuning of classifiers by taking into account patterns that generalize across individual learning tasks.

To the best of our knowledge, the results reported in this paper are the first for the task of automatically classifying the semantic relations for German nominal compounds. While it is always difficult to compare results across different datasets, languages and learning algorithms, the F-scores obtained in our study can be regarded as state-of-the-art results when compared to the studies mentioned in Section 2. The highest accuracy for any dataset of nominal compounds thus far (79.3%) was obtained by Tratz and Hovy (2010) on their dataset containing 17509 instances.

7 Conclusions

This paper has reported on novel results for the automatic classification of semantic relations that hold between the constituents of nominal compounds in German. The experiments use a dataset with a hybrid annotation scheme that models se-

mantic relations using a combination of prepositional paraphrases and semantic properties. To the best of our knowledge, it is the first study of its kind for German and its results are comparable to state-of-the-art results obtained for English on the same task.

Acknowledgements

The second and third author of the present paper would like to thank Verena Henrich and Christina Hoppermann for joint work on the dataset of German compounds used in this paper. We are very grateful to our student assistants Kathrin Adlung, Nadine Balbach, and Tabea Sanwald, who helped us with the annotations reported in this paper. Financial support was provided by the German Research Foundation (DFG) as part of the Collaborative Research Center 'Emergence of Meaning' (SFB 833) and by the German Ministry of Education and Technology (BMBF) as part of the research grant CLARIN-D.

References

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Corina Dima, Verena Henrich, Erhard Hinrichs, and Christina Hoppermann. 2014. How to tell a schneemann from a milchmann: An annotation scheme for compound-internal relations. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.

Tim Finin. 1980. The semantic interpretation of nominal compounds. In *Proceedings of the 1st National Conference on Artificial Intelligence*.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19(4):479–496.

Birgit Hamp and Helmut Feldweg. 1997. Germanet - a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Citeseer.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT - The GermaNet Editing Tool. In *Proceedings of the ACL 2010 System Demonstrations*, pages 19–24, Uppsala, Sweden, July. Association for Computational Linguistics.

Erhard Hinrichs, Verena Henrich, and Reinhild Barkey. 2013. Using part–whole relations for automatic deduction of compound-internal relations in GermaNet. *Language Resources and Evaluation*, 47(3):839–858.

Graeme Hirst and David St-Onge. 1997. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Christiane Fellfaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332, Cambrige, MA, USA. MIT Press.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Compound Nouns*. Ph.D. thesis, Macquarie University.

Judith N. Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 60–67. Association for Computational Linguistics.

Diarmuid Ó Séaghdha and Ann Copestake. 2013. Interpreting compound nouns with kernel methods. *Natural Language Engineering*, 19(03):331–356.

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, Computer Laboratory, University of Cambridge. Published as University of Cambridge Computer Laboratory Technical Report 735.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Karen Spärck Jones. 1983. Compound noun interpretation problems. In Frank Fallside and William A. Woods, editors, *Computer Speech Processing*. Prentice-Hall, Englewood Cliffs, NJ.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. 2011. Mulan: A Java Library for Multi-Label Learning. *Journal of Machine Learning Research*, 12:2411–2414.

Ben Verhoeven, Walter Daelemans, and Gerhard B van Huyssteen. 2012. Classification of Noun-Noun Compound Semantics in Dutch and Afrikaans. In *Proceedings of the Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa*, pages 121–125, Pretoria, South Africa, 11/2012.

Yannick Versley and Yana Panchenko. 2012. Not Just Bigger: Towards Better-Quality Web Corpora. In S. Sharoff & A. Kilgariff, editor, *the 7th Web as Corpus Workshop at WWW2012 (WAC7)*, pages 45–52, Lyon, Frankreich.

Beatrice Warren. 1978. *Semantic Patterns of Noun-Noun Compounds*. Acta Universitatis Gothoburgensis, Göteborg.

Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Amsterdam, 3 edition.

Too Colorful To Be Real
The meanings of multi word patterns

Konrad Szcześniak
University of Silesia
Grota Roweckiego 5
41-205 Sosnowiec, Poland
konrad.szczesniak@us.edu.pl

Abstract

This study focuses on the semantics of word patterns and schematic constructions. Examples of constructions with purportedly rich meanings are shown to convey readings less complex than is claimed in the literature. Finally, it is argued that there are no known mechanisms that could equip constructions with rich semantic content. One potential candidate, pragmatic strengthening, capable of endowing constructions with meanings does not go beyond fairly sparse readings already known to occur in grammatical forms.

1 Introduction

Among the main tenets of Construction Grammar is the proposal of an all-embracing constructicon featuring not only "unicellular" lexical items, but also items ranging from complex words, through multi-word phrases and partially filled phrases to completely schematic syntactic patterns. This represents a radical departure from the modular view of language where the lexicon is separate from syntax. Because there is no clear divide separating typically lexical items from typically syntactic patterns, the solution has been to postulate a comprehensive store where all language forms are stored. As Langacker put it,

> "There is no meaningful distinction between grammar and lexicon. Lexicon, morphology, and syntax form a continuum of symbolic structures, which differ along various parameters but can be divided into separate components only arbitrarily." (Langacker, 1987: 3)

One consequence of this move was that now all language forms are believed to carry meanings, a view in line with the Symbolic Thesis which states that "grammar is symbolic in nature, consisting in the conventional symbolization of semantic structure." (Langacker 1987: 2). While previously it used to be assumed that only lexical items were capable of semantic content, now many cognitive linguists argue that also forms formerly considered to be closed-class items have symbolic properties.

More strikingly, many constructionist analyses are predicated on the premise that closed-class forms may have any kind of meaning. In many cases this is either an implicit assumption, but some authors make it an open assertion, as in Kay & Michaelis (to appear), who propose that "[p]robably any kind of meaning that occurs can be the semantic contribution of a construction." Goldberg (2006) points out that the concern with meanings of constructions is a hallmark of constructionist approaches which

> "emphasize the detailed semantics and distribution of particular words, grammatical morphemes, and cross-linguistically unusual phrasal patterns; the hypothesis behind this methodology is that an account of the rich semantic/pragmatic and complex formal constraints on these patterns readily extends to more general, simple, or regular patterns." (Goldberg 2006: 5)

Similarly, Wierzbicka (2006) claims that there exist "[l]inks between culture and grammar*"* and that "grammatical categories of a language also encode meaning" (p. 171), which she demonstrates by means of many items, among which an "extremely rich and elaborate system of expressive derivation applicable to proper names (specifically, names of persons)" (p. 171) (to be discussed below here). To take another example, in a study of future constructions, Hilpert (2008) signals that they "are viewed as linguistic forms that are endowed with rich meanings that include, but may well go beyond, future time reference." (p. 1)

I believe that expecting all constructions, substantive and schematic ones alike, to have equally rich meanings is a mistake resulting from an unwarranted conclusion drawn from the continuum view. It is one thing to establish the fuzziness of the boundary, and quite another to conclude that it means the absence of that boundary. To take this tack is to commit the continuum fallacy, which involves arguing that if two extremes are connected by small intermediate differences and if at no step can one indicate a decisive difference, then the extremes are the same. To use an analogy, inability to specify at what temperature cold turns to hot should not lead to the conclusion that cold is really the same as hot. But this is more or less what happens when the fuzziness of the distinction is taken as a justification of viewing all language forms as constructions and granting them equal semantic potential.

In what follows, I will look at examples of closed-class elements claimed to express meanings that are more typical of lexical items. Each analysis will conclude with the observation that the rich meanings are not really dedicated semantic effects associated with the forms in question. The meanings are either more general or are only some among many other readings the constructions serve to convey.

2 Constructions with implausible meanings

2.1 Diminutive morphology

One interesting example of grammatical elements associated with a fairly colorful meaning is the diminutive morphology of Russian names reported by Wierzbicka (2006: 171). She observes that the English system is limited, allowing only derivations such as *Johnny* for *John*, while in Russian *Ivan* has a large number of derivations including *Vanja*, *Vanečka*, *Vanjuša*, *Vanjuška*, or *Vanjušečka.* At first blush, the news is rather sensational. Here are fine shades of endearment conveyed by not only one but a series of morphemes which seem to be very close to the closed-class end of the continuum—they are conceptually dependent grams, they are *not* minimal free forms, and most obviously they are not open to additions. What Wierzbicka does not mention is that the elaborate system that generates a series of diminutive names in Russian is not limited to names of persons. Diminutive morphology is a rather commonplace phenomenon found in language after language (and those languages that have elaborate sets of diminutive morphemes also tend to apply them to names). In Russian, the suffixes *-uša* and *-uška (-уша, -ушка,* in Wierzbicka's examples *Vanjuša* and *Vanjuška)* are found equally easily in general nominal word formation, in words like *izbuša and izbuška* (*избуша, избушка*), both diminutive forms of *izba* (*изба* 'hut'). It is natural for many nouns in Russian to come with a series of diminutives like *реченька*, for *reka* (*река* 'river'), which also features a form contain-

ing the suffix *-uška (речушка)*, or the suffix *-čka (-чка)* in *rečka* (*речка*) found in Wierzbicka's example *Vanječka*. Also, combinations of diminutive suffixes like *uš-ečka* (*уш-ечка*) (as in Wierzbicka's example *Vanjušečka)* can be found in nouns like *starušečka (старушечка) or babušečka (бабушечка*) both meaning 'old woman', derived from *stara* (*стара* 'old') and *baba* (баба 'grandmother'), respectively.

Thus, the above advantage of Russian over English is not because affectionate forms of names are somehow incompatible with Anglo tradition, but simply because English has a modest diminutive morphology. English does have quite a few diminutive suffixes (*-en* in *kitten*, *-let* in *starlet*, *-ock* in *bullock*, *-ling* in *duckling*), but they are far from being fully productive.

In this connection, one could also cite the case of Portuguese as an example of a system of diminutives with strange meanings. In Portuguese, diminutives are applied to participles, as in *cansadinho* for *cansado* (tired). But *-inho* is not an exotic participle-specific suffix; Portuguese is merely an example of a language allowing a general suffix to be applied to a category other than noun, which is typical for most languages.

2.2 The *give*-gerund CP construction

Another example of a closed-class construction that seems to be associated with remarkably contentful meanings is the *give*-gerund construction. It is a composite predicate (CP) pattern, a subtype of the fairly large group of light verb constructions, which are characterized by a broad semantic common denominator. However, unlike the super-category they belong to, *give*-gerund patterns seem to have a specialized semantic contribution. In her study of light verbs, Kearns (2002) gives a number of examples of *give*-gerund predicates (*give John a beating/flogging/whipping/thrashing*) and suggests that the verbs in gerund form denote actions involving 'bodily harm'. This observation seems consistent with a considerable number of examples like the following:

(1) a. The patrol officer tried to pin down his arms so that his comrade could *give him a good battering*.

b. We go in there and *give them a kicking*.

c. Give him a serious hiding for that kind of attitude.

d. I have a good mind to walk out there and *give you a sound licking*.

A quick search through uses of the construction reveals that the construction allows practically any native root with the meaning of 'beat' (*spank, belt, smack, cane*). This could give the impression that the construction is indeed dedicated to the expression of causing harm. However, Trousdale (2008) notes that the range of verbs allowed in the construction is much broader. First, he points out that "there is a considerable subset of give-gerund CPs which involve not physical harm but verbal castigation, as in *he gave him a dressing down*." (Trousdale 2008: 41) Examples of this subset are attested frequently:

(2) a. She *gave me a severe tongue lashing*.

b. I'm going *to give him a good chewing out* when I get home!

c. The police *gave the child a stern talking to*.

Trousdale also shows that some uses can be ambiguous, as in the following example, where seeing to can mean either 'beating' or 'having sex':

(3) I'll give her a seeing to. (Trousdale 2008: 35)

Further, there are examples, where the object is subjected to an action involving physical effort or a procedure:

(4) a. Throw the potatoes in the pan, put the lid on and *give them a vigorous shaking*.

b. I offered to *give the tree a pruning*.

c. ...gave himself a brushing down in front of the mirror. (Norman Collins, *Love In Our Time*)

d. *Give the lawn a thorough soaking*.

e. She *gave the shirt a quick ironing*.

If there is anything these uses have in common, it is the sense that the object is affected by the action, which is a fairly general semantic element, one that is perfectly natural and typical of grammatical forms. Being subjected to an action and becoming affected as a result is a pervasive recurring theme that is the main semantic contribution of grammatical categories such as the accusative case (Dowty 1991; Levin & Rappaport Hovav 1993), resultative construction (Levin & Rappaport Hovav, 1993) or to take a less obvious case, the malefactive dative (Janda 1993; Wierzbicka 1988).

More seriously, even if the non-harm meanings in (51) can be dismissed as a handful of unproblematic exceptions (or extensions of the bodily harm prototype), it should be recognized that the meanings credited to the construction do not come from the construction itself, but from the lexical material inserted in the slots the construction leaves open. There is hardly anything surprising about open-class lexical items carrying contentful meanings. Claims of rich constructional meanings would be more compelling if the construction featured gerundive verbs that do not so much as implicate physical harm – this would constitute evidence that the bodily harm meanings in question were conveyed by the construction independently of the lexical material. One potential candidate of a non-harm gerund would be *seeing to*, as at first impression, it does not seem to be a synonym of *beating*. However, the meaning 'take care of' of *see to* can also be interpreted as a euphemistic expression of the intention to confront someone facing trouble. To lend some credence to their arguments, proponents of the 'bodily harm' meaning would need to demonstrate that the construction functions similarly to the double object construction (*I told her a joke; they sold us a car*), which expresses transfer of possession but which also allows non-transfer verbs (such as *build* or *bake*). Uses like *He built her a home* or *The child sang us a song* are interpreted as conveying transfer of possession by means of the syntactic pattern, not the verb. The only requirement that the verb needs to meet is general compatibility with the thematic core "X causes Z to have Y" (Pinker 1989: 82). To put it another way, the verb does not have to express the same meaning as the thematic core; it should simply not clash with the semantic structure of the construction. In the case of the *give*-gerund construction, 'bodily harm' is not its semantic structure.

2.3 The *way* construction

The *way* construction (*They clawed their way to freedom*) has been a mainstay of much cognitive theorizing and has also been featured in many influential constructionist discussions. The construction merits mention here because it is widely recognized as an example of a construction with a fairly rich semantic content. According to Goldberg, the construction codes motion taking place despite some obstacle or difficulty. She argues that the construction should carry the "presupposition that the motion was difficult in some way" and that the motion involved "the creation of a path" (2010: 53). In fact, a quick review of instances of the way construction confirms this characterization:

(5) a. What of those who struggled their way through the fierce winds...?

However, this reading is at best an implicature, and so cannot be the construction's stable contribution. A more thorough treatment is offered in Szczesniak (2013), but for our purposes here, it is quite easy to question the 'difficulty/obstacle' reading by demonstrating that many uses are attested that do not convey any sense of experienced difficulty:

(6) a. Attired in jugglers' costumes, the two ***frolicked*** their way to a splendid victory. (Spokane Daily Chronicle, April 24, 1978)

b. Inspiring gay athlete Blake Skjellerup has ***whizzed*** his way to the 2010 New Zealand Senior Speed Skating title. (http://www.gaynz.com/articles/publish/2/article_9355.php)

c. Schultz ***rollicked*** his way to the front of the stage, swinging his unruly mop of hair around like a young Eddie Vedder and hucking himself over the edge. (http://www.theblueindian.com/show-coverage/show-photos-videos/music-midtown-2011-a-retrospective/)

Implicatures are triggered by specific contexts, and as the above examples illustrate, it is perfectly possible for many uses not to trigger the implicature of 'motion in the face of an obstacle' at all.

3 Pragmatic Strengthening

One could go on reviewing grammatical constructions in this fashion and show each one to have less contentful semantics than is claimed in the literature. This, however, would be to dismiss only known cases, with the theoretical possibility being that there may exist yet undiscovered examples of constructions whose meanings may in fact contravene familiar kinds of semantic content found in closed-class forms. While it may never be possible to rule out such potential cases of semantically rich constructions, it is necessary to at least attempt to demonstrate that there are no mechanisms capable of endowing closed-class items with such meanings.

The only apparent possibility for richer meanings to actually occur in grammatical forms is to suppose that while some colorful meanings start out as conversational implicatures (as is the case with the examples above), they may eventually turn into entailments. Such developments have been known to occur through what is termed pragmatic strengthening (Traugott, 1988). Pragmatic strengthening, which can be viewed as the opposite of desemanticization, is a pervasive process observed in countless examples of forms whose meanings evolved from conversational implicatures to conventional implicatures. For example, the adverb *hwilum* 'at times' became the temporal connective *while*, which subsequently acquired the concessive function (Traugott 1988: 407). This was possible when the connective *while* was used to juxtapose two events standing in some logical opposition to each other. Because pragmatic strengthening seems to provide an open door to the theoretically impossible meanings becoming in fact possible at some future point, any claim that rules out excessive semantic capabilities in schematic constructions should contend with this challenge.

However, there are reasons to believe that pragmatic strengthening does not represent a problem for the present account. Studies on pragmatic strengthening report only two kinds of meanings that can emerge as a result of pragmatic strengthening. First, an item can acquire meanings that are otherwise familiar examples of grammatical meanings, such as tense reference. It has been pointed out that volitional verbs tend to take on future tense meanings. This is the case of the English *wyllan* ('want/wish') becoming the future tense auxiliary *will* (Bybee et al. 1994) or the Serbian and Croatian *hteti*/*htjeti* ('want') becoming the future tense marker (Corbett & Browne 2009), as in the follow-

ing example, where the verb clearly expresses future rather than intention.

(7) *Hoću li dugo čekati*?
Want1SG if long wait?
'Will I wait a long time?'

Second, an item may acquire non-truth-conditional functions involved in construal operations, as is the case of the concessive *while* (Traugott 1988), causal *since* (Molencki 2007), concessive *albeit* (Sorva, 2007), or the scalar *even* expressing a 'reversal of expectations' (Traugott 1988). These are markers that convey the speaker's attitude or perception of the proposition. Traugott sums up the tendency by observing that "[m]eanings tend to become increasingly situated in the speaker's subjective belief-state/attitude toward the situation". They can include "the speaker belief in the truth or probability of the proposition" or "some surprise factor on the speaker's part" (1988: 410).

These two kinds of meanings are precisely those that are commonly found in grammatical forms. In other words, the effects of pragmatic strengthening are hardly surprising. They represent meanings that can be predicted based on what we already know about the semantic content of grams. The sense of predictability is further enhanced by the cross-linguistic recurrence of the same pragmatic-strengthening motifs, whose range is by no means unlimited. As Bybee (2010: 171) notes, "inferences that are preferred in context are often very similar across cultures". For example, the evolution of future tense forms mentioned above is found to have occurred in other non-related languages too. Future tense originating from lexical items with volitional meanings has also evolved in Syrian Arabic, where the verbal noun *bi-wuddi* ('I want/desire') has developed into the *b*-prefix marking the future (Jarad 2013).

What the studies on pragmatic strengthening do not report are rich truth-conditional (non-construal) meanings like the 'difficulty' reading proposed for the *way* construction or the 'manipulation/mental coercion' reading ascribed to the *into*-gerund construction (*Jocelyn sweet-talked Kevin into buying her a chihuahua*.), much less their cross-linguistic attestations. Indeed, authors who champion pragmatic strengthening confine its scope to grammatical meanings. For example, Brinton and Traugott (2005: 68) state that "content is not enriched, but is 'bleached' (it gradually becomes backgrounded as grammatical meanings are enriched)." Thus, if pragmatic strengthening is incapable of infusing grammatical forms with richer meanings, there do not seem to exist any theoretical reasons to suppose that such meanings are in fact possible. In other words, pragmatic strengthening does not provide a means for endowing syntactic constructions with overly expansive meanings.

4 Conclusion

One could place a bold wager that no rich semantic or pragmatic effects proposed in constructionist analyses are true contributions of schematic grammatical constructions. As closed-class forms, schematic constructions are simply unable to convey more than what constructions have been traditionally known to convey. Although the Construction Grammar framework deserves the credit for drawing attention to the semantics of constructions, numerous semantic characterizations proposed within the framework are rather beyond belief, precisely because they are at odds with the implications of the lexicon-syntax distinction. The distinction, which has been de facto consigned to history, may still be very relevant to constructionist analyses.

Reference

Brinton, Luarel J. & Traugott, Elizabeth C., 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

Bybee, Joan, 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Bybee, Joan, Perkins, R. & Pagliuca, W., 1994. *The Evolution of Grammar. Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Corbett, Greville & Browne, Wayles, 2009. Serbo-Croat. In: B. Comrie, ed. *The World's Major Languages*. New York: Routledge, pp. 330-346.

Dowty, David, 1991. Thematic proto-roles and argument selection. *Language,* Volume 67, p. 547–619.

Goldberg, Adele, 2006. *Constructions At Work: The Nature of Generalization in Language*. Oxford: Oxford University Press..

Hilpert, Martin, 2008. *Germanic Future Constructions. A usage-based approach to language change*. Amsterdam: John Benjamins.

Janda, Laura, 1993. *A Geography of Case Semantics: The Czech Dative and the Russian Instrumental*. Berlin: Mouton.

Jarad, Najib I., 2013. The Evolution of the b-Future Marker in Syrian Arabic. *Lingua Posnaniensis,* LV(1), p. 69–85.

Kay, Paul & Michaelis, Laura A., to appear. Constructional Meaning and Compositionality. In: *Semantics: An International Handbook of Natural Language Meaning: Volume 2*. Berlin: Mouton de Gruyter.

Kearns, Kate, 2002. [1988] Light verbs in English. Ms.

Langacker, Ronald W., 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites, Volume 1*. Stanford, CA: Stanford University Press.

Levin, Beth & Rappaport Hovav, Malka, 1993. *Unaccusativity At the Syntax-Lexical Semantics Interface*. Cambridge, MA: MIT Press.

Molencki, Rafał, 2007. The evolution of since in medieval English. In: U. Lenker & A. Meurman-Solin, eds. *Connectives in the History of English*. Amsterdam: John Benjamins, pp. 97-114.

Pinker, Steven, 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.

Sorva, Elina, 2007. Grammaticalization and syntactic polyfunctionality. The case of albeit. In: U. Lenker & A. Meurman-Solin, eds. *Connectives in the History of English*. Amsterdam: John Benjamins, pp. 115-144.

Szcześniak, Konrad, 2013. You can't cry your way to candy: Motion events and paths in the x's way construction. *Cognitive Linguistics,* 24(1), p. 159–194.

Traugott, Elizabeth C., 1988. Pragmatic Strengthening and Grammaticalization. *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pp. 406-416.

Trousdale, Graeme, 2008. Constructions in grammaticalization and lexicalization: Evidence from the history of a composite predicate construction in English. In: G. Trousdale & N. Gisborne, eds. *Constructional Approaches to English Grammar*. Berlin: Mouton de Gruyter , pp. 33-70.

Wierzbicka, Anna, 1988. *The Semantics of Grammar*. Amsterdam: John Benjamins.

Wierzbicka, Anna, 2006. *English. Meaning and Culture*. Oxford: Oxford University Press.

Verb-Noun Collocations in PolNet 2.0

Zygmunt Vetulani
Adam Mickiewicz University
Fac. of Mathematics and Computer Science
ul. Umultowska 87
61-614 Poznań, Poland
vetulani@amu.edu.pl

Grażyna Vetulani
Adam Mickiewicz University
Faculty of Modern Languages
al. Niepodległości 4
61-874 Poznań, Poland
gravet@amu.edu.pl

Abstract

In this paper we present works contributing to transformation of PolNet, a Polish wordnet developed at the Adam Mickiewicz University in Poznań, into a Lexicon Grammar of Polish. The current step consists in inclusion of verb-noun collocations and relations linking the verbal synsets to noun synsets.

1 Credits

This work was done within the Polish National Program for Humanities and was covered by the grant 0022/FNiTP/H11/80/2011.

2 Introduction

The "PolNet-Polish WordNet" project started at the Adam Mickiewicz in 2006 with the objective to produce a lexicographical database inspired by the Princeton Wordnet (Miller et al., 1990). In the Princeton WordNet (and other similar systems) the basic entities are *synsets*, i.e. classes of synonyms related by some relations.[1] The main, organizing, relations between synsets are hyponymy and hyperonymy. PolNet was built from scratch within the so called "merge development model" (development algorithm was published in 2007 (Vetulani, Z. et al., 2007). In December 2011/January 2012 we have released the first public version of the main deliverable of the project "PolNet - Polish Wordnet". This first release is freely distributed as PolNet 1.0 under a CC license.[2] The noun part of the PolNet 1.0 consisted of the noun synsets partially ordered by the hyponymy/hyperonymy relation and the verb part organized by the predicate-argument structures relating the verb synsets with the noun synsets. In the present extension (from PolNet 1.0 to PolNet 2.0) we continue to use this organization.[3]

The present development from PolNet 1.0 to PolNet 2.0 consists in completing several gaps and eliminating bugs, but first of all in extending substantially the verbal component with the inclusion of concepts (synsets) represented (in many cases uniquely) by compound construction in form of verb-noun collocations[4]. This extension brings to PolNet some 2000 new synsets for the most important verb-nouns collocations (corresponding to 400 predicative nouns), some of which closely related to the already existing verb synsets of the PolNet 1.0.

3 Verb synsets in PolNet 1.0

Rather than translating the Princeton WordNet (Miller et al., 1990) according to the so called "expand model" we decided to develop PolNet following the so called "merge model" in order to limit the non-controlled and non-desired import of conceptualization reflected in English.[5] The team - formed of computer scientists and

[1] For the fundamental concept of synset ("a class of synonyms") cf e.g. (Miller et al., 1990) or (Vossen et al., 1998).

[2] Accessible through www.ltc.amu.edu.pl (follow the link to LTC 2011).

[3] The main statistics about PolNet 1.0 are as follows:
Nouns: 11,700 synsets (20,300 word+meaning pairs, 12,000 nouns)
Verbs: 1,500 synsets (2,900 word+meaning pairs, 900 verbs)

[4] By *verb-noun collocations* we mean compound verbal structures made of a support verb and a predicative noun.

[5] Such methodological choice was possible because Polish is rich in high quality dictionaries what makes the work "from scratch" feasible.

lexicographers - explored first of all traditional resources (dictionaries).[6] This work was inspired by and benefited from the methodology and tools of the EuroWordNet and Balkanet projects (DebVisDic systems generously was made accessible by Karel Pala /Masaryk University in Brno)[7].

The release PolNet 1.0 consisted of noun and verb synsets. Selection of words to both parts, nominal and verbal, was done on the basis of word frequencies observed in the disposable corpus (IPI PAN corpus, cf. Przepiórkowski (2004)). In the present development we continue to use the frequency-based methods, however with use of new resources, mainly ackwired from online sources and web crawling (the IPI PAN corpus appeared no longer sufficient for our purposes).

Initially, PolNet was conceived for nouns only. However, by the end of the first phase of the project some amount of verb synsets were included as well. With inclusion of verb synsets we brought to PolNet the ideas inspired by the FrameNet (Fillmore et al., 2002) and the VerbNet (Palmer, 2009) projects, as well as by the works of Polański (1992) and Gross (1994). The verbal part becames the backbone of the whole network in PolNet 2.0 (its organizing part is the system of semantic roles).

Our decision to include possibly all relevant grammatical information connected with verbs appeared to be a challenging task whose significance consisted in transforming the initial lexical ontology [8] in a lexicon-grammar [9] organized by relations attached to verb synsets. The role of these relations was to encode the way in which verbs and nouns combine in simple sentences. This work, pioneering for Polish, was done (in PolNet 1.0) in a relatively short time due to the high quality description of Polish verbs from "Syntactic-generative Dictionary of Polish Verbs" (Polański, 1992) /in five volumes, published between 1980 and 1992/. The formalism proposed by Polański is not machine-readable and the content required substantial human preprocessing before its integration to PolNet. The task appeared not trivial at all and required important engineering decisions.

In PolNet, as in other wordnets, lexical units are grouped into synsets on the basis of the relation of synonymy. In opposition to nouns, where the interest is mainly in the hierarchical relations (hyperonymy/hyponymy) between concepts, for verbs the main interest is in relating *verb synsets* (representing predicative concepts) to *noun synsets* (representing general concepts) in order to show what the semantic/morpho-syntactic connectivity constraints corresponding to the respective argument positions are. Inclusion of this information gives to PolNet the status of a lexicon grammar. This approach imposes strong granularity restrictions on synonymy of verbal synsets referring to the concept of valency structure. By *valency structure* we mean in this paper *the structured information on the arguments opened by the predicative word including both semantic constraints on the arguments as well as the surface morpho-syntactic properties of the text fillers of argument positions (like case, numer, gender, preposition etc.)*[10]. **Synonymous are solely such verb/meaning units in which corresponding *semantic roles* take the same values (this condition is necessary but not sufficient for verb synonymy).** In particular, the valency structure of a verb is one of its formal indices of the meaning (all members of a sysnset share the valency structure). This permits to consider valency structure as a property of a synset.

3.1 PolNet as a lexical ontology

The extended PolNet may be considered as a situational-semantic network of concepts (represented by synsets). Indeed, as it is often admitted, verb synsets may be considered as representing situations (events, states), whereas semantic roles (Agent, Patient, Beneficent,...) provide information on the ontological nature of various actors participating, actively or passively, in these situations (events, states). Abstract roles as Manner and Time refer to concepts which position the situation (event, state) in time, space and possibly also with respect to some abstract, qualitative landmarks. The abstract role Cause indicates the reason for some situation (event, state). Formally, the semantic roles are functions (in mathematical

[6] The ambitious, large scale project plWordnet developed at the Wrocław Technical University follows different methodology.

[7] Cf. (Pala et al., 2007).

[8] Cf. (Gangemi, Navigli and Velardi, 2003).

[9] The concept of lexicon-grammar was first developed for French by Maurice Gross (lexique-grammaire) in the 1970s; cf. (Gross, 1994).

[10] Some authors do not take into account the morpho-syntactic layer. We do.

sense) associated to the argument positions in the syntactic pattern(s). Values of these functions are ontology concepts represented by synsets (where possible) or by appropriate formal ontology concepts (here adapted from the EuroWordNet top ontology (Vossen et al., 1998) and the SUMO upper ontology (Pease, 2011)).

3.2 TWO EXAMPLES (different valency structures for two meanings of the verb "szanować")

3.2.1. The example below (simplified /no semantic information/) presents a fragment of the code with the valency information for one of the senses of the verb "szanować" (close to English "respect"). The "frame" lines describe the valency structure in terms of semantic roles and morpho-syntactic features. An important formal constraint is that any two "frame" lines should be compatible, i.e. a role must hold the same formal parameters in all "frame" lines.

```
<VALENCY> /*fragment*/
<FRAME>Agent(N) _ Benef(Acc)</FRAME>
/*"Andrew szanuje/respects/ Sonię"*/
<FRAME>Agent(N) _ Benef(Acc)
Cause('za'+Acc)</FRAME> /*"Andrew szanuje
Sonię za mądrość"*/
<FRAME>Agent(N) _ Object(Acc)
Cause('za'+Acc)</FRAME> /* Człowiek szanuje
prawo za ład i porządek*/
</VALENCY>
```

Fig. 1. A fragment of the PolNet 1.0 code with the valency information (one of meanings of "szanować").

The element of the code "Agent(N)" says that the agent position (subject) requires a form in Nominative. Similarly "Acc" stands for Accusative and "'za'+Acc" for a form in Accusative preceded by the preposition 'za'.

3.2.2. The following code corresponds to another meaning of "szanować" (close to English "protect")

```
<VALENCY>
<FRAME>Agent(N)_ Object(Acc)</FRAME>
/*Mądry człowiek szanuje/protect/ przyrodę */
</VALENCY>
```

Fig. 2. A fragment of the PolNet 1.0 code with the valency information (another meaning of "szanować").

The above two distinct meanings of "szanować" are represented by different synsets.

4 Granularity problem for verb synsets. Collocations in PolNet 2.0

The problem of granularity of the verbnet part of PolNet re-appears as an important theoretical issue at the present passage from PolNet 1.0 to PolNet 2.0, marked first of all by inclusion of verb-noun collocations. The verb-noun collocations included into PolNet were taken from the "Syntactic dictionary of verb-noun collocations in Polish" (Vetulani, G. 2000 and 2012). Adding verb-noun collocations to PolNet was non-trivial because of specific phenomena related with collocations in Polish.

The challenging issue of verb synsets granularity is closely connected with synonymy which is fundamental for the concept of wordnet. Let us notice the fact, well described in the literature, that while there is consensus through the wordnet community concerning the principle that *synonymy is the basis of organization of the (wordnet) database in synsets*, (i.e. synonyms should belong to the same synsets), there is no consensus among linguists on the concept of synonymy. The spectrum of solutions is large, starting with the very restrictive definition proposed by Leibnitz (based on *truth value invariability* – leading to small synsets, often containing just one word), up to the solution applied by Miller and Fellbaum /in Vossen, 1998/) which postulates a very weak understanding of this concepts (based on *invariability test with respect to just one linguistic context* – often leading to very large synsets).

The application of the above principle has important consequences. This is because the morpho-syntactic requirements of predicative words are not invariant with respect to the traditional relation of synonymy. For example the simple word "respektować" and its equivalent in form of the collocation "mieć respekt" (both corresponding to one of the meanings of "to respect" in English), do not have the same morpho-syntactic requirements. It is so because "respektować" requires the direct object in accusative, whereas "mieć respekt" - in genitive and preceded by the preposition "dla".

Within the concept of synonymy used in the PolNet project, *respektować* and *mieć respect,* should be put into different synsets of PolNet

because they do not share the valency structure and the synset of PolNet are supposed to contain complete syntactic and semantic information about words, the same for all synset members.[11] Having the valency structure directly stored with the synset (as a part of its description) is an important advantage when (e.g.) one uses PolNet as the grammatical database to be directly consulted by parsers and other language processing software[12].

5 Solution

The above problem, typical of highly inflected languages, appears when in order to paraphrase a sentence with a single-word predicate we replace it by a collocation what often implies change of the grammatical case of an argument. Such a case is non-existent in English and other low inflectional languages.

In PolNet 2.0 we have applied a solution which seems optimal from the practical (language engineering) point of view. This consists to store collocations and their corresponding single word equivalents in separate synsets if their valency structures are different. These synsets will be related by the transformational relation which describes the difference of their morpho-syntactic properties.

Let us consider "upoważnić" ("to authorize") vs "dać pełnomocnictwo" ("to give authorisation") in the sentences: ("*Piotr upoważnił adwokata(Acc) do zakupu*" vs. ("*Piotr dał pełnomocnctwo do zakupu adwokatowi*" *(Dat)*

The Fig. 3. below presents the valency structures /simplified/ for the verb "upoważnić" ("to authorize") in opposition to the valency structure for "dać pełnomocnictwo" ("to give authorisation"). We observe the grammatical case transformation of the direct object between both considered sentences.

"*Piotr upoważnił* ***adwokata(Acc)*** *do zakupu*"

<VALENCY>
<FRAME>Agent(N) _ **Patient(Acc)** Purpose('do'+G)
</FRAME>
</VALENCY>

"*Piotr dał pełnomocnctwo do zakupu* ***adwokatowi(D)***"
<VALENCY>
<FRAME>Agent(N) _ **Patient(D)** Purpose('do'+G)
</FRAME>
</VALENCY>

Fig. 3. Case transformation of the Patient

In PolNet 2.0 we store the verb "upoważnić" and the collocation "dać pełnomocnictwo" in different synsets related by an external (inter-synset) relation describing the direct object case transformation necessary for paraphrasing (TRANS_CASE_PATIENT(A,D)). At the same time, collocations like "dać pełnomocnictwo", "udzielić pełnomocnictwa", "nadać pełnomocnictwo" (and some other) will belong to the same synset.

6 Future work

We are now developing, cleaning and extending the PolNet system. The public release of PolNet 2.0 including over 7000 new collocations is scheduled for the end of 2014.

Reference

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002. The FrameNet Database and Software Tools. In: *Proceedings of the Third International Conference on Language Resources and Evaluation. Vol. IV*. LREC: Las Palmas.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in

[11] The concept of verb synset of PolNet 1.0 is presented as follows in (Vetulani&Vetulani, 2013): "In opposition to nouns, where the focus is on the relations between concepts (represented by synsets), and in particular on hiperonymy/hyponymy relations, for verbs the main interest is in relating verbal synsets (representing predicative concepts) to noun synsets (representing general concepts) in order to show what are connectivity constraints corresponding to the particular argument positions. This approach imposes granularity restrictions on verbal synsets and more precisely on the synonymy relation. Synonymous will be only such verb+meaning pairs in which the same *semantic roles* take as value the same concepts (this condition is necessary but not sufficient). In particular, the valency structure of a verb is one of formal indices of the meaning (so, all members of a given sysnset share the valency structure). This approach permits to formally encode valency structure as a property of a synset."

[12] This is a capital argument in favor of the Lexicon Grammar approach. We did positively verified this solution in the implementations of parsers of the POLINT family (e.g. POLINT-112-SMS) (Vetulani & Marciniak, 2011) where easy access to valency information permitted to define simple heuristics to speed-up parsing through smart reduction of search space.

WordNet (PDF). *Proc. of International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE 2003)* (Catania, Sicily (Italy)). pp. 820–838.

Maurice Gross. 1994. Constructing Lexicon-Grammars. In: Beryl T. Sue Atkins, Antonio Zampolli (eds.) *Computational Approaches to the Lexicon*, Oxford University Press. Oxford, UK, pp. 213–263.

George A. Miller, Richard Beckwith, Christiane D. Fellbaum, Derek Gross, and Katherine J. Miller. 1990. WordNet: An online lexical database. *Int. J. Lexicograph*. 3, 4, pp. 235–244.

Karel Pala, Aleš Horák, Adam Rambousek, Zygmunt Vetulani, Paweł Konieczka, Jacek Marciniak, Tomasz Obrębski, Paweł Rzepecki, and Justyna Walkowska. 2007. DEB Platform tools for effective development of WordNets in application to PolNet, in: Zygmunt Vetulani (ed.) *Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2005, Poznań, Poland*, Wyd. Poznańskie, Poznań, pp. 514-518.

Martha Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*. Sept. 2009, Pisa, Italy: GenLex.

Adam Pease. 2011. *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA. ISBN 978-1-889455-10-5.11

Kazimierz Polański (ed.). 1992. *Słownik syntaktyczno - generatywny czasowników polskich* vol. I-IV, Ossolineum, Wrocław,1980-1990, vol. V, Kraków: Instytut Języka Polskiego PAN.

Adam Przepiórkowski. 2004. *Korpus IPI PAN. Wersja wstępna / The IPI PAN CORPUS: Preliminary version*. IPI PAN, Warszawa.

Grażyna Vetulani. 2000. *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych*. (In Polish). Poznań: Wyd. Nauk. UAM.

Grażyna Vetulani. 2012. *Kolokacje werbo-nominalne jako samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I.* (In Polish). Poznań: Wyd. Nauk. UAM.

Zygmunt Vetulani. 2012. Wordnet Based Lexicon Grammar for Polish. *Proceedings of the Eith International Conference on Language Resources and Evaluation (LREC 2012)*, May 23-25, 2012. Istanbul, Turkey, (Proceedings), ELRA: Paris. ISBN 978-2-9517408-7-7, pp. 1645-1649.

Zygmunt Vetulani and Jacek Marciniak. 2011. Natural Language Based Communication between Human Users and the Emergency Center: POLINT-112-SMS. In: Zygmunt Vetulani (ed.): *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2009. Revised Selected Papers*. LNAI 6562, Springer-Verlag Berlin Heidelberg, str. 303-314.

Zygmunt Vetulani and Grażyna Vetulani. (2013): Through Wordnet to Lexicon Grammar, in: Fryni Kakoyianni Doa (Ed.). *Penser le lexique-grammaire : perspectives actuelles*, Editions Honoré Champion, Paris, France, 531-545.

Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Paweł Konieczka, Paweł Rzepecki, and Jacek Marciniak. 2007. PolNet - Polish WordNet project algorithm, in: Zygmunt Vetulani (ed.) *Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2005, Poznań, Poland,* Wyd. Poznańskie, Poznań, pp. 172-176.

Piek Vossen, Laura Bloksma, Horacio Rodríguez, Salvador Climent, Nicoletta Calzolari, and Wim Peters. 1998. The EuroWordNet Base Concepts and Top Ontology, Version 2, Final, January 22, 1998 (Euro WordNet project report) (http://www.vossen.info/docs/1998/D017.pdf; access 22/04/2011). (also: Piek Vossen (et al.) (2003): *EuroWordNet General Document, Version 3. Final, July 22*).