Proceedings of the LREC Workshop

Towards a Shared Task for Multiword Expressions (MWE 2008)

Marrakech, Morocco

1 june 2008

Programme

09.15 - 09.30	Opening
09.30 - 10.30	Resource session I
	A Resource for Evaluating the Deep Lexical Acquisition of English Verb-Particle Constructions *Timothy Baldwin*
	A Lexicographic Evaluation of German Adjective-Noun Collocations *Stefan Evert*
	Description of evaluation resource -- German PP-verb data *Brigitte Krenn*
	Reference Data for Czech Collocation Extraction *Pavel Pecina*
10.30 - 11.00	Coffee break
11.00 - 13.30	Resource session II
	A Lexicon of shallow-typed German-English MW-Expressions and a German Corpus of MW-Expressions annotated Sentences *Dimitra Anastasiou and Michael Carl*
	The VNC-Tokens Dataset *Paul Cook, Afsaneh Fazly and Suzanne Stevenson*
	Multi-Word Verbs of Estonian: a Database and a Corpus *Heiki-Jaan Kaalep and Kadri Muischnek*
	A French Corpus Annotated for Multiword Nouns *Eric Laporte, Takuya Nakamura and Stavroula Voyatzi*
	An Electronic Dictionary of French Multiword Adverbs *Eric Laporte and Stavroula Voyatzi*
	Cranberry Expressions in English and in German *Beata Trawinski, Manfred Sailer, Jan-Philipp Soehn, Lothar Lemnitzer and Frank Richter*
	Standardised Evaluation of English Noun Compound Interpretation *Su Nam Kim and Timothy Baldwin*
	Interpreting Compound Nominalisations *Jeremy Nicholson and Timothy Baldwin*
	Paraphrasing Verbs for Noun Compound Interpretation *Preslav Nakov*
13.30 - 14.30	Lunch break
14.30 - 14.50	Introduction to shared task and baseline results
14.50 - 15.40	Shared task participants
	An Evaluation of Methods for the Extraction of Multiword Expressions *Carlos Ramisch, Paulo Schreiner, Marco Idiart and Aline Villavicencio*
	A Machine Learning Approach to Multiword Expression Extraction *Pavel Pecina*
15.40 - 16.00	Thoughts on the first MWEVAL
16.00 - 16.30	Coffee break
16.30 - 17.30	Interactive group discussion
17.30 -	Workshop summary and General discussion

Organisers

Nicole Grégoire, *University of Utrecht (The Netherlands)*

Stefan Evert, *University of Osnabrück (Germany)*

Brigitte Krenn, *Austrian Research Institute for Artificial Intelligence (ÖFAI) (Austria)*

Programme Committee

Iñaki Alegria, *University of the Basque Country (Spain)*

Timothy Baldwin, *Stanford University (USA); University of Melbourne (Australia)*

Colin Bannard, *Max Planck Institute (Germany)*

Francis Bond, *NTT Communication Science Laboratories (Japan)*

Gaël Dias, *Beira Interior University (Portugal)*

Kyo Kageura, *University of Tokyo (Japan)*

Rosamund Moon, *University of Birmingham (UK)*

Diana McCarthy, *University of Sussex (UK)*

Eric Laporte, *University of Marne-la-Vallee (France)*

Preslav Nakov, *University of California, Berkeley (USA)*

Jan Odijk, *University of Utrecht (The Netherlands)*

Stephan Oepen, *Stanford University (USA); University of Oslo (Norway)*

Darren Pearce, *University of Sussex (UK)*

Pavel Pecina, *Charles University (Czech Republic)*

Scott Piao, *University of Manchester (UK)*

Violeta Seretan, *University of Geneva (Switzerland)*

Suzanne Stevenson, *University of Toronto (Canada)*

Beata Trawinski, *University of Tuebingen (Germany)*

Kiyoko Uchiyama, *Keio University (Japan)*

Begoña Villada Moirón, *University of Groningen (The Netherlands)*

Aline Villavicencio, *Federal University of Rio Grande do Sul (Brazil)*

Table of Contents

A Resource for Evaluating the Deep Lexical Acquisition of English Verb-Particle Constructions

Timothy Baldwin

Department of Computer Science and Software Engineering
University of Melbourne
Victoria 3010 Australia

tim@csse.unimelb.edu.au

Abstract

This paper describes a dataset which provides the platform for a task on the extraction of English verb particle constructions with basic valence information.

1. Introduction

With growing interest in multiword expressions (MWEs: Sag et al. (2002)) and the extraction of different types of MWE (Evert and Krenn, 2001; Baldwin, 2005a; Baldwin, 2005b; van der Beek, 2003), there is an increasing need for standardised datasets across which to compare different techniques. This paper presents a dataset intended for use in evaluating the extraction of English verb particle constructions with basic subcategorisation information.

2. Task Definition

English **verb-particle constructions**, or VPCs, consist of a head verb and one or more obligatory **particles**, in the form of intransitive prepositions (e.g. *hand in*), adjectives (e.g. *cut short*) or verbs (e.g. *let go*) (Villavicencio and Copestake, 2002; Huddleston and Pullum, 2002); for the purposes of the dataset, we assume that all particles are prepositional—by far the most common and productive of the three types—and further restrict our attention to single-particle VPCs (i.e. we ignore VPCs such as *get along together*).

We distinguish between compositional and non-compositional VPCs in this research, and restrict our attention exclusively to non-compositional VPCs. With compositional VPCs, the semantics of the verb and particle both correspond to the semantics of the respective simplex words, including the possibility of the semantics being specific to the VPC construction in the case of particles. For example, *battle on* would be classified as compositional, as the semantics of *battle* is identical to that for the simplex verb, and the semantics of *on* corresponds to the continuative sense of the word as occurs productively in VPCs (c.f. *walk/dance/drive/govern/... on*). Note that this distinction was not made in the original Baldwin (2005a) research that provides the foundation for this dataset.

English VPCs can occur in a number of subcategorisation frames, with the two most prevalent and productive valences being the simple transitive (e.g. *hand in the paper*) and intransitive (e.g. *back off*). For the purposes of this dataset, we focus exclusively on these two valence types.

Given the above, we define the English VPC extraction task to be the production of triples of the form $\langle V, P, S \rangle$, where V is a verb lemma, P is a prepositional particle, and $S \in \{intrans, trans\}$ is the valence; additionally, each triple has to be semantically non-compositional. The triples are generated relative to a set of putative token instances for each of the intransitive and transitive valences for a given VPC. That is, a given triple should be classified as positive iff it is associated with at least one non-compositional token instance in the provided token-level data.

3. Corpora and Annotation

The dataset is based on the research of Baldwin (2005a), where VPC token instances were variously identified in the written portion of the British National Corpus (BNC: Burnard (2000)) by tagger-, chunker-, chunk grammar-, and parser-based extraction methods. These were then combined together to form a type-level hypothesis, including a prediction of the valence of the VPC. We used the Baldwin (2005a) method to identify 2,898 novel VPC types[1] which were associated with one or more high-confidence token instances (as identified by weighted voting across the predictions of the individual extraction techniques) for the intransitive and/or transitive valences. We presented the token instances to an annotator and asked them to filter out any compositional VPCs. These annotations were then vetted by a second annotator for final inclusion in the lexicon of the English Resource Grammar, an implemented HPSG under development at CSLI (Flickinger, 2002).

In the dataset, we have provided a single file for each of 4,090 candidate VPC triples (corresponding to 2,898 unique VPCs), containing up to 50 sentences containing the given VPC. All sentences are tokenised according to the Penn Treebank format, and cardinal and ordinal numbers are additionally normalised to the tokens `--CNUMB--` and `--ONUMB--`, respectively. Evaluation is relative to the files `intrans.gold` and `trans.gold`, containing the gold-standard sets of intransitive and transitive VPC triples, respectively.

4. Summary

In this paper, we have presented a dataset for standardised evaluation of English VPC extraction with basic valence information.

[1] Outside the original 1,000 VPC types targeted in the original research.

5. References

Timothy Baldwin. 2005a. The deep lexical acquisition of English verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Timothy Baldwin. 2005b. Looking for prepositional verbs in corpus data. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 180–9, Colchester, UK.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proc. of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, pages 188–95, Toulouse, France.

Dan Flickinger. 2002. On building a more efficient grammar by exploiting types. In Stephan Oepen, Dan Flickinger, Jun'ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering*. CSLI Publications, Stanford, USA.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, pages 1–15. Springer-Verlag, Hiedelberg/Berlin, Germany.

Leonoor van der Beek. 2003. The extraction of Dutch determinerless PPs. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, UK.

Aline Villavicencio and Ann Copestake. 2002. Verb-particle constructions in a computational grammar of English. In *Proc. of the 9th International Conference on Head-Driven Phrase Structure Grammar (HPSG-2002)*, Seoul, Korea.

A Lexicographic Evaluation of German Adjective-Noun Collocations

Stefan Evert

Institute of Cognitive Science, University of Osnabrück
49069 Osnabrück, Germany
stefan.evert@uos.de

Abstract

This paper describes a small database of 1,252 German adjective-noun combinations, which have been annotated by professional lexicographers with respect to their collocational status and their usefulness for the compilation of a bilingual dictionary. The database is a random sample taken from the most frequent ($f \geq 20$) adjective-noun pairs in a standard newspaper corpus (*Frankfurter Rundschau*). It is particularly useful for the evaluation and development of ranking techniques for multiword candidates. Suitable corpus frequency data (instances of adjective-noun cooccurrences from the same corpus) are made available together with the database.

1. Introduction and background

The work presented here was motivated by two comparative studies that evaluated the usefulness of different association measures for the identification of German adjective-noun collocations (Lezius, 1999; Evert et al., 2000).[1] Both studies seemed to confirm results from previous comparative evaluations carried out for other languages and other types of collocations, e.g. Daille (1994) and Krenn (2000). In particular, the following observations were made:

1. The most useful measure for collocation identification is *log-likelihood* (Dunning, 1993), justifying its well-established role as a default association measure in computational linguistics.
2. Log-likelihood is significantly better than the *chi-squared* measure (even if Yates' continuity correction is applied), as has been claimed by Dunning (1993).
3. A simple ranking of candidates by their cooccurrence *frequency* achieves surprisingly good results, although precision is significantly lower than for log-likelihood.
4. Contrary to the claims of Church and Hanks (1990), *Mutual Information* (MI) is very poorly suited for collocation identification.
5. Many other association measures (including *t-score*) are very close to log-likelihood, but none of them achieves significantly better results for any n-best list of candidates. This observation led to the hypothesis that log-likelihood represents an upper limit for collocation identification methods based on cooccurrence frequency data (the "sonic barrier" hypothesis).

However, both studies also had considerable shortcomings, so that these findings have to be qualified. The most serious problems, which motivated the follow-up study described in this paper, are the following:

- Lezius (1999) only looked at short 100-best lists of candidates, and many of the observed differences are not significant.[2] It is also not clear whether the results can be generalised to practically relevant 1000-best or 2000-best lists.
- Evert et al. (2000) aimed at a complete manual annotation of all recurrent adjective-noun combinations (with $f \geq 2$) in a given corpus, so that recall and baseline precision can be computed. For practical reasons, they chose an unrealistically small corpus of German law texts (approx. 800,000 running words). Real-life applications are likely to use much larger corpora and higher frequency thresholds, which may favour association measures like chi-squared and MI that are oversensitive to low-frequency data.
- Both studies failed to give a precise definition of collocations and did not supply clear guidelines to annotators. As a consequence, inter-annotator agreement was very low (though not reported in the original publications) and it was often impossible to resolve differences by discussion. This raises considerable doubt as to which aspects of the interplay between collocativity and statistical association have been evaluated, and whether a comparison with other studies is meaningful at all.

For these reasons, a follow-up study was designed in order to verify the findings of Lezius (1999) and Evert et al. (2000). The new study was based on a 40-million-word newspaper corpus (*Frankfurter Rundschau*), and candidate collocations were examined by professional lexicographers. This approach ensures a consistent and practically relevant definition of collocations and enables a direct comparison with other studies based on lexicographic (Smadja, 1993) or terminological (Daille, 1994) expert judgements.
The remainder of this paper is organised as follows. Section 2. summarises the initial results of this follow-up study, which have not been published before and provide an important reference point for future experiments with the database. The new adjective-noun database is described in Section 3., while Section 4. documents the file format and availability of the resource.

[1] Following the terminology of the cited studies, we understand *collocations* as a fuzzy concept that encompasses lexicalised, partly lexicalised and other "habitual" word combinations. It is similar in meaning to the current usage of the term *multiword expressions*, but may also include conventionalised word combinations even if they do not show the typical linguistic hallmarks of lexicalisation, i.e. non-compositionality, non-substitutability and non-modifiability (Manning and Schütze, 1999, 184).

[2] In the original study, no significance tests were carried out.

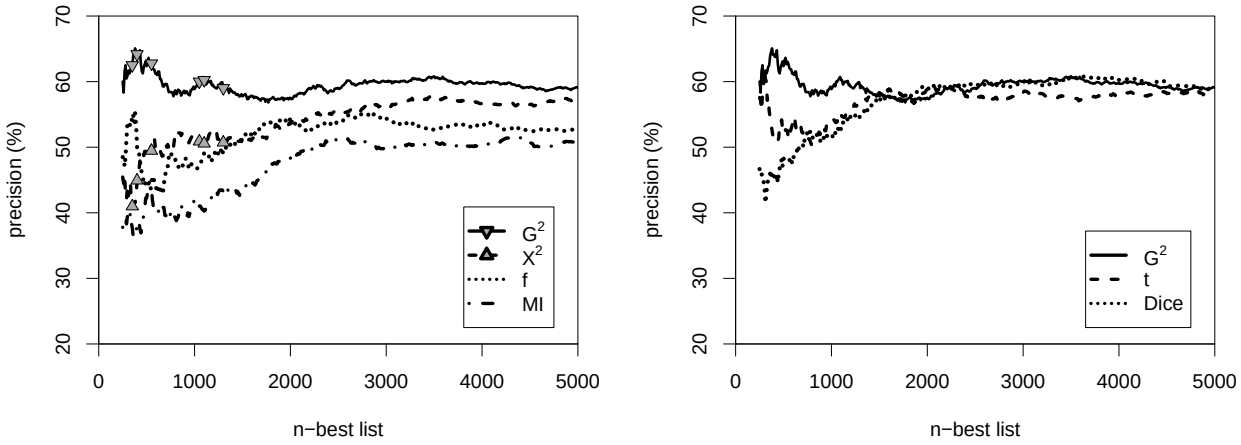


Figure 1: Evaluation results for the identification of adjective-noun collocations for lexicographic purposes. True positives are all word pairs that are considered useful for the compilation of a bilingual dictionary. The following association measures have been evaluated: log-likelihood (G^2), chi-squared with Yates' correction (X^2), Mutual Information (MI), t-score (t), Dice coefficient (Dice) and frequency ranking (f). Grey triangles indicate significant differences between log-likelihood and chi-squared ($\alpha = .05$).

2. The original experiment

For the follow-up experiment, German adjective-noun combinations were extracted from the *Frankfurter Rundschau* corpus (a detailed description of the corpus and extraction procedure is given in Section 3.). After application of a frequency threshold ($f \geq 5$), 5,000-best lists of collocation candidates were prepared according to 7 standard association measures. These measures were selected in order to verify observations made by previous studies. They include log-likelihood, chi-squared, t-score, MI, as well as the Dice coefficient. See Evert (2004, Ch. 3) or `http://www.collocations.de/` for full descriptions of all relevant association measures.

The ranked collocation candidates were manually evaluated by professional lexicographers with respect to their usefulness for the compilation of a bilingual (German-English) dictionary. Since annotation of all 13,533 candidates in the pooled n-best lists would have been prohibitively time-consuming, evaluation was based on a 15% random sample, using the RSE methodology of Evert and Krenn (2005).

The initial results were in accordance with previous studies, as the precision graphs in Figure 1 show (see Evert and Krenn (2005) or Evert (2004) for a detailed explanation of such evaluation graphs). Log-likelihood is the best association measure for this task (left panel). It is significantly better than chi-squared, at least for n-best lists up to $n = 1500$. Frequency ranking performs surprisingly well, but has significantly lower precision than log-likelihood, and MI is worse than frequency ranking. The right panel shows a clear "sonic barrier" effect: for $n \geq 1500$, log-likelihood, t-score and Dice have virtually indistinguishable performance, despite their entirely different mathematical properties.

In summary, this experiment seemed to confirm the results of Lezius (1999) and Evert et al. (2000). The only surprising observation was that log-likelihood achieves almost constant precision ($\approx 60\%$) for all n-best lists. While it is apparently very useful for selecting a large set of 5,000 promising candidates (on par with t-score and Dice), it does not seem to be able to make any further distinctions between these candidates. At the time, this was interpreted as supporting evidence for the "sonic barrier" hypothesis.

One possible explanation for the nearly constant precision of log-likelihood was the fact that the evaluation criterion of "usefulness for dictionary compilation" mixes entirely different types of collocations, ranging from non-compositional multiword expressions to regularly formed, but frequent combinations (which might provide good material for usage examples in the dictionary). In a second evaluation, true positives were therefore restricted to "true" collocations, which are at least partly lexicalised and need to be listed in the dictionary (if only for contrastive reasons). The results were entirely surprising, as the precision curves in Figure 2 show.

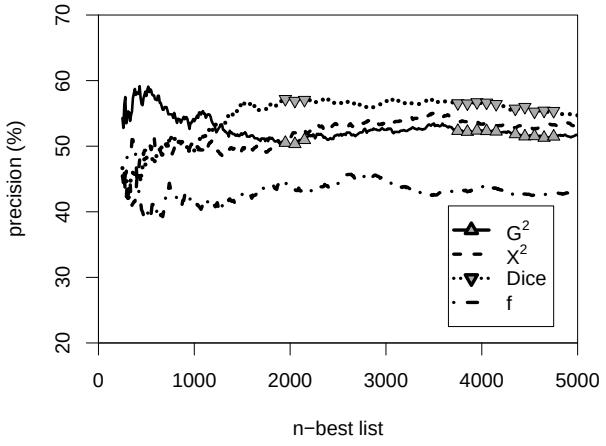


Figure 2: Evaluation results for the identification of "true" adjective-noun collocations, which need to be listed in a bilingual (German-English) dictionary. Grey triangles indicate significant differences between Dice and log-likelihood ($\alpha = .05$).

The precision achieved by log-likelihood is somewhat lower than before, but still almost constant across all n-best lists. Chi-squared is less affected by the modified evaluation criterion and is even slightly better than log-likelihood for $n \geq 2000$, contradicting the argument of Dunning (1993). Most unexpectedly, however, the Dice coefficient (which has never figured prominently as an association measure) obtains significantly higher precision than log-likelihood. With this experiment, the "sonic barrier" hypothesis was falsified: there is indeed room for improvement over log-likelihood.

The choice of association measures for the lexicographic evaluation had been based on the literature on collocation extraction. It seemed quite plausible that other, previously neglected association measures might give even better results than Dice. In order to support experiments with a wide range of different association measures, the manually annotated database was extended to cover (a random sample of) all frequent adjective-noun combinations ($f \geq 20$) in the *Frankfurter Rundschau* corpus. Since the full data set would be biased towards the measures considered in the original experiment, only this high-frequency subset has been publically released and is described in the following sections.

3. Data preparation and manual annotation

The German adjective-noun database (codenamed `La11t`) has been derived from the *Frankfurter Rundschau* corpus, containing approx. 40 million tokens (words and punctuation) of text from German newspaper articles published in the years 1992–1993.[3] The corpus was part-of-speech tagged with TreeTagger (Schmid, 1995) and lemmatised with IMSLex (Lezius et al., 2000). Adjective-noun combinations – consisting of the head of a noun phrase and a prenominal modifying adjective – were extracted using the part-of-speech patterns described and evaluated by Evert and Kermes (2003). Only 8,546 adjective-noun pairs with cooccurrence frequency $f \geq 20$ were retained as collocation candidates.

A random sample of 1,252 candidates ($\approx 15\%$) was manually annotated by four professional lexicographers of Langenscheidt KG, Munich. The main criterion was usefulness for the compilation of a large bilingual (German-English) dictionary, but finer distinctions were also made by the annotators. Each candidate was classified into one of the following 6 categories:

1. *True collocations:* these candidates are at least partly lexicalised and need to be listed in a dictionary. They can be equated with the notion of multiword expression in computational linguistics. (Ex.: *autofreie Zone* 'zone in which no cars are allowed', *böses Blut* 'bad blood', *das gelbe Trikot* 'the yellow jersey')

2. *Habitual combinations:* these candidates have some idiosyncratic properties (often semi-compositional), but usually allow limited substitution of components with semantically related words. Only some items from such a series need to be listed in the dictionary. Habitual combinations fall into the grey area between multiword expressions and free combinations. (Ex.: *brütende Hitze* 'stifling heat', *neuer Anlauf* 'another go', *technische Daten*, 'technical specification')

3. *Familiar combinations:* mostly free, but frequent combinations without contrastive relevance. They often provide good examples to illustrate the usage of a headword. (Ex.: *ehemaliger Schüler* 'former pupil', *günstiges Angebot* 'bargain, good offer', *unbekanntes Ziel* 'unknown destination')

4. *Candidates with unclear status:* these items may assist lexicographers in the compilation process, but are probably not directly relevant for a bilingual dictionary (Ex.: *neuer Meister* 'new champion', *übrige Zeit* 'remaining time')

5. *Non-collocational:* recurrent combinations that are clearly not relevant for a bilingual dictionary, although they might help lexicographers and translators understand the usage of a headword. (Ex.: *Deutsche Bundesbank* 'Central Bank of Germany', *erstes Semester* 'first term at university', *heißer Sommer* 'hot summer')

6. *Trash:* mostly tagging and lemmatisation errors, as well as some combinations that are idiosyncratic for the corpus used. (Ex.: *[unter] anderem Werke [von]*: adverbial misinterpreted as adjective, *Höchster Stadtpark*: district *Höchst* misinterpreted as superlative of adjective *hoch* 'high', *[Die] verliebte Wolke* 'cloud in love': name of a stage play)

For each candidate, the annotators were given up to 10 randomly selected corpus examples. Due to time constraints, an evaluation of inter-annotator agreement could not be carried out, but the four lexicographers discussed all decisions among themselves. In some cases, lemmatisation errors or incomplete extraction of a larger multiword expression were considered as true positives if the correct form could easily be reconstructed from the corpus examples. Table 1 shows the number and percentage of candidates for each of the six categories. The baseline precision of the entire database ranges from 29.3% (if only true collocations in category 1 are accepted as true positives) to 50.9% (if all useful candidates in categories 1–3 are accepted).

1	2	3	4	5	6
367	153	117	45	537	33
29.3%	12.2%	9.4%	3.6%	42.9%	2.6%

Table 1: Number of candidates and corresponding percentage for each annotation category in the `La11t` database.

4. Availability and use

The `La11t` database is made available as a TAB-delimited text file with a single header row specifying variable names

[3] The *Frankfurter Rundschau* corpus is part of the ECI Multilingual Corpus I distributed by ELSNET. See `http://www.elsnet.org/eci.html` for more information and licensing conditions.

for the columns. This is the native format of the UCS toolkit (Evert, 2004); it also works well with statistical software such as R (R Development Core Team, 2008) and spreadsheet programs like Microsoft Excel. The table columns are:

1. `l1` = adjective (lemma)
2. `l2` = noun (lemma)
3. `n.cat` = collocational status (category assigned by lexicographers, cf. Section 3.)

Since the German words contain non-ASCII characters, versions in Unicode (UTF-8) and Latin1 (ISO-8859-1) encoding are provided. The database can be downloaded from the Resources section of `http://multiword.sf.net/`. It may be used freely for academic research and all non-commercial purposes under the terms of the Creative Commons Attribution-Noncommercial (CC-BY-NC) license, version 3.0 unported.

The `La11t` database is primarily useful for the evaluation of association measures and other ranking methods for collocation and multiword candidates. It also supports the optimisation of association measures with machine learning techniques, which can either take the form of a two-way classification task (with true positives belonging to category 1, categories 1–2, or categories 1–3) or of a multi-way classification task that distinguishes between all six categories. As a simplified problem, three-way classification into category groups 1, 2–4 and 5–6 is suggested.

In order to facilitate such experiments, cooccurrence frequency data from the same *Frankfurter Rundschau* corpus are provided together with the database in two formats: (a) a list of *cooccurrence tokens* with adjective and noun lemma, partially disambiguated morphosyntactic information, and the surface realisation of the expression; and (b) a table of *pair types* with their frequency signatures[4] in the UCS data set format (Evert, 2004). It has to be noted that these resources contain no data for some of the adjective-noun candidates, or indicate a cooccurrence frequency far below the threshold of $f \geq 20$. The reason is that the frequency data were obtained from a re-annotated version of the corpus in which some tagging and lemmatisation errors have been corrected (by using improved releases of the tagger and morphology).

5. Acknowledgements

Our thanks are due to the lexicographers at the Redaktion Wörterbücher, Langenscheidt KG, Munich for their enthusiastic support of this project, the annotation work they carried out and their willingness to release the data without further restrictions.

[4] A frequency signature consists of the cooccurrence frequency of a pair type, the marginal frequencies of its two components, and the sample size. It provides the same information as a 2×2 contingency table, and automatic translation between the two data structures is possible.

6. References

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

Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques.* Ph.D. thesis, Université Paris 7.

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

Stefan Evert and Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 83–86.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Stefan Evert, Ulrich Heid, and Wolfgang Lezius. 2000. Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In Werner Zühlke and Ernst G. Schukat-Talamazzini, editors, *KONVENS-2000 Sprachkommunikation*, pages 215 – 220. VDE-Verlag.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. See `http://www.collocations.de/` for software and data sets.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*, volume 7 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI & Universität des Saarlandes, Saarbrücken, Germany.

Wolfgang Lezius, Stefanie Dipper, and Arne Fitschen. 2000. IMSLex – representing morphological and syntactical information in a relational database. In Ulrich Heid, Stefan Evert, Egbert Lehmann, and Christian Rohrer, editors, *Proceedings of the 9th EURALEX International Congress*, pages 133–139, Stuttgart, Germany.

Wolfgang Lezius. 1999. Automatische Extrahierung idiomatischer Bigramme aus Textkorpora. In *Tagungsband des 34. Linguistischen Kolloquiums*, Germersheim, Germany.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

R Development Core Team, 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. See also `http://www.r-project.org/`.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, March.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Description of evaluation resource -- German PP-verb data

Brigitte Krenn
Austrian Research Institute for Artificial Intelligence
Freyung 6, 1010 Vienna, Austria
brigitte.krenn@ofai.at

Abstract

A description of the German PP-verb data (German_PNV_Krenn available from http://multiword.sourceforge.net/ is presented. The data comprise preposition-noun-verb triples which were extracted from the Frankfurter Rundschau corpus making use of syntactic structure. For processing the partial parser YAC and the IMSLex morphology were employed. The resulting triples were manually annotated distinguishing Funktionsverbgefüge and figurative expressions from other, non-lexicalized word combinations. Linguistic criteria for identifying Funktionsverbgefüge and figurative expressions from the PNV data are presented and borderline cases are discussed.

1. Introduction

In this contribution we present a description of the German PP-verb data (German_PNV_Krenn available from http://multiword.sourceforge.net). The data comprise preposition-noun-verb (PNV) triples that have been extracted from the Frankfurter Rundschau corpus[1] making use of syntactic structure, i.e., the prepositional head P of a PP, the nominal head N of the NP governed by the preposition and the main verb V co-occurring with the PP in the same clause. No distinction between arguments and modifers is made. To avoid inflation of marginal frequencies, each main verb in a clause is only paired with the nearest PP. For syntactic pre-processing the partial parser YAC (Kermes 2003) was employed. The IMSLex morphology (Lezius et al., 2000) was used for lemmatising the nominal head and the verb. Fused preposition-article combinations were normalised so that the definite article is indicated by a "+" character (e.g. both "im"="in dem" and "ins"="in das" are represented as "in+" in the data set). See Evert (2004, p. 39f) for details of the candidate extraction procedure.

The triples have been manually annotated by the author with respect to MWE, distinguishing two types of lexical collocations, Funktionsverbgefüge (FVG) and figurative expressions (figur), from non-collocative PNV combinations.[2] The resulting data set comprises 21796 PNV combinations of which 549 are classified as FVG and 600 as figurative expressions. The reliability of the annotation (collocation versus non-collocation) has been validated in Krenn et al. (2004), achieving kappa agreement scores above 75% for annotators with thorough linguistic training, and scores between 60% and 70% for non-specialised students in a computational linguistics degree.

The results on intercoder agreement clearly show that distinguishing FVG and figurative expressions from non-collocative word combinations in PNV data requires expert knowledge, and even then, classification has a certain potential for errors, especially when it comes to distinguishing FVG and figur.

It is also important to be aware that reducing the PP-Verb combinations to PNV and normalising the PNV data to their morphological bases obliterates the underlying collocations. Consider for instance the FVG *ins Rollen bringen/kommen* (‘set the ball rolling’, ‘start’). Only the specific surface forms *ins* and *Rollen* can be part of the FVG.. Thus, identifying the underlying collocation variants from the PNV triples only requires native language competencies.

Another imprecision in the data originates from the nearest neighbour pairing of PP and verb. The procedure successfully covers relevant PNV triples in verb final clauses, but may lead to imprecision in verb second clauses. Consider the two sentences

a) *sie hat ihm ihr Auto mit Vorbehalt zur Verfügung gestellt* (verb final),
b) *sie stellt ihm ihr Auto mit Vorbehalt zur Verfügung* (verb second).

Both contain an instance of the FVG zur Verfügung stellen. However, only from example a) the PNV triple ”zu+:Verfügung stellen” will be extracted.

Moreover, some of the extracted PP-verb combinations are part of larger MWEs. See Table 4 for a list of such units.

In the remainder of this contribution, we discuss the linguistic criteria applied for distinguishing Funktionsverbgefüge (section 2) and figurative expressions (section 3) from other PP-verb combinations. A decision tree that supports the distinction between FVG and figur is presented in section 4, and border cases between the two types of collocations are discussed.

[1] The FR corpus is part of the ECI Multilingual Corpus I distributed by ELSNET. Fr more information and licensing conditions se http://www.elsnet.org/eci.html.

[2] Collocations in our terms are lexically motivated word combinations that constitute phrasal units with restrictions in their semantic compositionality and morpho-syntactic flexibility.

2. Funktionsverbgefüge (FVG)

Funktionsverbgefüge are particular verb-object collocations constituted by a nominal and a verbal collocate, the predicative noun and the so called function verb, light verb, or support-verb. For a discussion of FVG in the literature see (Krenn, 2000, p.74). In a vast number of cases the predicative noun is part of a PP, which brings us back to the PNV data set.

Semantically, **FVG function as predicates** comparable to main verbs in sentences. In some cases, FVG may be paraphrased by adjective-copula constructions, e.g.

in Kraft treten ~ wirksam werden (come into force),

and more often by main verbs where the predicative noun is derived from the verb, e.g.

zu Besuch kommen ~ besuchen (visit),
in Auftrag geben ~ beauftragt werden (commission),
unter Beweis stellen ~ beweisen (attest).

Some FVG can be used as active paraphrases of passive constructions, e.g.

zur Anwendung kommen (active) ~ *angewandt werden* (be applied, passive).

Even though the vast majority of **predicative nouns** are de-verbal or de-adjectival, abstract primary nouns with argument structure can also function as predicative nouns. The noun usually combines with more than one verb. Accordingly, FVG with identical predicative noun form more abstract types. An example for such a type is given in Table 1 with the predicative noun *Betrieb* and the corresponding verbs.

predicative phrase	verbs	AA	caus	meaning
in Betrieb	gehen	incho	-	go into operation
	nehmen	incho	+	put into operation
	setzen	incho	+	start up
	sein	neut	-	be running
	bleiben	contin	-	keep on running
	lassen	contin	+	keep (something) running
außer Betrieb	gehen	termin	-	go out of service
	nehmen	termin	+	take out of service
	setzen	termin	+	stop'
	sein	neut	-	be out of order
	bleiben	contin	-	stay out of order
	lassen	contin	+	keep out of order

Table 1: Variations of an FVG

The **support-verb** is considered to be a main verb that has lost major parts of its lexical semantics and mainly contributes Aktionsart and information on causativity to the FVG, while the predicative noun contributes the core meaning. A number of typical support verbs can be identified. Breidt (1993), for instance, presents the following list:

bleiben, bringen, erfahren, finden, geben, gehen, gelangen, geraten, halten, kommen, nehmen, setzen, stehen, stellen, treten, ziehen.

A generally acknowledged list of support-verbs, however, does not exist. Varying lists of FVG are presented in (Herrlitz, 1973; Persson, 1975; Yuan, 1986).

Mesli (1989)[3] distinguishes four **Aktionsarten** (AA): inchoative (incho, begin of process or state), terminative (termin, end of process or state), continuative (contin, continuation of process or state) and neutral (neut). Aktionsart in FVG is mainly expressed by the support-verbs, but prepositions may also determine AA, See Table 1, where the prepositions *in* and *außer* express inchoativity and terminativity. While the verbs *gehen, nehmen, setzen* express change of process.

Causativity (caus) increases the argument structure by one. In Table 1 causative variants are marked with +, noncausative variants with –. There are two verb pairings in the example that express causative-noncausative alternation:

nehmen, setzen (take, put) versus *gehen* (go);
lassen (let) versus *bleiben* (stay).

More examples for causative-noncausative alternation are

setzen (set) versus *kommen*, *geraten*, *treten* (come, get, come);
bringen (bring) versus *kommen* (come);
stellen (put) versus *stehen* (stand).

Note, not all predicative nouns combine with all verbs required to realize the full range of Aktionsart and causativity.

3. Figurative expressions (figur)

Figurative expressions emerge during language use by reinterpretation of the literal meaning of a word combination, and may become conventionalized in the course of time. Similar to FVG, the process of lexicalization is also associated with restrictions in semantic compositionality and syntactic flexibility, and PP and verb constitute a unit functioning as **semantic predicate**. Other than for FVG where the semantics is mainly determined by the noun and the verb adds Aktionsart and causativity, in figurative expressions both noun and verb equally contribute to the core meaning of the whole unit. **Nouns** in figurative expressions are either **concrete** or permit concrete interpretation such as Anfang (begin), Ende (end), Liste (list), Weg (path), Zeit (time), Lebensgefahr (danger of life).

As show in Table 2, some figurative expressions too have **causative** and **noncausative** variants.

[3] We refer to Mesli because of her thorough discussion of Aktionsart and causativity in FVG.

NP		verb	caus
in den	Mittelpunkt (focus)	stellen	+
im		stehen	-
ins	Zentrum (focus)	stellen	+
im		stehen	-
an die	Spitze (top)	stellen	+
an der		stehen	-

Table 2: Figurative expressions with spatial nouns and causative-noncausative variation

Alternatively, there are examples where the noncausative variant is realized with *stehen*, but for the causative variants other verbs than *stellen* are used, e.g.

auf dem {Programm, Spielplan} stehen (‘be in the programme’),
auf {das Programm, den Spielplan} setzen (‘put in the programme’),.

Other verb pairs expressing causative-noncausative alternation are:

bringen − kommen (bring − come),
legen − liegen (lay − lie).

In some cases, only the noncausative variant exists, which may be due to a higher degree of lexicalization, e.g.:

unter (die) Räder kommen (`fall into the gutter'),
zu Tode kommen (`die'),
zum Zug kommen (`get a chance'),
i*m Regen stehen* (`be left out in the cold').

A major group of PNV-combinations with figurative interpretation contains **nouns** that **represent body parts** (Table 3). In case pronominal modification is required, it is indicated with (…). Obligatory determiners are given.

body part	figurative expression
Arm (arm)	unter die Arme greifen (`help somebody out with something').
Augen (eyes)	vor Augen {führen, halten} (`to make something concrete to somebody'), vor Augen liegen (`be visible/see'), aus (den) Augen verlieren (`lose sight of')
Beine, Füße (legs, feet)	auf (…) {Beine, Füße} stellen (`to put something in motion'), auf (…) {Beinen, Füßen} stehen (`stand on one's own two feet')
Fersen (heels)	auf den Fersen bleiben (`be at someone's heels')
Finger (finger)	auf die Finger schauen (`keep a sharp eye on someone')
Gesicht (face)	ins Gesicht schreiben *etwas ist jemanden ins Gesicht geschrieben* (`see something in someone's face'), zu Gesicht stehen (`to suit someone')
Hand (hand)	in die Hand {bekommen, drücken, nehmen} (`get hold of', `(discretely) give', `take something in hand'), aus der Hand geben (`to hand over'), auf der Hand liegen (`be obvious') in die Hände fallen (`fall into someone's hands'), in (…) Hände kommen (`come under the influence/control of someone'), in (…) Händen liegen (`be in someone's hands')
Haut (skin)	unter die Haut gehen (`get under someone's skin')
Herz (heart)	ans Herz legen (`enjoin someone to do something'), am Herzen liegen (`have at heart'), ins Herz schließen (`take to heart'), übers Herz bringen (`have the heart to do something')
Kopf (head)	auf den Kopf fallen *er ist nicht auf den Kopf gefallen* (`he is quite smart'), in den Kopf setzen (`put something into one's head/get something into someone's head'), auf den Kopf stellen (`turn things inside out')

Table 3: Figurative expressions containing nouns denoting body parts

Our data set also contains a number of PNV triples which belong to **larger units**. The respective PNV triples and the full word combination they are a part of are given in Table 4. All examples are classified as figurative expressions in the data set.

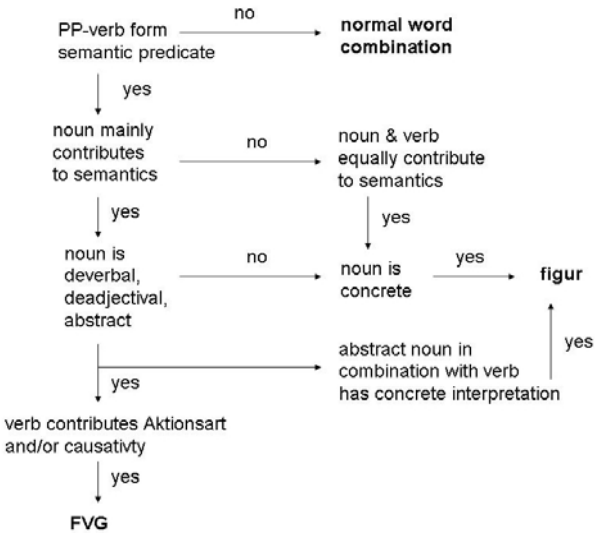
PNV triple	underlying MWE
in:Teufel bringen	in Teufels Küche bringen
in+:Dunkel bringen	Licht ins Dunkel bringen
mit:Sack kommen	mit Sack und Pack kommen
zwischen:Bein werfen	Prügel zwischen die Beine werfen
unter:Scheffel stellen	sein Licht unter den Scheffel stellen
auf:Messer stehen	auf (des) Messers Schneide stehen
in:Bauch stehen	die Beine in den Bauch stehen
auf:Kopf treffen	den Nagel auf den Kopf treffen
auf:Nummer gehen	auf Nummer sicher gehen
auf:Stirn treiben	Schweis auf die Stirn treiben
aus:Segel nehmen	jemanden den Wind aus den Segeln nehmen
durch:Rechnung machen	jemanden einen Strich durch die Rechnung machen
im:Pfeffer liegen	wissen wo der Hase im Pfeffer liegt

Table 4: Examples of PNV triples contained in larger lexicalized units

4. Summary and conclusion

The following decision tree has been designed to help classify the reference data. First of all, those PP-verb combinations which function as semantic predicates are separated from the others. For the former it is investigated whether noun and verb equally contribute to the semantics of the predicate, thus separating potential FVG from figurative expressions. In a next step the noun is investigated for being abstract or concrete, and the verb is analyzed with respect to Aktionsart and causativity.

A number of PP-verb combinations exist that show characteristics of FVG, but are also comparable to figurative expressions. A distinction of these cases is hard, and may result in arbitrary or inconsistent annotations.



The given data set is not free of such inconsistencies. Thus, when comparing extraction methods on the data set, one should separate the evaluation of their performance in identifying collocations (FVG and figure) as opposed to non-collocations, and in distinguishing between FVG and figur.

To get a flavor of border cases see the following examples. Consider for instance *am Anfang stehen* (at the beginning stand, `be at the beginning'). *Anfang* is on the one hand derived from the verb *anfangen* (begin), on the other hand spatial interpretation of the word combination suggests itself, and thus speaks for a classification as figurative expression. The figurative aspect is even more prevalent in the word combination *in den Anfängen stecken* (`be at the first stage'). Similarly *vor der Auflösung stehen* (`be in its final stages') is figurative, but can be paraphrased by the passive construction *aufgelöst werden*, which is an indicator for FVG.

Other examples are *an Bord gehen* (‘go on board’) ~ *borden* (board), *über Bord gehen* (‘go overborad’), *am Pranger stehen* (‘be in the stocks’), *an den Pranger stellen* (‘to pillory’). Both *Bord* and *Pranger* are conrete nouns. The PP-verb combinations have a strong figurative reading. *an Bord gehen* and *am Pranger stehen*, *an den Pranger stellen* can be paraphrased by their related verbs, i.e. *borden*, *angeprangert werden* (passive), and *anprangern*.

All in all, the classification of these border cases requires further investigation. Looking at the pragmatics of their use possibly may lead to better insights. Extracting concordances from corpora should be the first step.

5. References

Breidt, E. (1993). Extraction of N-V-collocations from text corpora: A feasibility study for German. In Proceedings of the 1st ACL Workshop on Very Large Corpora, Columbus, Ohio. (a revised version is available from http://arxiv.org/abs/cmp-lg/9603006).

Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut fuer maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from http://www.collocations.de/phd.html.

Herrlitz, W. (1973). Funktionsverbgefüge vom Typ “in Erfahrung bringen”. Ein Betrag zur generativ-transformationonellen Grammatik des Deutschen. *Lingusitsiche Arbeiten*, (1).

Kermes, H. (2003). Off-line (and On-line) Text Analysis for Computational Lexicography. Ph.D. thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts fuer Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.

Krenn, B. (2000). The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations, volume 7 of Saarbrücken Dissertations in Computational Linguistics and Language Technology. DFKI & Universitaet des Saarlandes, Saarbruecken, Germany. Available from http://www.dfki.de/lt/diss_order.php.

Krenn, B., Evert, S.; Zinsmeister, H. (2004). Determining intercoder agreement for a collocation identification task. In Proceedings of KONVENS 2004, Vienna, Austria, pp. 89-96.

Lezius, W., Dipper, S., Fitschen, A. (2000). IMSLex - representing morphological and syntactical information in a relational database. In U. Heid, S. Evert, E. Lehmann, and C. Rohrer (eds.), Proceedings of the 9th EURALEX International Congress, Stuttgart, Germany, pp. 133-139.

Persson, I. (1975). Das System der kausativen Funktionsverbgefüge. Liber, Malmö.

Yuan, J. (1986). Funktionsverbgefüge im heutigen Deutsch. Eine Analyse und Kontrastierung mit ihren chinesischen Entsprechungen. Sammlung Groos 28.

6. Acknowledgements

The author would like to thank Stefan Evert for his efforts to redo the syntactic pre-processing of the corpus; Hannah Kermes for adapting her parser to meet the needs for extracting the PP verb data; Heike Zinsmeister for conducting the tests on intercoder agreement as part of one of her computational linguistics lessons.

Reference Data for Czech Collocation Extraction

Pavel Pecina

Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic
pecina@ufal.mff.cuni.cz

Abstract

We introduce three reference data sets provided for the MWE 2008 evaluation campaign focused on ranking MWE candidates. The data sets comprise bigrams extracted from the *Prague Dependency Treebank* and the *Czech National Corpus*. The extracted bigrams are annotated as collocational and non-collocational and provided with corpus frequency information.

1. Motivation

Gold standard reference data is absolutely essential for empirical evaluation. For many tasks of Computational Linguistics and Natural Language Processing (such as machine translation or word sense disambiguation) standard and well designed reference data sets are widely available for evaluation and development purposes. Since this has not been the case for the task of collocation extraction, we decided to develop a complete test bed on our own with the aim to use it for evaluation of methods for collocation extraction (Pecina and Schlesinger, 2006).

In this paper we presents three sets of bigrams extracted from the *Prague Dependency Treebank*: one set consists of dependency (syntactical) bigrams, the second one of surface (adjacent) bigrams, and the third one contains instances of the second one in the *Czech National Corpus*. The extracted bigrams are annotated as collocational and non-collocational (and also assigned to finer-grained categories). The reference sets are associated with corpus frequency information for easy computation of association measure scores. All the data sets are publicly available from the MWE wiki page[1].

2. Prague Dependency Treebank

The *Prague Dependency Treebank 2.0* (PDT) is a moderate sized corpus provided with manual morphological and syntactic annotation. By focusing only on two-word collocations, PDT provides sufficient evidence of observations for a sound evaluation. By default the data is divided into training, development, and evaluation sets. We ignored this split and used all data annotated on the morphological and analytical layer: a total of 1 504 847 tokens in 87 980 sentences and 5 338 documents.

2.1. Treebank Details

The Prague Dependency Treebank[2] has been developed by the Institute of Formal and Applied Linguistics and the Center for Computational Linguistics, Charles University, Prague and it is available from LDC[3] (catalog number LDC2006T01). It contains a large amount of Czech texts with complex and interlinked annotation on morphological, analytical (surface syntax), and tectogrammatical (deep syntax) layer. The annotation is based on the long-standing Praguian linguistic tradition, adapted for the current Computational Linguistics research needs.

Morphological Layer

On the morphological layer each word form (token) is assigned a *lemma* and a *morphological tag*. Combination of the lemma and the tag uniquely identifies the word form. Two different word forms differ either in lemmas or in morphological tags. Lemma has two parts. First part, the *lemma proper*, is a unique identifier of the lexical item. Usually it is the base form (e.g. first case singular for a noun, infinitive for a verb, etc.) of the word, possibly followed by a number distinguishing different lemmas with the same base forms (different word senses). Second part is optional. It contains additional information about the lemma (e.g. semantic or derivational information). Morphological tag is a string of 15 characters where every position encodes one morphological category using one character. Description of the categories and range of their possible values are summarized in Table 1. Details of morphological annotation can be found in (Zeman et al., 2005).

Pos	*Name*	*Description*	*# Values*
1	**POS**	**Part of speech**	**12**
2	SubPOS	Detailed part of speech	60
3	**Gender**	**Gende**r	**9**
4	Number	Number	5
5	Case	Case	8
6	PossGender	Possessor's gender	4
7	PossNumber	Possessor's number	3
8	Person	Person	4
9	Tense	Tense	5
10	**Grade**	**Degree of comparison**	3
11	**Negation**	**Negation**	2
12	Voice	Voice	2
13	Reserve1, 2	Reserve	-
14	Reserve2	Reserve	-
15	Var	Variant, style	10

Table 1: Morphological categories encoded in Czech tags.

Analytical Layer

Analytical layer of PDT serves to encode sentence *dependency structures*. Each word is linked to its *head word* and assigned its *analytical function* (dependency type). If we think of a sentence as a graph with words as nodes and dependency relation as edges, the dependency structure is

[1] http://multiword.wiki.sourceforge.net/
[2] http://ufal.mff.cuni.cz/pdt2.0/
[3] http://www.ldc.upenn.edu/

Id	*Form*	*Lemma*	*Full Tag*	*Parent Id*	*Afun*	*Id*	*Lemma Proper*	*Reduced Tag*	*Parent Id*	*Afun*
1	Zbraně	zbraň	NNFP1-----A----	0	ExD	1	zbraň	NF-A	0	Head
2	hromadného	hromadný	AANS2----1A----	3	Atr	2	hromadný	AN1A	3	Atr
3	ničení	ničení_^(*3it)	NNNS2-----A----	1	Atr	3	ničení	NN-A	1	Atr

Table 2: Example of annotated and normalized expression (*weapons of mass destruction*). A normalized form consists of a lemma proper (lemma without technical suffixes) and a reduced morphological tag (positions 1, 3, 10, 11 of the full tag).

a tree – a directed acyclic graph having one root. Details of analytical annotation can be found in (Hajič et al., 1997).

2.2. Collocation Candidate Data Sets

Two collocation candidate data sets were obtained from PDT. Both were extracted from morphologically normalized texts and filtered by a frequency filter and a part-of-speech filter. Details of these steps are the following:

Morphological Normalization
The usual role of morphological normalization is to canonize morphological variants of words so that each word (lexical item) can be identified regardless its actual morphological form. This technique has been found very beneficial for example in information retrieval, especially on morphologically rich languages such as Czech.Two basic approaches to this problem are: *stemming*, where a word is transformed (usually heuristically) into its *stem* which often does not represent a meaningful word, and *lemmatization*, where a word is properly transformed into its base form (lemma) by means of morphological analysis and disambiguation.
The latter approach seems more reasonable in our case (manually assigned lemmas are available in PDT) but it is not completely adequate. By transforming words only into lemmas we would loose some important information about their lexical senses that we want to preserve and use to distinguish between occurences of different collocation candidates. For example *negation* and *grade* (degree of comparison) significantly change word meanings and differentiate between collocation candidates (eg. “secure area” vs. “insecure area”, "big mountain“ vs. ”(the) highest mountain“). Indication of such morphological categories is not encoded in a lemma but rather in a tag. With respect to our task, we decided to normalize word forms by transforming them into combination of a *lemma* (lemma proper, in fact; the technical suffixes in PDT lemmas are omitted) and a *reduced tag* that comprises the following morphological categories: *part-of-speech*, *gender*, *grade*, and *negation* (highlighted in Table 1). For similar reasons and also in order to decrease granularity of collocation candidates, we simplified the system of Czech analytical functions by merging some of them into one value.

Part-of-Speech Filtering
A part-of-speech filter is a simple heuristic that improves results of collocation extraction methods a lot (Justeson and Katz, 1995): the collocation candidates are passed through a filter which only lets through those patterns that are likely to be 'phrases' (potential collocations). Justenson and Katz (1995) filtered the data in order to keep those that are more likely to be collocations than others; for bigram collocation extraction they suggest to use only patterns A:N (adjective–noun) and N:N (noun–noun). We, however, deal with a broader notion of collocation in our evaluation and this constraint would be too limitative. We filter out candidates having such part-of-speech patterns that *never* form a collocation (at least in our data), in other words to keep the cases with part-of-speech patterns that can *possibly* form a collocation. This step does not effect the evaluation because it can be done prior to all extraction methods. The list of employed patterns is presented in Table 3. It was proposed congruently by our annotators before the annotation process described in Section 2.3.

Frequency Filtering
As mentioned earlier our motivation to create the reference data set was empirical evaluation of methods for collocation extraction. To ensure that the evaluation is not biased by low-frequency data, we limit ourselves only on collocation candidates occurring in PDT more than five times. The less frequent candidates do not meet the requirement of sufficient evidence of observations needed by some methods (they assume normal distribution of observations and/or become unreliable when dealing with rare events). Moore (2004) argues that these cases comprise majority of all the data (the well-known Zipfian phenomenon) and should not be excluded from real-world applications.

PDT-Dep
Dependency trees from the treebank were broken down into the dependency bigrams. From all PDT sentences we obtained a total of 635 952 different dependency bigram types (494 499 of them were singletons). Only 26 450 of them occur in the data more than five times. After applying the frequency and part-of-speech pattern filter we obtained a list of 12 232 collocation candidates (consisting of a normalized head word and its modifier, plus their dependency type) further referred to as *PDT-Dep*.

PDT-Surf
Although collocations form syntactic units by definition, we can attempt to extract collocations also as *surface bigrams* (pairs of adjacent words) without guarantee that they form such units but with the assumption that majority of bigram collocations can not be modified by insertion of another word and in text they occur as surface bigrams (Manning and Schütze, 1999, chapter 5). This approach does not require the source corpus to be parsed, which is usually a time-consuming process accurate only to a certain extent. A total of 638 030 surface bigram types was extracted from PDT, 29 035 of them occurred more then five times and after applying the part-of-speech filter we obtained a list of 10 021 collocation candidates (consisting of normalized components) further referred to as *PDT-Surf*. 974 of these bigrams do not appear in *PDT-Dep* test sets (if

we ignore the syntactical information).

2.3. Manual Annotation

Three educated linguists, familiar with the phenomenon of collocations, were hired to annotate the reference data sets extracted from PDT in paralell. To consolidate their notion of collocation we adopt the definition from Choueka (1988): "A collocation expression is a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components." It is relatively wide and covers a broad range of lexical phenomena such as idioms, phrasal verbs, light verb compounds, technological expressions, proper names, and stock phrases. It requires collocations to be syntactic units – subtrees of sentence dependecy trees in case of dependecy syntax used in PDT.

The dependency bigrams in *PDT-Dep* were assessed first. The annotation was performed independently and without knowledge of context. To minimize the cost of the process each collocation candidate was presented to each annotator only once although it could appear in many different contexts. The annotators were instructed to judge any bigram which could *eventually* appear in context where it has a character of collocation, as a *collocation*. E.g. idiomatic expressions were judged as collocations although they can also occur in contexts where they have a literal meaning. Similarly for other types of collocations. As a result the annotators were quite liberal in their judgments which we exploited in combining their outcomes.

During the assessment the annotators also attempted to classify each collocation into one of the following categories. This classification, however, was not intended as a result of the process but rather as a way how to clarify and simplify the annotation. Any bigram that can be assigned to any of the categories was considered a collocation.

1. stock phrases
   zásadní problém (major problem), konec roku (end of a year)
2. names of persons, organizations, geographical locations, and other entities
   Pražský hrad (Prague Castle), Červený kříž (Red Cross)
3. support verb constructions
   mít pravdu (to be right), činit rozhodnutí (make decision)
4. technical terms
   předseda vlády (prime minister), očitý svědek (eye witness)
5. idiomatic expressions
   studená válka (cold war), visí otazník (hanging question mark ~ open question)

The surface bigrams from *PDT-Surf* were annotated in the same fashion but only those collocation candidates that do not appear in *PDT-Dep* were actually judged (974 items). Technically we removed the syntactic information from *PDT-Dep* data and transfer the annotations to *PDT-Surf*, if a surface bigram from *PDT-Surf* appears also in *PDT-Dep* it is assigned the same annotation from all three annotators.

Inter-annotator Agreement

The interannotator agreement among all the categories of collocations (plus a 0 category for non-collocations) was relatively low: the average accuracy between two annotators on *PDT-Dep* was as low as 72.88%, the average Cohen's κ was estimated as 0.49. This demonstrates that the notion of collocation is very subjective, domain-specific, and also somewhat vague. Since we did not distinguish between different collocation categories – ignoring them (considering only two categories: *true collocations* and *false collocations*) increased the average accuracy up to 80.10% and the everage Cohen's κ to 0.56. The three annotators were employed to get a more precise and objective idea about what can be considered a collocation by combining their independent outcomes. Only those candidates that *all* three annotators recognized as collocations (of any type) were considered *true collocations* (full agreement required). The *PDT-Dep* reference data set contained 2 557 such bigrams (21.02%) and *PDT-Surf* data set 2 293 (22.88%). For comparison of these reference data set see Figure 1.

Pattern	*Example*	*Translation*
A:N	trestný čin	*criminal act*
N:N	doba splatnosti	*term of expiration*
V:N	kroutit hlavou	*shake head*
R:N	bez problémů	*no problem*
C:N	první republika	*First Republic*
N:V	zranění podlehnout	*succumb*
N:C	Charta 77	*Charta 77*
D:A	volně směnitelný	*free convertible*
N:A	metr čtvereční	*squared meter*
D:V	těžce zranit	*badly hurt*
N:T	play off	*play-off*
N:D	MF Dnes	*MF Dnes*
D:D	jak jinak	*how else*

Table 3: Part-of-speech patterns for filterring collocation candidates (A – adjectives, N – nouns, C – numerals, V – verbs, D – adverbs, R – prepositions, T – particles).

3. Czech National Corpus

At the time of multi-billion word corpora, a corpus of the size of PDT is certainly not sufficient for real-world applications. We attempted to extract collocations also from larger data – a set of 242 million tokens from the *Czech National Corpus*. This data, however, lacks of any manual annotation, hence we settle for an automatic part-of-speech tagging (Hajič, 2004) and extracted collocation candidates as surface bigrams similarly as in the case of *PDT-Surf*.

3.1. Corpus Details

The *Czech National Corpus* (CNC) is an academic project with the aim to build up a large computer-based corpus, containing mainly written Czech[4]. The data we used comprises of two synchronous (containing contemporary written language) corpora SYN2000 and SYN2005 (ICNC, 2005) each containing about 100 million running words (excluding punctuation).

3.2. Automatic Preprocessing

SYN2000 and SYN2005 are not manually annotated, neither on morphological nor analytical layer. Manual anota-

[4]http://ucnk.ff.cuni.cz/

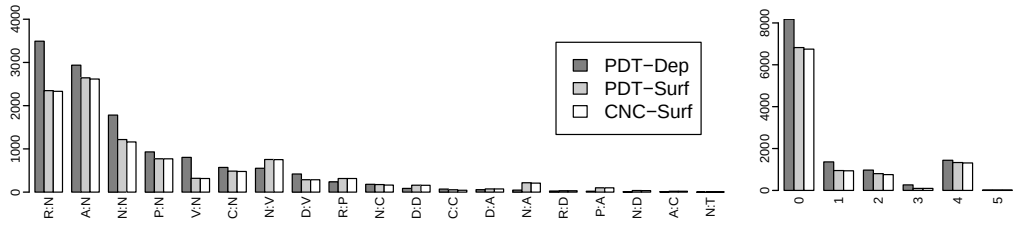


Figure 1: Part-of-speech pattern distribution in the reference data sets(left) and distribution of collocation categories in the reference data sets assigned by one of the annotators (right).

tion of such amount of data would be unfeasible. These corpora, however, are processed by a part-of-speech tagger.

3.3. Collocation Candidate Data Set

CNC-Surf

From the total of 242 million tokens from SYN2000 and SYN2005 we extracted more then 30 million surface bigrams (types). We followed the same procedure as for PDT reference data and after applying the part-of-speech and frequency filters, the list of collocation candidates contained 1 503 072 surface bigrams. Manual annotation of such amount of data was infeasible. To minimize the cost we selected only a small sample of it – already annotated bigrams from the *PDT-Surf* reference data set – a total of 9 868 surface bigrams further called *CNC-Surf*. All these bigrams appear also in *PDT-Surf*, the remaining 153 do not occur in the corpora more than five times. The major difference is only in the frequency counts provided with the data set. This reference data set contains 2 263 (22.66%) *true collocations* – candidates that all three annotators recognized as collocations (of any type). For comparison with the reference data sets extracted from PDT see Figure 1.

Reference Data Set	*PDT-Dep*	*PDT-Surf*	*CNC-Surf*
sentences	87 980	87 980	15 934 590
tokens	1 504 847	1 504 847	242 272 798
words (no punctuation)	1 282 536	1 282 536	200 498 152
bigram types	635 952	638 030	30 608 916
after frequency filtering	26 450	29 035	2 941 414
after part-of-speech filtering	12 232	10 021	1 503 072
collocation candidates	12 232	10 021	9 868
sample size (%)	100	100	0.66
true collocations	2 557	2 293	2 263
baseline precision (%)	21.02	22.88	22.66

Table 4: Statistics of the three reference data sets and the corpora they were extracted from.

4. Summary

We prepared three reference data sets for the task of identifying collocation candidates. All of them consist of two-word collocation candidates. *PDT-Dep* and *PDT-Surf* were extracted from the manually annotated Czech *Prague Dependency Treebank* and differ only in the character of bigrams. *PDT-Dep* consists of dependency bigrams and *PDT-Surf* of surface bigrams. Both were filtered by the same part-of-speech pattern filter and frequency filter. Manual annotation was done exhaustively – no sampling was needed, true collocations are indicated in all data.

CNC-Surf reference data set was extracted from much larger data from the *Czech National Corpurs* and comprises surface bigrams also appearing in *PDT-Surf*. It can be considered as a random sample from the full set of collocation candidates filtered by the same part-of-speech pattern filter and frequency filter as the PDT reference data.

Acknowledgments

This work has been supported by the Ministry of Education of the Czech Republic, projects MSM 0021620838.

5. References

Yaacov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*.

Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alla Bémová. 1997. A manual for analytic layer tagging of the prague dependency treebank. Technical Report TR-1997-03, ÚFAL MFF UK, Prague, Czech Republic.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, volume 1. Charles University Press, Prague.

ICNC. 2005. Czech national corpus. Institute of the Czech National Corpus Faculty of Arts, Charles University, Praha, http://ucnk.ff.cuni.cz.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on EMNLP*, Barcelona, Spain.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, Australia.

Dan Zeman, Jiří Hana, Hana Hanová, Jan Hajič, Emil Jeřábek, and Barbora Vidová Hladká. 2005. A manual for morphological annotation, 2nd edition. ufal technical report no. tr-2005-27. Technical Report TR-2005-27, ÚFAL MFF UK, Prague, Czech Republic.

A Lexicon of shallow-typed German-English MW-Expressions and a German Corpus of MW-Expressions annotated Sentences

Dimitra Anastasiou, Michael Carl
Institut für Angewandte Informationsforschung
Martin Luther Str. 14, 66111 Saarbrücken, German
dimitraa@coli.uni-sb.de, carl@iai.uni-sb.de

Abstract

We describe a bilingual German-English lexicon of idiomatic Multiword Expressions (MWEs) and a corpus of German sentences. The lexicon consists of 871 idiomatic MWEs-entries, 598 of which are verb phrases (VPs). The corpus includes 536 German sentences and it was assembled from three different resources: a subset of the Europarl corpus, a mixture of manually constructed data and examples filtered from the Web and sentences extracted from the digital lexicon of the German language in the 20th century. The sentences of the corpus include idiomatic MWEs which are entries in the bilingual lexicon. In our paper we give a brief introduction to semantics and syntax of German idiomatic MWEs and their relationship. We describe the German-English bilingual lexicon and we focus on the syntactic patterns in which verbal idiomatic MWEs can occur according to the topological field model. We also look at the corpus and at each data set separately. Last but not least, we give a statistical analysis of the realisation of continuous (when MWEs form a chain) and discontinuous (when alien elements intervene among the idiom's parts) MWEs in the German corpus.

1. Introduction

We describe two types of resources: a bilingual German-English lexicon of idiomatic Multiword Expressions (MWEs) and a corpus of German sentences. These sentences include idiomatic MWEs which are entries of the bilingual lexicon.

In section 2 we briefly discuss general properties of German idiomatic MWEs including some of their basic semantic and syntactic properties and in section 3 we give an overview of the idiomatic MWEs contained in the German-English bilingual lexicon. In section 4 we introduce the topological field model and we describe the syntactic patterns of verbal idiomatic MWEs giving corresponding examples. Section 5 refers to the corpus of German sentences which we collected from various resources[1]: Europarl corpus[2] (EP), manually constructed data and real examples (MDS) and the digital lexicon of the German language in the 20th century[3] (DWDS).

2. Properties of German idiomatic MWEs

Idioms are semantically categorised into non-compositional, partially compositional and strictly compositional idioms (see Rothkegel, 1989; Keil, 1997). The major characteristic of (non-compositional) idioms is that they are meaningful linguistic units whose meaning is not a function of the constituent words (Erbach, 1991).

Wasow et al. (1983) argue that syntactic flexibility is tied to semantic transparency, i.e. the strictly compositional idioms have only relative fixedness. Fraser (1970) and Jackendoff (1977) state that idioms can in principle undergo any syntactic operation their literal counterparts can undergo.

In this paper we focus on syntax and particularly, on the syntactic permutations of verbal idioms, i.e. shifts of idiom's components throughout the sentence.

3. Lexicon of idiomatic MWEs

We have manually collected 871 idiomatic MWEs – lexicon entries from our resources: EP, MDS and DWDS.

Each idiom entry consists of four tab-separated columns: set of lemmas and set of Part-of-Speech – type (Table 1).

German	Gtype	English	Etype
Blut und Wasser schwitzen	Verb	be in a cold sweat	Verb

Table 1: Fields of an idiomatic MWE lexicon entry

Sometimes, an idiom can be used both literally and idiomatically, e.g. *be in a cold sweat*, but when it is used in its literal meaning, it has bizarre side-effect and/or it should have referred to extraordinary situations (Burger, 2007). Also, they are less common than those used in their idiomatic meaning.

For most of the 871 entries, the German (**Gtype**) and the English type (**Etype**) are identical. The distribution of these entries is shown in table (2).

DE & EN equal PoS	**826**
Verbs	598
itj (Interjections)	163
noun (Noun phrases)	37
p (Prepositional phrases)	28

Table 2: Distribution of entries of equal type

We label verbal idioms, despite being VPs as *verb*.

Proverbs and sayings take the type 'interjection' (*itj*), as they maintain their lemma's forms. These entries are

[1] The lexicon and the corpus of German sentences is the basis of the first author's PhD thesis.

[2] http://www.statmt.org/europarl/

[3] http://www.dwds.de/

essentially opaque and cannot be modified.

Gt	German	Et	English
itj	das Maß ist voll	itj	enough is enough
itj	alle Jahre wieder	itj	year after year

Table 3: Examples of interjections

In our typology, entries of type *p* (PPs) function similarly to interjections (*itj*), as they are not usually modified when realised in the German or English sentence. However, their syntactic function is slightly different.
Table (4) shows the distribution of entries where the language sides have different types.

German type	English type	45
verb	itj	15
itj	verb	14
itj	noun	3
itj	p	3
noun	adjective	3
noun	p	2
p	adverb	2
p	itj	2
p	noun	1

Table 4: Distribution of entries of different type

MWEs are 'shallow typed' as shown in tables (2), (3) and (4).
Table (5) shows the structures of German verbal idioms and their occurrence in the bilingual lexicon. The most common structure is VP with PP constituent (PP-V), e.g.:

(1) *wieder auf dem Damm sein*
be back on one's feet

followed by VP with NP constituent (NP-V), such as:

(2) *Blut und Wasser schwitzen*
be in a cold sweat

and then by VP with both constituents (NP-PP-V), like:

(3) *den Bock zum Gärtner machen*
set the fox to keep the geese

They are represented in verb final form, as in the lexicon.

Types of German VPs - MWEs	Occurence
PP - V	230
NP - V	198
NP - PP - V	131
PP - Adverb -V	13
PP - PP- V	9
Adverb - V	4
Adjective - V	4
Subordinate Clause - V	4
Adverb - NP - V	2
NP - Adverb - PP - V	2
Adjective - NP -V	1

Table 5: German verbal idioms

4. Patterns of idiomatic verbal MWEs

Idiomatic verbal MWEs can be realised in two different ways: in a continuous or discontinuous way. In the former case the idiom's constituents occur side by side, while in the latter case alien element(s) intervene(s) among the idiom's constituents. The verb can be on the right of the MWE (typically in German subordinate clauses) or shifted to the left of the nominal or prepositional phrase, as it is the case in main clauses. The possible verb position in German clauses (and consequently when realising idiomatic verbal MWEs) can be formalised based on the *topological field model* (Drach, 1963) and the grammar of Duden (1998). According to this model, the German main clause can be divided into five fields, each of which may contain a certain number of syntactic constituents. The five fields and their constituents are presented below.

- The *pre-field* (VF) contains only one syntactic constituent;
- the *left bracket* (LK) holds the finite verb (in main clauses) or a subordinating conjunction. The LK can be empty in cases of relative clauses or indirect wh-questions.
- the *middle field* (MF) includes diverse permutations of various kinds of syntactic constituents and subordinate clauses;
- the *right bracket* (RK) consists of participles or infinitive forms in case the finite verb is an auxiliary or a modal verb;
- the *post-field* (NF) contains subordinate clauses or coordinated main clauses.

More information about the topological field model can be found in Dürscheid (2000).
The pattern of continuous realisation of idiomatic verbal MWEs is the following:

(4) [4]$\mathbf{iNP}_{MF}$ / $\mathbf{iPP}_{MF}$ / [$\mathbf{iNP}_{MF}$ – $\mathbf{iPP}_{MF}$] $\mathbf{iV}_{RK}$

The most common continuous realisation is the subordinate clause in German which is shown in (5) below. The same pattern applies to idiomatic verbal MWEs in the case of the perfect tense (6) with the finite auxiliary verb ***hat*** or modal verb and in the case of the passive voice (if the passivisation is feasible). The less common continuous realisations are the participle form of the verb, e.g. **ins Fettnäpfchen *tretender*** and the topicalisation (7).

(5) *Obwohl er öfters **ins Fettnäpfchen tritt**, hat er sein Ziel erreicht.*
*Although he often **puts his foot in it**, he reached his goal.*

(6) *Er **hat** während seines Studiums immer **ins Fettnäpfchen getreten.***
*During his studies he always **puts his foot in it.***

[4] The symbols starting with a small *i* stand for *idiom's* + PoS, i.e. *iNP*: idiom's NP, *iPP*: idiom's PP, *iV*: idiom's verb.

(7) ***Ins Fettnäpfchen treten*** *will doch keiner!*
No one wants to ***put their foot in it****!*

Most of the discontinuous realisations of idiomatic verbal MWEs correspond to the pattern (8):

(8) $\mathbf{V}_{LK}$ (Adjectve/Adverb/Participle/Pronoun/ Prepositional Adverb/NP/PP/Subclause)*$_{MF}$ **iNP**$_{MF}$ / **iPP**$_{MF}$ / [**iNP**$_{MF}$ – **iPP**$_{MF}$] (Subclase*$_{NF}$ – V*$_{RK}$)

Sentence (9) exemplifies the pattern. The finite verb ***tritt*** occurs in the LK, and the PP ***ins Fettnäpfchen*** is located at the end of the MF. An optional subordinate clause appears between the verb and the idiom's component PP.

(9) *Er* ***tritt****, obwohl er das nicht will,* ***ins Fettnäpfchen****.*
He ***puts****, although he does not want it,* ***his foot in it****.*

5. German Sentences Corpus

A corpus of 536 German sentences was assembled from three different resources:

1. a subset of the Europarl corpus (**EP**)
2. a mixture of manually constructed data and examples filtered from the Web (**MDS**)
3. sentences extracted from the **DWDS**

The corpus contains sentences with idiomatic MWEs. These MWEs are stored in the bilingual German-English lexicon. We emphasise the importance of including sentences with idioms of all syntactic categories and of every possible permutation in our German corpus. The stronger the permutations of idioms are, the more difficult it is for a Machine Translation (MT) system to translate at a later stage.
In the corpus, each sentence appears on one line. The sentences are categorised according to the data set name and the (continuous and discontinuous) type of idioms. Idiomatic MWEs in the sentences are marked with XML tags. The idiomatic parts are surrounded by angled brackets.

5.1 Europarl Corpus (EP)

The English-German EP corpus consists of 1,313,096 sentences. We preferred a manual search rather than a (semi) automatic one by using fixed criteria for the sake of higher accuracy, as after performing the (semi) automatic search, unwanted junk had been collected. We randomly picked the first 5,000 sentences and found 80 sentences containing idiomatic MWEs: 63 continuous MWEs and 17 discontinuous ones. Koehn (2002) describes Europarl corpus in detail.

5.2 Mixture of Data Sets (MDS)

This mixture data set includes in total 275 idiomatic MWEs, 205 continuous and 70 discontinuous ones. One part of the data was manually constructed; the other part was extracted from corpora stored in web-interfaces. The advantages and disadvantages of the resources are briefly described below.

5.2.1 Manually Constructed Data

A group of students was assigned to manually construct sentences containing idiomatic MWEs. This data set includes various possible permutations of MWEs, stretching the components to every part of the sentence. These permutations are not easy to find in standard corpora, because they are unusual. Despite the strong permutations, the sentences are grammatically correct.
However, sometimes the sentences are very simple and semantically obsolete, since they were mainly constructed to test an automatic MWE matching programme.

5.2.2 Real Examples (Search Engine)

The real examples were mainly searched in *Google*. Our methodology of building this part of MDS is described in three steps:

1. We input into the search tool consecutively more idiom's parts starting from the most distinctive one, i.e. ***Fettnäpfchen*** from the idiomatic phrase ***ins Fettnäpfchen treten***.
2. After one or more attempts and having found the idiomatic phrase in question, we manually extract the sentences containing this idiom and copy them to our corpus file.
3. Exceptionally the idiomatic phrase may be used in its literal meaning. We discard those sentences keeping in our corpus only the sentences where the idiom is used in its idiomatic meaning.

The real example-sentences are either very long with unimportant context or too short, so that their meaning is incomplete. They were carefully selected in respect of their length.
Some sentences were also extracted from the lexicon portal of the University of Leipzig[5]. Here, all data is automatically – though carefully – collected from publicly available sources. The corpora are stored in a uniform schema in a MySQL database. The functionality of a database entails efficient indexing methods and allows the storage of very large resources (Quasthoff, 2006).
After inputting into the portal's search tool the idiom's part ***Fettnäpfchen***, frequency information, co-occurrence statistics and examples which contain the input unit appear in the webpage. There is also a link to the idiomatic phrase ***ins Fettnäpfchen treten*** which leads to another webpage with similar information. Of course, we can input from the beginning the whole idiomatic phrase. When reversed, i.e. ***treten ins Fettnäpfchen***, no results were found. When we input ***getreten*** (participle form of ***treten***) there is a link to ***Fettnäpfchen*** recognizing it as a significant co-occurence.

[5] http://wortschatz.uni-leipzig.de/

5.3 DWDS

The digital lexicon of the German language in the 20th century (DWDS) is a web-interface which was developed by the Berlin-Brandenburg Sciences Academy. It contains a dictionary, several corpora, and word information. The dictionary consists of 130,000 entries. The main DWDS corpus includes 100 million tokens in 79,830 documents. The examples are chronologically ordered and the time span is the whole 20th century. We have extracted 131 sentences (91 with continuous and 40 with discontinuous idiomatic MWEs) mainly from the German newspaper-corpus *DIE ZEIT*. This section alone consists of 106 million tokens in more than 200,000 articles.

We followed the same methodology as for the real examples. DWDS corpora are not downloadable.

5.4 Idiom text types

We have examined in which text types idioms occur by means of a small sample of 50 sentences. Most of them were found in political texts followed by general newspaper articles and literature texts.

5.4 Statistical analysis of idiomatic verbal MWEs

Most of the MWEs in our data are continuous and verbal, and most of them occur in the MDS corpus. Table (6) shows the realisation of the continuous MWEs, and table (7) the realisation of the discontinuous MWEs. Alien elements, such as a pronoun/ NP/ PP/ adverb/ adjective/ subordinate clause, intervene between the verb and the idiom's nominal or prepositional part producing the discontinuous pattern V-X-NP/PP, where X is the alien element (consider example 9 in section 4). However, there are also some verb-final discontinuous realisations of the form [NP]/PP-X-V.

	EP	**MDS**	**DWDS**
NP-V	8	65	15
PP-V	29	106	60
NP-PP-V	4	21	6
PP-PP-V	-	-	1
NP-Adj-V	-	-	1
Proverb	6	6	-
NP	4	-	2
PP	12	4	-
NP-PP	-	-	6
Interjection	-	3	-

Table 6: Realisation of continuous MWEs

	EP	**MDS**	**DWDS**
V-NP	1	8	13
V-PP	16	25	18
V-NP-PP	-	22	9
V-PP-PP	-	1	-
V-PP-Adv	-	1	-
PP-V	-	5	-
NP-PP-V	-	2	-
Proverb	-	2	-
PP	-	4	-

Table 7: Realisation of discontinuous MWE

6. References

Burger, H. ([3]2007). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Schmidt Erich, Berlin.

Drach, E. (1963) [1940]. *Grundgedanken der deutschen Satzlehre*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.

DUDEN Redaktion. (1998). *Grammatik der deutschen Gegenwartssprache*. Mannheim, Germany.

Dürscheid, C. (2000). *Syntax: Grundlagen und Theorien*. Wiesbaden.

Erbach, G. (1991). *Lexical Representation of Idioms*. IWBS Report 169, IBM TR-80.91 – 023, Germany.

Fraser, B. (1970). Idioms within a Transformational Grammar. *Foundations of Language*, 6, pp. 22--42.

Jackendoff, R.S. (1977). *X-bar Syntax: A Study of Phrase Structure*. Cambridge: MIT Press.

Keil, M. (1997). Wort für Wort. Repräsentation und Verarbeitung verbaler Phraseologismen. *Sprache und Information*, 35, Niemeyer Verlag, Tübingen.

Klappenbach, R.; Steinitz, W. (1964). *Wörterbuch der deutschen Gegenwartssprache.* Deutsche Akademie der Wissenschaften zu Berlin. Incorporated into the Website of the Digital Lexicon of the German language in the 20th century: http://www.dwds.de.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit X*. Phuket, Thailand, pp. 79--86.

Quasthoff, U.; M. Richter; C. Biemann. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, pp. 1799--1802.

Rothkegel, A. (1989). *Polylexikalität. Verb-Nomen-Verbindungen und ihre Behandlung in EUROTRA*, EUROTRA-D Working Papers, No 17, Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes.

Wasow, T.; I. Sag; G.. Nunberg. (1983). Idioms: An Interim Report. In *Proceedings of the Thirteenth International Congress of Linguistics*. CIPL, Tokyo, pp. 102--115.

The VNC-Tokens Dataset

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson

Univeristy of Toronto
Toronto, Canada
{pcook, afsaneh, suzanne}@cs.toronto.edu

Abstract

Idiomatic expressions formed from a verb and a noun in its direct object position are a productive cross-lingual class of multiword expressions, which can be used both idiomatically and as a literal combination. This paper presents the VNC-Tokens dataset, a resource of almost 3000 English verb–noun combination usages annotated as to whether they are literal or idiomatic. Previous research using this dataset is described, and other studies which could be evaluated more extensively using this resource are identified.

1. Verb–Noun Combinations

Identifying multiword expressions (MWEs) in text is essential for accurately performing natural language processing tasks (Sag et al., 2002). A broad class of MWEs with distinct semantic and syntactic properties is that of idiomatic expressions. A productive process of idiom creation across languages is to combine a high frequency verb and one or more of its arguments. In particular, many such idioms are formed from the combination of a verb and a noun in the direct object position (Cowie et al., 1983; Nunberg et al., 1994; Fellbaum, 2002), e.g., *give the sack*, *make a face*, and *see stars*. Given the richness and productivity of the class of idiomatic verb–noun combinations (VNCs), we choose to focus on these expressions.

It is a commonly held belief that expressions with an idiomatic interpretation are primarily used idiomatically, and that they lose their literal meanings over time. Nonetheless, it is still possible for a potentially-idiomatic combination to be used in a literal sense, as in: *She **made** a **face** on the snowman using a carrot and two buttons*. Contrast the above literal usage with the idiomatic use in: *The little girl **made** a funny **face** at her mother*. Interestingly, in our analysis of 60 VNCs, we found that approximately half of these expressions are attested fairly frequently in their literal sense in the British National Corpus (BNC).[1] Clearly, automatic methods are required for distinguishing between idiomatic and literal usages of such expressions, and indeed there have recently been several studies addressing this issue (Birke and Sarkar, 2006; Katz and Giesbrecht, 2006; Cook et al., 2007).

In order to conduct further research on VNCs at the token level, and to compare the effectiveness of the varying proposed methods for their treatment, an annotated corpus of VNC usages is required. Section 2 describes our dataset, VNC-Tokens, which consists of almost 3000 English sentences, each containing a VNC usage (token) annotated as to whether it is literal or idiomatic. Sections 3, 4, and 5 respectively describe previous research conducted using VNC-Tokens, other work on idioms which could make use of this dataset, and possible ways in which VNC-Tokens could be extended. We summarize the contributions of the VNC-Tokens resource in Section 6.

[1] `http://www.natcorp.ox.ac.uk`

2. The VNC-Tokens Dataset

The following subsections describe the selection of the expressions in VNC-Tokens, how usages of these expressions were found, and the annotation of the tokens.

2.1. Expressions

We began with the dataset used by Fazly and Stevenson (2006), which includes a list of VNCs. We eliminated from this list any expression whose frequency in the BNC is less than 20 or does not occur in at least one of two idiom dictionaries (Cowie et al., 1983; Seaton and Macaulay, 2002). This gave 60 candidate expressions.

Two expert judges, both native English-speaking authors of this paper, examined the candidate expressions and eliminated 7 of them. The idiomatic meaning of *blow one's (own) horn*, *get the bird*, and *pull one's hair (out)* were not familiar to one judge, and therefore could not be annotated with confidence.[2] For the expressions *catch one's breath*, *cut one's losses*, and *push one's luck* the annotators agreed that a literal interpretation was not possible, while they judged that *give a lift* does not have a clear idiomatic meaning. This gave a final set of 53 expressions.

2.2. Sentence Extraction

To identify usages of a VNC in text, we first parsed the BNC (Collins, 1999), and then looked for sentences containing the component verb and noun from one of our 53 VNCs in a direct object relation. For each expression, 100 sentences containing its usage were randomly selected, and for expressions with less than 100 usages, we extracted all sentences.

This dataset was originally created using the BNC World edition for which licenses are no longer available. A number of files occurring in this version of the BNC are not part of the newer BNC XML edition. Therefore the 8 sentences extracted from these files have been eliminated.

We observed that there were a number of duplicates in our selected sentences. To ensure consistency across the expressions, we therefore also extracted any sentence which contained the same text as any one of the sentences in our dataset. Thus, all expressions have all duplicates included

[2] *Pull one's hair (out)* is a verb-particle construction. Although such expressions may be, to varying degrees, idiomatic, they were not the focus of this study.

for any originally selected sentence. The final dataset consists of 2984 VNC tokens, of which 2920 are unique occurrences.

2.3. Token Annotation

Each instance of the 53 chosen expressions was annotated by the two judges as one of literal, idiomatic, or unknown. During annotation the judges were presented with the single sentence containing the VNC usage; sentences in the surrounding context were not included. If the judge was unable to determine the class of a token based on the sentence in which it occurs, the judge chose the unknown label.

The idiomaticity of an expression is not binary. Expressions may be more or less idiomatic, falling on a continuum ranging from completely literal expressions, i.e., *get the present*, to semantically opaque idioms, i.e., *get the sack* (which has the idiomatic interpretation of losing one's employment). For usages falling towards the middle of this continuum, the human annotators were instructed to choose the most appropriate label according to their judgement, as opposed to using the unknown label.

This dataset was originally intended for use in Cook et al. (2007). The 53 selected expressions were divided into three sets: development, test, and skewed. Skewed contains expressions for which one of the literal or idiomatic meanings is very infrequent, while the expressions in development and test are more balanced across the senses.

The primary annotator annotated all the tokens in each subset of the data. These preliminary annotations were used to divide the expressions into the three sets. The secondary annotator then annotated the sentences in the development set. The judges then discussed tokens on which they disagreed to achieve a consensus annotation. They also discussed the annotation process at length to improve the quality and consistency of their annotations. The primary judge then re-examined their own annotations for the test set to ensure consistency, while the secondary judge annotated these items. Again, disagreements were discussed to come to consensus annotations as well as to refine the annotation process. Consensus annotations were then determined for the skewed set in the same manner as for the test set.

A number of issues arose during reconciliation of disagreements that are worth noting, particularly with respect to usages that fall somewhat towards the middle of the literal–idiomatic continuum. For example, there are idiomatic usages of the expression *have word* that have a meaning that is somewhat related to its literal meaning, as in: *At the moment they only had the word of Nicola's husband for what had happened*.[3] The final annotation for this sentence was idiomatic since the idiomatic meaning was judged to be much more salient than the literal meaning, as in: *In contrast, the French, for example, have two words for citizenship*. Further towards the literal end of the continuum are certain usages of expressions such as *hit the road*. This expression may be used in a clear literal sense, as in: *Gina Coulstock, 18, stumbled, fell heavily and was knocked out when she hit the road*. It may also be used with the idiomatic meaning of departure, as in: *The marchers had hit the road before 0500 hours, and by midday they were limping back to Heumensoord.* However, this expression may also be used in a more intermediate sense, as in: *You turn right when we hit the road at the end of this track*. Such usages of *hit the road*, and similar usages of other expressions, were judged to be figurative extensions of literal meanings, and were therefore classified as literal.

The items in each of the development, test, and skewed sets, along with their number of usages in each sense, are given in Table 1. The observed agreement and unweighted Kappa score for each set, and over all sets, before the judges discussed their disagreements, is given in Table 2.[4]

3. Previous Research Using VNC-Tokens

The only research to date which has made use of VNC-Tokens is that of Cook et al. (2007). They perform an extensive token-based study of VNCs using an earlier version of the development and test subsets of VNC-Tokens for development and evaluation of their methods. Their study is based on the observation that the idiomatic meaning of a VNC tends to be expressed in a small number of preferred lexico-syntactic patterns, referred to as canonical forms (Riehemann, 2001). For example, while both the idiomatic and literal interpretations are available for the phrase *kicked the bucket*, only the literal meaning is possible for *kicked a bucket* and *kicked the buckets*.

Cook et al. hypothesize that idiomatic usages of a VNC will usually occur in one of that expression's canonical forms, while the literal meaning will be expressed in a wider variety of forms. Drawing on established unsupervised methods for determining the canonical forms of a VNC (Fazly and Stevenson, 2006), Cook et al. propose three unsupervised methods for distinguishing literal and idiomatic VNC usages that incorporate their hypothesis.

Their CFORM method relies solely on information about canonical forms, and simply classifies a usage of an expression as idiomatic if it occurs in one of that expression's canonical forms, and as literal otherwise. Their other two methods, $\text{DIFF}_{\text{I-CF, L-NCF}}$ and $\text{DIFF}_{\text{I-CF, L-COMP}}$, incorporate lexical co-occurrence information along with the syntactic information provided by canonical forms. In these methods, three co-occurrence vectors approximating each of the meaning of the target token to be classified, the literal meaning of the expression, and that expression's idiomatic meaning are formed. The vector representing the target is then compared using cosine to those for the literal and idiomatic meanings, and the target is assigned the

[3] All examples in this subsection are taken from the BNC and occur in VNC-Tokens.

[4] We expected the inter-annotator agreement scores would have been at least as high for the test subset as for the development subset, due to the discussion that took place after annotating the development expressions. However, as Table 2 shows, this is not so. The observed agreement for each development expression is above 80%, while for three test expressions this is not the case. For the expressions *have word* and *hold fire* the judges systematically disagreed on the label for one particular sense of each of these expressions. For the expression *make hit*, the low agreement may have been a result of the proportionally large number of questionable usages (see Table 1). Eliminating these three expressions gives an observed agreement and unweighted Kappa score of 89% and 0.83, respectively, for the remaining test expressions.

Subset	Expression	I	L	Q	Total
Dev.	blow trumpet	19	10	11	40
	find foot	48	5	12	65
	get nod	23	3	2	28
	hit road	25	7	17	49
	hit roof	11	7	11	29
	kick heel	31	8	7	46
	lose head	21	19	21	61
	make face	27	14	67	108
	make pile	8	17	3	28
	pull leg	11	40	22	73
	pull plug	45	20	15	80
	pull weight	27	6	17	50
	see star	5	56	9	70
	take heart	61	20	6	87
	Total	362	232	220	814
Test	blow top	23	5	0	28
	blow whistle	27	51	3	81
	cut figure	36	7	1	44
	get sack	43	7	29	79
	get wind	13	16	4	33
	have word	80	11	8	99
	hit wall	7	56	4	67
	hold fire	7	16	8	31
	lose thread	18	2	6	26
	make hay	9	8	11	28
	make hit	5	9	12	26
	make mark	72	13	12	97
	make scene	30	20	15	65
	pull punch	18	4	10	32
	Total	388	225	123	736
Skewed	blow smoke	0	52	3	55
	bring luck	24	0	0	24
	catch attention	100	0	0	100
	catch death	22	1	0	23
	catch imagination	45	0	0	45
	get drift	19	0	11	30
	give notice	95	0	6	101
	give sack	15	3	9	27
	have fling	21	0	0	21
	have future	100	0	0	100
	have misfortune	78	0	0	78
	hold fort	22	0	3	25
	hold horse	2	20	4	26
	hold sway	100	0	1	101
	keep tab	54	1	7	62
	kick habit	40	0	3	43
	lay waste	32	0	1	33
	lose cool	28	0	3	31
	lose heart	51	0	1	52
	lose temper	104	0	0	104
	make fortune	100	0	0	100
	move goalpost	13	2	8	23
	set fire	98	0	3	101
	take root	83	15	1	99
	touch nerve	24	0	6	30
	Total	1270	94	70	1434
All	Total	2020	551	413	2984

Table 1: Number of tokens annotated idiomatic (I), literal (L), and unknown (Q), as well as the total number of tokens (Total), for each expression, grouped by subset of VNC-Tokens.

Set	Observed Agreement (%)	Kappa
Development	89	0.83
Test	78	0.65
Skewed	93	0.67
All	88	0.76

Table 2: Percent observed agreement and unweighted Kappa score for each set.

meaning of the more similar vector. In both DIFF methods, the co-occurrence vector for the idiomatic meaning is created by considering the words in a 5-word window on either side of all canonical form usages of that expression. In this way they obtain an unsupervised, but noisy, estimate of the idiomatic meaning. The two DIFF methods estimate the literal meaning of an expression in differing ways. $\text{DIFF}_{\text{I-CF, L-NCF}}$ approximates the literal meaning using non-canonical form usages in a similar manner to the estimate of the idiomatic meaning. $\text{DIFF}_{\text{I-CF, L-COMP}}$ assumes that a literal VNC usage is compositional, and averages the co-occurrence vectors for each of the component verb and noun in a VNC to estimate its literal meaning.

Cook et al. compare their methods to a baseline which classifies every token as idiomatic. They also compare against a slightly modified version of the supervised method proposed by Katz and Giesbrecht (2006), which classifies a token according to the gold-standard labels of the k nearest tokens according to cosine distance between their co-occurrence vectors. Cook et al. find all three of their unsupervised methods to outperform the baseline of 62% accuracy, with CFORM achieving the highest accuracy of 72%. The CFORM method performs as well as the supervised method with k set to 1; however, using the 5-nearest neighbours in a supervised setting achieves the best performance of 76% accuracy.

Fazly et al. (2008) extend the work of Cook et al. in several ways. Fazly et al. represent the context of a token as the full set of words from the sentence in which it occurs, in an effort to overcome data sparseness problems reported by Cook et al. Consequently, they compare tokens using a set-based similarity measure, Jaccard index. Fazly et al. examine the performance of their methods on all three subset of VNC-Tokens, and present a detailed analysis of their results. They too find CFORM to have the highest unsupervised performance on the test subset. However, their results on the previously-unused skewed subset indicate that their unsupervised method using context outperforms CFORM on expressions that are predominantly used idiomatically.

4. Related Work on Idioms

Two approaches to distinguishing between literal and non-literal tokens have recently been proposed that could be evaluated more extensively using the VNC-Tokens dataset. Katz and Giesbrecht (2006) perform a token-based study of the German expression *ins Wasser fallen* which when used literally means *to fall into water*, but which also has an idiomatic interpretation of *to fail to happen*. They propose a supervised method to distinguish between literal and idiomatic usages of this expression, which is quite similar to,

and in fact was the motivation for, the supervised 1-nearest neighbour method considered by Cook et al. (2007). The main difference between these two approaches is that Katz and Giesbrecht employ singular value decomposition to reduce the dimensionality of the co-occurrence vectors. They evaluate their method on 67 instances of *ins Wasser fallen* found in a corpus of text from a German newspaper, and report an accuracy of 72% on this task which has a baseline of 58%. One of the main shortcomings of this study is that it only presents results for one expression. The VNC-Tokens dataset addresses this by allowing for a more extensive evaluation, although not on German idioms.

Birke and Sarkar (2006) propose a minimally-supervised method for distinguishing between literal and non-literal usages of verbs. Their algorithm relies on seed sets of literal and non-literal usages of verbs that are automatically obtained from readily-available lexical resources. The class of a target verb token is then determined using the similarity between the context of that token and each of the seed sets. Although the annotations in VNC-Tokens are for the combination of a verb and its direct object, it may still be an appropriate resource for evaluating this algorithm. For many expressions in VNC-Tokens, such as *blow the whistle* and *move the goalposts*, the verb is used in a non-literal sense when the VNC is idiomatic, and in a literal sense when the VNC is literal. For other expressions, such as *get the nod* and *make a pile*, this may not be the case depending on the definitions of literal and idiomatic employed—the verb may be contributing a literal meaning even when the VNC it forms with its direct object is idiomatic. Nevertheless, some of the expressions in VNC-Tokens would be appropriate, and would allow for a more extensive evaluation of Birke and Sarkar's algorithm.

Hashimoto et al. (2006) build an unsupervised classifier that exploits manually-encoded lexical knowledge to distinguish between literal and non-literal usages of Japanese idioms, which they evaluate on a relatively small dataset of 309 tokens. However, since their classifier draws on specific properties of Japanese idioms, it is not clear that a more extensive evaluation of their method could be conducted using the English expressions in VNC-Tokens.

5. Future Extensions to VNC-Tokens

While annotating the items in VNC-Tokens, the human judges had access to only the sentence in which a VNC usage occurs (see Section 2.3). This limitation of the annotation process resulted in 413 tokens being assigned the unknown label. Had the annotators had access to more of the surrounding context of each token, far fewer items would have been labelled unknown. As future work, we intend to re-visit those tokens annotated as unknown, and attempt to label them as idiomatic or literal by examining a broader context of their usage.

VNC-Tokens currently consists of at most 100 usages of each of 53 expressions (see Section 2.2). For expressions which occur more than 100 times in the BNC, 100 tokens were randomly selected. VNC-Tokens could be expanded by including additional tokens for these expressions. This would require human effort to annotate the new tokens, but would not be an arduous task as the judges are already familiar with the expressions and the issues involved in their annotation. To expand VNC-Tokens by adding new expressions would be a substantially larger effort. This would require re-running the extraction software and then having human judges annotate the new tokens. Annotating instances of a novel expression would likely be more difficult than annotating new instances of an expression already in VNC-Tokens, as the specific properties of the newly-added expressions may give rise to new annotation issues.

6. Summary

This paper describes the VNC-Tokens dataset, a resource which facilitates research on potentially-idiomatic verb–noun combinations, a productive and common cross-lingual class of MWE. We have described one study which used VNC-Tokens for evaluation, and have shown how two similar studies could also be evaluated more extensively using this resource. Finally, we have identified several ways in which this resource could be expanded in the future.

7. References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-2006*, pages 329–336, Trento, Italy.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic.

Anthony P. Cowie, Ronald Mackin, and Isabel R. McCaig. 1983. *Oxford Dictionary of Current Idiomatic English*, volume 2. Oxford University Press.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of EACL-2006*, pages 337–344, Trento, Italy.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2008. Unsupervised type and token identification of idiomatic expressions. Submitted to *Computational Linguistics*.

Christiane Fellbaum. 2002. VP idioms in the lexicon: Topics for research using a very large corpus. In S. Busemann, ed., *Proc. of the KONVENS 2002 Conference*, Saarbruecken, Germany.

Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006. Detecting Japanese idioms with a linguistically rich dictionary. *Language Resources and Evaluation*, 40(3–4):243–252.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/Coling Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Susanne Riehemann. 2001. *A Constructional Approach to Idioms and Word Formation*. Ph.D. thesis, Stanford.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico.

Maggie Seaton and Alison Macaulay, editors. 2002. *Collins COBUILD Idioms Dictionary*. HarperCollins Publishers, second edition.

Multi-Word Verbs of Estonian: a Database and a Corpus

Heiki-Jaan Kaalep, Kadri Muischnek
University of Tartu
Liivi 2, Tartu, Estonia
E-mail: Heiki-Jaan.Kaalep@ut.ee, Kadri.Muischnek@ut.ee

Abstract

The paper describes two interrelated language resources: a database of 13,000 Estonian multi-word verbs (MWV) and a 300,000 word corpus with annotated MWVs. Both resources have been manually post-edited, and are meant to be used by a wide audience, from corpus linguists to language engineers. The paper gives a short overview of the types of MWVs in Estonian, followed by a description of some grammatical features – word order and inflection – of Estonian and their manifestation in the MWVs. The database is a table that has 13,000 rows and 11 columns and contains information about the source (dictionary or corpus) of the MWV, its linguistic category, frequency in the text corpus, and morphological description. The text corpus contains the morphological analysis of the source text and the annotated MWVs. The layout of the corpus is essentially a table, a row standing for a running word and the columns filled by annotations. The corpus contains 8,200 instances of tagged MWVs and 34,100 simplex main verbs, meaning that roughly every fifth predicate is represented by a MWV. The number of different types of the MWVs in the corpus is 3,500.

1. Introduction

In order to provide an automatic treatment of a language phenomenon, one must first gain a less formal, linguistic understanding of it. Usually it involves making a list of the items one is interested in (morphemes, words, grammar rules etc.) and investigating their behaviour in a real life speech or text corpus. When we are interested in multi-word units (MWU), we face problems that are somewhat similar to those faced by the lexicographers, and researchers interested in morphological analysis and disambiguation: which items should be in the lexicon, how does their form vary, and how do they behave in texts?

Being interested in multi-word verbs (MWV) of Estonian, we have created two interrelated, harmonised resources that complement each other: a database of MWVs and a corpus where the MWVs are tagged. Both of the resources are meant to be used by a wide audience, from corpus linguists to language engineers.

2. Multi-word verbs of Estonian

Estonian belongs to the Finnic group of the Finno-Ugric language family. Typologically it is an agglutinating language. The word order in Estonian reveals remarkable heterogeinity, the written language having tendency towards verb-second pattern. One can find a detailed description of the grammatical system of Estonian in (Erelt, 2003).

In this section we will give a short overview of the types of the Estonian MWVs followed by a brief description of some grammatical features of Estonian posing problems for the automatic treatment of the MWV-s.

In our database we distinguish between the following types of Estonian MWVs:

1. Particle verb (marked *yv* in the database) consisting of an uninflecting particle and a verb (e.g. English *back up*)
2. Expression consisting of a noun (phrase) and a verb (marked as *nv* in the database); could be divided further into idiomatic expressions (e.g. English *kick the bucket*) and collocations (e.g. English *answer the question*).
3. Support verb construction (marked as *sv* in the database) - combinations of a verb and its object (or, occasionally, some other argument), where the nominal component denotes an action of some kind and the verb is semantically empty in this context (e.g. English *take a walk*).
4. Catenative verb construction (marked as *av* in the database) consisting of a verb and an infinitive (e.g. *make do*).

Word order of the MWVs

The heterogenous word order of Estonian means that the components of a MWV can occur in various permutations in a clause and they can be separated from each other by several intervening words as it is the case with the particle verb *üle minema* ‘go over’ in example (1).

(1) *Peavalu **läks** alles järgmisel päeval **üle**.*
Headache go-PST only next-ADE day-ADE over
‘The headache stopped only the next day’

In the examples (2-5) an idiomatic MWV *sõjakirvest välja kaevama* ‘dig out the hatchet, i.e. start the quarrel’ consisting of three components occurs with four different word order variants. In real-life sentences intervening words can occur between all the components of this MWV.

(2) *Jaan **kaevas sõjakirve välja**.*
Jaan-NOM dig-PST hatchet-GEN out
’Jaan started the quarrel’
(3) ***Sõjakirve kaevas välja** Jaan.*
hatchet-GEN dig-PST out Jaan
(4) *Jaan **kaevas välja sõjakirve**.*
Jaan dig-PST out hatchet-GEN
(5) *Kui Jaan **sõjakirve välja kaevas**...*
When Jaan hatchet-GEN out dig-PST

Inflectional variation of the MWVs

Estonian being an agglutinative language means that the verbal component of a MWV inflects freely in texts. In the database it is recorded in its base form and there are

principally two possible ways of matching the database with the texts: either the morphological tagging of the text, or generating all possible forms of the verb in the database.
The non-verbal component of the particle verbs and catenative verbs does not inflect.
However, if a MWV consists of a verb and a NP, the latter may inflect, albeit with various degrees of freedom, which in turn depend on syntactic and semantic features. The rigidness of NPs of MWVs is an important characteristic, and should be recorded accordingly. The MWVs can be divided into subclasses depending on the inflectional behaviour of the nominal component. Among these MWV-s support verb constructions are distinguished as a special subclass. The remaining MWV-s fall into the subclasses of opaque idioms, transparent idioms and collocations. The nominal components of all opaque idiomatic expressions and part of the transparent idiomatic expressions are always frozen in the same case and number and can therefore be treated much like particle verbs in the database.
The flexibility of the nominal components of part of the transparent idioms, most of the support verb constructions and most of the collocations depends on the type of the syntactic relationship with the verb. If the nominal component is formally in the object position of the verb, it can undergo the so-called object case alternations.
Here a few words should be said about the case alternation of the object NP in Estonian in general. Three case forms are possible for the object NP – partitive (both in singular and plural), nominative (singular and plural) and genitive (only singular).
Partitive is the unmarked case form of the object – the 'partial object', as it is often called. The nominative and genitive forms are grouped together under the label 'total object'.
Total object can be found only in an affirmative clause; it cannot be used in a negative clause. The case alternation of the object is used to express the distinction between telic-atelic aspect of the clause. If the verb denotes telic activity (an activity that can have a result), and the activity described in the clause is perfective, then the total object is used:

(6) *Mees ehitas suvilat*
Man built summer-house-PART
'The man was building a summer-house.' (imperfective activity)
(7) *Mees ehitas suvila*
Man built summer-house-GEN
'The man built a summer-house.' (perfective activity)

The nominal components of the transparent idioms are divided 75-25 between the forms of partial object and total object. E.g. in the example (8), the transparent idiom with the nominal component in the form of the total object was used to describe a perfective activity.

(8) *Esinemisele **pani punkti** ilutulestik*.
Show-ALL put period-GEN fireworks
'The fireworks put an end to the show.'

Some of the transparent idioms behave like regular verb-object combinations in this respect, while others show irregular variation, and there are those whose nominal components are frozen in the partitive case. Thus the transparent idioms do not form a homogenous group with respect to the case alternation of the nominal component.
As a practical solution, the information about the variability of the nominal component is recorded separately for each MWV together with the information about the relevant morphological categories (cf sect 3.1).
In support verb constructions, the case alternation of the object is regularly used to express the aspect of the clause:

(9) *Žürii alles **teeb otsust***.
Jury still makes decision-PART
'The jury is still making the decision.' (imperfective)
(10) *Žürii **tegi** lõpuks **otsuse***.
Jury made at-last decision-GEN
'The jury made the decision at last.' (perfective)

Different support verb constructions differ from each other (just like ordinary verbs do) in whether they express an atelic or telic activity. Some support verb constructions are generally used to emphasize the process of the activity (atelic activity), not its result. Such expressions don't normally show case alternation in texts.
In addition to the object case alternations, the nominal components of these three groups can undergo number alternations. Especially the support verb constructions make extensive use of number alternation of the nominal component, whereas the plural form of the noun denoting an action usually refers to several events.

3. Database of MWVs

This database contains multi-word expressions, consisting of a verb and a particle or a verb and its complements. The expressions consisting of a verb and its subject are not included. The multi-word units consisting of a verb and an infinite form of a verb are included irregularly.
The present version of the database contains ca 13,000 expressions.
The database has been compiled on the basis of:
1. Dictionaries and wordlists, aimed at human users, namely:
1.1. Phraseology Dictionary (Õim, 1991),
1.2. The Explanatory Dictionary of Estonian (EKSS 1988-2000),
1.3. Filosoft thesaurus (http://www.filosoft.ee/ thes_et/),
1.4. A list of particle verbs (Hasselblatt, 1990),
1.5. Index of the Thesaurus of Estonian (Saareste, 1979),
1.6. Dictionary of Synonyms (Õim, 1993).
2. The MWVs, extracted automatically from corpora totalling 20 million tokens and post-edited manually. This collocation extraction experiment is described in (Kaalep, Muischnek, 2003).
3. The MWVs found during manual post-editing of the corpus of MWVs (see section 4)

3.1 Database Layout

The database is a table, with every row having 11 fields. The fields are delimited with colons. If a field is empty it means that this information is missing at the moment.

The fields contain the following information:

Field 1
The expression itself. The verbal component of the expression is recorded in the supine form, the traditional form of presenting the Estonian verbs in the dictionaries. As for the expressions consisting of a verb and a noun or a noun phrase, the noun can be 'frozen' in a certain case form or allow certain case alternations. If the nominal component is 'frozen', then it is recorded in the database in this certain case form. If the nominal component can undergo certain case alternations, it is recorded in the database in the partitive case form, but the information about the case alternation is given in the morphological analysis (see field 11).

Field 2
The subtype of the expression. The possible subtypes are:
yv – particle verb
nv – expression consisting of a noun (phrase) and a verb; could be divided further into idiomatic expressions and collocations
tv – support verb construction
av – catenative verb construction

Fields 3-9
Indication that the expression was recorded in a certain dictionary/wordlist and/or was retrieved with collocation extraction methods:
field 3: Phraseology dictionary (Õim, 1991)
field 4: The Explanatory Dictionary of Estonian (EKSS 1988-2000)
field 5: Filosoft thesaurus (http://www.filosoft.ee/thes_et/)
field 6: A list of particle verbs (Hasselblatt, 1990)
field 7: Index of the Thesaurus of Estonian (Saareste, 1979)
field 8: Dictionary of Synonyms (Õim, 1993)
field 9: Automatically extracted collocations

Field 10
If the expression was found and tagged in the corpus of MWVs (see section 4), the number in this field shows the number of its occurrences in the corpus; otherwise, the frequency is zero.

Field 11
Morphological analysis of the expression. This information is needed by programs that tag MWVs in texts: the components of a MWV may be separated by several words, and the form of its components may vary in various ways, depending on the morphosyntactic type of the component and the rigidity of the MWV itself.
The field is delimited by the <morf> and </morf> tags.
The morphological analysis is similar to the one used in the corpus of MWVs.

4. Corpus

A part of a morphologically tagged corpus from http://www.cl.ut.ee/korpused/morfkorpus has been automatically tagged and manually post-edited also for the MWVs. Table 1 shows the composition of the corpus and the number of MWV instances, compared with the number of sentences and simplex main verb instances (auxiliary and modal verbs are excluded from counts). It is worth noting that roughly 20% of all the predicates used in the texts are MWVs.

	tokens	sentences	MWVs	simplex main verbs
fiction	104,000	9,000	3,800	17,000
press	111,000	9,500	2,500	14,500
popular science	98,000	7,300	1,900	12,600
total	313,000	25,800	8,200	34,100

Table 1. Corpus with MWVs tagged.

4.1 Corpus Layout

Here is an example of a sentence '*Nad jätavad ülikooli pooleli*' ('They leave university in-half', i.e. 'They quit the university') containing a MWV '*pooleli jätma*' ('leave in-half', i.e. 'quit'), as it is represented in the corpus:

```
Nad    tema+d //_P_ pl nom //
jätavad    jät+vad //_V_ main indic pres ps3 pl ps af //
                                        #->pooleli jätma#
ülikooli    üli_kool+0 //_S_ com sg gen //
pooleli    pooleli+0 //_D_ //
```

Figure 1: Corpus layout

The text is in 2 columns, delimited by the tabulation character:
1. Wordform and its morphological analysis; this column is actually just a copy from the Morphologically tagged corpus (http://www.cl.ut.ee/korpused/morfkorpus).
2. MWV, surrounded by # and being in a canonical form, i.e. the form used in dictionaries. MWV is situated on the same row with the verbal component. Immediately after the first #, there is an arrow (<- or ->), indicating the direction where the other parts of the MVW are to be found (in our example, the adverb '*pooleli*').
In rare cases, two or more MWVs are tagged on the same row. This happens when the same verb is used in several MWVs at the same time, e.g. *pass out and away.*

4.2 Tagging

Before tagging the MWVs, the corpus had been morphologically analyzed and manually disambiguated (Kaalep, Muischnek 2005). Thus it was possible to automatically tag the candidate MWVs in the texts, according to what could be found in the database of MWVs. It was then the task of a human annotator to select the right expressions, and occasionally to tag new ones, missing from the database and thus having not been tagged automatically. The tagged version was checked by another person, in order to minimize accidental mistakes.

5. Database vs. Corpus

Table 2 serves to compare the lexicon of MWVs based on the corpus with the entries of the DB.

MWV types in the DB	13,000
MWV types in the corpus	3,500
hapax legomena of MWVs in the corpus	2,100

Table 2. MWV types in the DB and corpus.

The small proportion of MWVs of the DB that can be found in real texts (compare rows 1 and 2) may be first explained by the small size of the corpus. The second reason is that the human-oriented dictionaries that were used when building the DB implicitly aimed at showing the phraseological richness of the language and thus contained a lot of idiomatic expressions well known to be rare in real-life texts.

The amount of MWS occurring only once in the entire corpus (*hapax legomena*) deserves some explanation.

From the literature, one may find a number of multiword unit (MWU) or collocation extraction experiments from a corpus that show that the extraction method yields many items, missing from the available pre-compiled lexicons. Some of the items may be false hits, but the authors (whose aim has been to present good extraction methods) tend to claim that a large number of those should be added to the lexicon.

(Evert, 2005) lists a number of authors, who have found that lexical resources (machine readable or paper dictionaries, including terminological resources) are not suitable for serving as a gold standard for the set of MWUs (for a given language or domain). According to (Evert, 2005), manual annotation of MWUs in a corpus would be more trustworthy, if one wants to compare the findings of a human (the gold standard) with those of a collocation extraction algorithm.

In lexicography, we may find a slightly conflicting view: not everything found in real texts deserves to be included in a dictionary. Producing a text is a creative process, sometimes resulting in *ad hoc* neologisms and MWUs that are never picked up and re-used after the final full stop of the text they were born in.

Unfortunately these two conflicting views mean that there is no general, simple solution for the problem of finding a gold standard for automatic treatment (extraction or tagging) of MWUs. It is normal that there is a discrepancy between a stand-alone lexicon and the vocabulary of a text.

6. Conclusion

This paper described two interrelated language resources: a database of Estonian multiword verbs and a corpus where these expressions are tagged.

The umbrella term “multiword verbs” covers particle verbs, support verb constructions and expressions consisting of a verb and a noun phrase. The latter category encompasses idiomatic expressions as well as collocations.

The database of MWV-s, based on the data of dictionaries as well as collocations extracted from text corpora, contains various types of linguistic information for ca 13,000 expressions.

A corpus of 300,000 words has been tagged for these MWV-s, indicating that roughly one in five predicates is represented by a MWV.

A closer look at the database and corpus indicates that the criteria for selecting MWUs to be included in a database or tagged in a corpus, might actually be in need of reconsideration, taking into account the experience from the field of lexicography.

7. Acknowledgements

The work on tagging MWVs has been supported in 2004-2007 by the national programs “Estonian Language and National Culture” and “Language technology for Estonian”, and by the Estonian Science Foundation.

8. References

EKSS (1988-2000) *Eesti kirjakeele seletussõnaraamat* (A-Žüriivaba). ETA KKI, Tallinn

Erelt, M. (editor) (2003) Estonian Language. *Linguistica Uralica Supplementary Series vol 1*. Estonian Academy Publishers, Tallinn.

Evert, S. (2005) *The statistics of word cooccurrences : word pairs and collocations*. URL: http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/

Filosoft - Tesaurus. http://www.filosoft.ee thes_et/

Hasselblatt, C. (1990) Das Estnische Partikelverb als lehnübersetzung aus dem Deutschen. Wiesbaden

Kaalep, H-J. Muischnek, K. (2003) Inconsistent Selectional Criteria in Semi-automatic Multi-word Unit Extraction. In *COMPLEX 2003, 7th Conference on Computational Lexicography and Corpus Research*, Ed. By F. Kiefer, J.Pajzs, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, pp. 27--36

Kaalep, H-J, Muischnek, K. (2005) The corpora of Estonian at the University of Tartu: the current situation. *Proceedings of the Second Baltic Conference on Human Language Technologies*. Institute of Cybernetics, Tallinn University of Technology. Institute of the Estonian Language. Editors: Margit Langemets, Priit Penjam. Tallinn: 267-272

Saareste, A. (1979) *Eesti keele mõistelise sõnaraamatu indeks*. Finsk-ugriska institutionen, Uppsala

Tael, K. (1988) *Sõnajärjemallid eesti keeles (võrrelduna soome keelega)*. Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut. Preprint KKI-56

Õim, A. (1993) Fraseoloogiasõnaraamat. ETA KKI, Tallinn

Õim, A. (1991) Sünonüümisõnastik. Tallinn

Abbreviations

ADE – adessive case
ALL – allative case
GEN – genitive case
PART – partitive case
PST – past tense

A French Corpus Annotated for Multiword Nouns

Éric Laporte, Takuya Nakamura, Stavroula Voyatzi
Université Paris-Est
IGM-Labinfo
5, Boulevard Descartes, Champs-sur-Marne
77454 Marne-la-Vallée Cedex 2 (France)

E-mail: eric.laporte@univ-paris-est.fr, nakamura@univ-mlv.fr, voyatzi@univ-mlv.fr

Abstract

This paper presents a French corpus annotated for multiword nouns. This corpus is designed for investigation in information retrieval and extraction, as well as in deep and shallow syntactic parsing. We delimit which kind of multiword units we targeted for this annotation task; we describe the resources and methods we used for the annotation; and we briefly comment on the results. The annotated corpus is available at *http://infolingu.univ-mlv.fr/* under the LGPLLR license.

1. Introduction

Recognizing multiword nouns such as *groupes de pression* 'lobbies' in texts is useful for information retrieval and extraction because of the information that such nouns can convey. In particular, in specialized languages, most of the technical and terminological information is concentrated in multiword nouns. In addition, such recognition is likely to help resolving prepositional attachment during shallow or deep parsing: some multiword nouns contain internal prepositional phrases, and in many cases, recognising them rules out analyses where they are complements of verbs, adjectives or other nouns (Blanc *et al*., 2007). In the case of English, the same is true for the analysis of noun sequences (Vadas & Curran, 2007).

The quality of the recognition of multiword nouns depends on algorithms, but also on resources. We created a corpus of French texts annotated with multiword nouns. This corpus is freely available on the web with LGPLLR license. In this article, we survey related work, we define the target of our annotation effort, we describe the method implemented and we analyse the corpus obtained.

2. Related work

Many problems related with the notion of multiword expression (MWE) in general have been studied by linguists and lexicologists (e.g. Downing, 1977; Sag *et al*., 2001; Girju, 2005; as regards French multiword nouns: Silberztein, 1993), but textual resources annotated for MWEs are still rare and small. In the Grace corpus (Rajman *et al*., 1997), most MWEs are ignored. In the French Treebank (Abeillé *et al*., 2003), multiword nouns are annotated as such. We are not aware of other available French corpora annotated with multiword nouns. In other languages, including English, corpora annotated with MWEs are rare and small as well. In the Penn Treebank (Marcus *et al*., 1993), even such frozen nouns as *stock market* are not annotated as MWEs. Subirats & Sato (2004) report an experiment of annotating MWUs, including multiword nouns, in a Spanish corpus, and Mota *et al*. (2004) and Ranchhod (2005) in a Portuguese corpus, but the resulting annotated corpora are not publicly available. The recognition of multiword nouns is essential to identifying meaningful units in texts, and the availability of a larger corpus of annotated text is likely to shed light on the problems posed by this task.

3. Target of annotation

The target of our annotation effort is defined by the intersection of two criteria: (i) multiword expressions and (ii) nouns. In this section, we define both criteria in more detail, we define the features that we included in the annotations, and we describe the corpus. More details are provided in the guidelines which are available along with the corpus.

3.1 The multiword unit criterion

For this work, we considered a phrase composed of several words to be a multiword expression if some or all of their elements are frozen together in the sense of Gross (1986), that is, if their combination does not obey productive rules of syntactic and semantic compositionality. In the following example, *assemblée générale* ('annual general meeting', lit. 'general assembly') is a multiword noun:

(1) *Notre **assemblée générale** se tiendra vendredi*
'Our **annual general meeting** will be held on Friday'

This criterion ensures a complementarity between lexicon and grammar. In other words, it tends to ensure that any combination of linguistic elements which is licit in the language, but is not represented in syntactic-semantic grammars, will be stored in lexicons.

Syntactic-semantic compositionality is usually defined as follows: a combination of linguistic elements is compositional if and only if its meaning can be computed from its elements. This is also our conception. However, in this definition, we consider that the possibility of computing the meaning of phrases from their elements is of any interest only if it is a better solution than storing the same phrases in lexicons, i.e. if they rely on grammatical rules with sufficient generality. In other words, we consider a combination of linguistic elements to be compositional if and only if its meaning can be computed

from its elements **by a grammar**. In example (1) above, the lack of compositionality is apparent from distributional restrictions such as:

* *Notre **assemblée partielle** se tiendra vendredi*
(*'Our **annual partial meeting** will be held on Friday')

The point is that this blocking of distributional variation (as well as other syntactic constraints) cannot be predicted on the basis of general grammar rules and independently needed lexical entries. Therefore, the acceptable combinations are meaning units and have to be included in lexicons as multiword lexical items.

We annotated multiword named entities (NE) denoting places, institutions, events etc. The status of named entities with respect to compositionality is not fully consensual. We complied with the usual view that, since they follow quite specific grammatical rules, they should be considered as MWEs. However, we did not tag person names consisting of a combination of one or several first names and possibly a last name, e.g. *Gordon Brown*.

We tagged multiword nouns of functions and titles, unless they have the form *N-role de Det N-institution*, where *N-institution* is a noun denoting an institution, *Det* is a determiner, and *N-role* is a noun denoting a role assumed by a member of this institution. We consider this construction as compositional. For example, in *président de l'Assemblée nationale* 'president of the National Assembly', only *Assemblée nationale* falls in the target of our annotation task, but in *ministre de l'Économie et des Finances* 'minister of Economy and Finance', the whole phrase does.

3.2 Delimitation

The general rule to determine the delimitation of an occurrence of a multiword noun is that all and only the elements frozen with the rest of the expression should be included.

Consequently, a sequence of words should not be tagged as a multiword noun when it is included in a larger MWE. For example, the verbal idiom *aller dans le bon sens* 'be a step in the right direction' apparently includes the multiword noun *bon sens* 'good sense', but only apparently, and does not fall in the target of our annotation task. Thus, annotating multiword nouns involves analysing sentences and detecting whether frozen nominal sequences are included in larger frozen units. Such larger frozen units may be verbal idioms, as in the example above, or belong to other types, such as frozen prepositional phrases, e.g. *sur le pied de guerre* 'on a war footing', *au grand jour* 'in broad daylight; in the open', *d'un bout à l'autre de* 'from one end to the other of'. In these phrases, *pied de guerre*, *grand jour* or *bout à l'autre* should not be tagged.

When a multiword place name contains a noun denoting the type of place, we considered this noun to be a part of the multiword. For example, the nouns *océan* 'ocean' and *rue* 'street' are not included in multiword nouns when they occur in *Cet océan grisâtre l'émouvait* 'That greyish ocean moved her' or *La rue de mon enfance est de l'autre côté de l'église* 'The street of my childhood is on the other side of the church', but we analyse *océan Atlantique* and *rue de la Paix* as proper nouns.

When a multiword noun is employed with a support verb, as in *Le nouveau président donne un coup de pied dans la fourmilière*, 'The new president is kicking the anthill', the resulting construction is usually classified among multiword expressions. However, we consider that the support verb, here *donne* 'gives', is not frozen, since the noun can occur without this verb with the same meaning, as opposed to what happens with a verbal idiom. Thus, in this example, *coup de pied dans la fourmilière* 'kick in the anthill' falls in the target of our annotation task.

When a multiword noun is coordinated with another one and appears as reduced because a common part is factored, we tagged it as if it were not reduced. For example, in *accidents ferroviaires et aériens* 'rail and air crashes', the noun *accidents* 'crashes' is factored; therefore, we tagged *accidents ferroviaires* 'rail crashes' on its own, and *aériens* with the same tags as if it had the form of *accidents aériens* 'air crashes', which produces the following form[1]:

<N fs='NA:mp'>accidents ferroviaires</N> et <N fs='NA:mp'>aériens</N>

When the factored part is in the plural only because of the factoring, we tagged the multiword nouns in the singular. For example, in *les océans Atlantique et Pacifique*, both *océans Atlantique* and *Pacifique* were marked as singular.

The rules above do not apply when the whole coordination is frozen, as in *ministère de l'économie et des finances* 'ministry of Economy and Finance', which is recognizable by the impossibility to permute the coordinated parts (there is no *ministère des finances et de l'économie* 'ministry of Finance and Economy').

3.3 The noun criterion

We annotated only expressions belonging to the noun part of speech. We recognized them through the usual criteria regarding their morphosyntactic context.

Many quotations behave as nouns or names. We considered they should be tagged if they are used as titles of works. For example, the quoted sequence should be tagged in *"Autant en emporte le vent" est un film de 1939* '"Gone with the wind" is a 1939 film', but not in *Et il répondit : "Pas encore"* 'And he answered:"Not yet"'.

3.4 Features

Two types of features were included in the annotations.

(i) Each occurrence of a multiword noun was assigned a subcategory among a closed list of 13. The definition of the subcategories is based on internal morphosyntactic structure, i.e. surface constituency of the internal structure of the multiword nouns. They were described as sequences of parts of speech and syntactic categories. For example, *opinion publique* 'public opinion' is assigned a subcategory identified by the mnemonic acronym *NA*, and defined as a noun followed by an adjectival phrase. The 13 subcategories are listed in Table 1.

When a multiword noun did not strictly match any of these structures, annotators were requested to select the closest

[1] For the XML notation, see the section 3.4.

structure. For instance, *agence nationale des travailleurs d'outre-mer* 'national agency for overseas workers' is assigned the *NDN* structure, in spite of the adjective *nationale*. In case of a coordination of prepositional phrases, the multiword noun is classified as if there were only one of them: *ministre de l'emploi, de la cohésion sociale et du logement* 'minister of employment, social cohesion and housing' is assigned the *NDN* structure.

Acronym	Definition	Examples
AN	Noun with a preposed adjectival phrase or numerical determiner	*premier ministre, 35 heures*
NA	Noun with a postposed adjectival phrase	*opinion publique*
NN	Sequence of two nouns, including borrowed nouns such as *business*	*assurance-vie, pôle environnement, show-business*
VV	Sequence of two verb forms	*savoir-faire*
XV	Verb form, with a non-verb preposed modifier	*bien-être, pis-aller*
VN	Verb followed by a noun	*porte-monnaie, faire-part*
PN	Preposition followed by a noun	*après-midi, sous-traitance*
XN	Word of another category (borrowed word, prefix...) followed by a noun	*plus-value, mi-temps, stock-option*
NDN	Noun followed by a prepositional phrase with the preposition *de*	*code du travail, bien de première nécessité*
NAN	Noun followed by a prepositional phrase with the preposition *à*	*gaz à effet de serre, rappel au règlement*
NPN	Noun followed by a prepositional phrase with a preposition other than *de* or *à*	*étranger en situation irrégulière, violence contre la personne*
AAN	Noun with two preposed adjectives (coordinated or not)	*petites et moyennes entreprises*
NAA	Noun with two postposed adjectives (coordinated or not)	*produit intérieur brut, conseil économique et social*

Table 1: Morphosyntactic subcategories of multiword nouns

(ii) Inflectional features (gender and number) were also encoded in the compact form of *:ms*, *:mp*, *:fs* and *:fp*.
The syntax of the encoding follows the XML language. All features are included in an *fs* attribute, as in *<N fs='AN:ms'>Premier ministre</N>* 'prime minister'.

3.5 The corpus

The corpus we annotated comprises:
(i) the complete minutes of the sessions of the French National Assembly on October 3-4, 2006, transcribed into written style from oral French (hereafter *AS*) [2];
(ii) Jules Verne's novel *Le Tour du monde en quatre-vingts jours*, 1873 (hereafter *JV*).
Errors (e.g. *mis enoeuvre* for *mis en oeuvre* 'implemented') have not been corrected. Statistics on the corpus are displayed in Table 2.

	size (Kb)	sentences	words (tokens)	words (types)
corpus *AS*	824	5 146	98 969	18 028
corpus *JV*	1 231	3 648	69 877	19 828
whole corpus	2 055	8 794	168 846	37 856

Table 2: Size of the corpus

4. Methodology

In order to annotate the corpus, we tagged the occurrences of the multiword nouns described in a morphosyntactic lexicon, following the same method as Abeillé *et al.* (2003), Subirats & Sato (2004), Mota et al. (2004) and Ranchhod (2005); we revised the annotation manually.

4.1 The lexicon

We used the same morphosyntactic lexicon as Abeillé *et al.* (2003), so that the two corpora can be used jointly for further research. This lexicon, Delacf (Silberztein, 1990), covers the inflected forms of 100 000 lemmas. It is freely available [3] for research and business with the LGPLLR license. It is the fruit of long-term work on the basis of conventional dictionaries, corpora and introspection (Gross, 1986).

4.2 Tagging

We tagged the corpus with the Unitex platform[4] (Paumier, 2006). We used transducers in order to tag the recognized sequences with morphosyntactic features.

4.3 Manual revision

The annotation was manually validated by three experts. This validation followed guidelines, which are available along with the corpus. It involved two operations.
(i) The sequences tagged with the aid of the lexicon and Unitex were checked in order to detect cases in which the recognized sequence is in fact a part of a larger MWU. For instance, when *court terme* 'short term' occurred within the multiword adverb *à court terme* 'in the short term', the tags around *court terme* were deleted. When *ministre de l'intérieur* 'ministry of interior' occurred within the complete title *ministre de l'intérieur et de l'aménagement du territoire* 'ministry of interior and territory development', the end tag *</N>* after *intérieur* was shifted to the end of the complete title. Cases of coordinated multiword nouns (cf. section 3.2) were processed

[2] http://www.assemblee-nationale.fr/12/documents/index-rapports.asp.
[3] http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Dictionnaires/downloads.html
[4] http://igm.univ-mlv.fr/~unitex.

manually during this operation.
(ii) The text was integrally reviewed in search for multiword nouns absent from the lexicon, and thus undetected by Unitex, e.g. *passage à l'euro* 'euro changeover' or *Passe de Cheyenne* 'Cheyenne Pass'.
The experts had meetings during the annotation process in order to make it consistent. In the end, one of them reviewed the annotated corpus entirely for consistency.

5. Results

The resulting corpus is annotated with 5 054 occurrences of multiword nouns. Table 3 displays their distribution in function of the parts of the corpus and of the subcategories based on morphosyntactic structures. The percentages correspond to membership in the subcategories.

Struct.	JV corpus	JV (%)	AS corpus	AS (%)
AN	131	11.2	206	5.2
NA	206	18.7	1393	35.3
NN	267	24.2	211	5.3
VV	1	0.1	4	0.1
XV	0	0.0	4	0.1
VN	8	0.7	18	0.5
PN	11	1.0	24	0.6
XN	142	12.9	63	1.6
NDN	322	29.2	1639	41.5
NAN	7	0.6	160	4.0
NPN	6	0.5	186	4.1
AAN	1	0.1	18	0.5
NAA	1	0.1	25	0.6
Total	1103	100.0	3951	100.0

Table 3 : No. of occurrences of multiword nouns by subcategory

6. Conclusion

This paper described the annotation of a French corpus for multiword nouns. Two types of features are included in the annotations: internal morphosyntactic structure and inflectional features. This annotated corpus can be used jointly with the French Treebank (Abeillé *et al*., 2003) for research on information retrieval and extraction, automatic lexical acquisition, as well as on deep and shallow syntactic parsing.

7. Acknowledgment

This task has been partially financed by CNRS and by the Cap Digital business cluster. We thank Anne Abeillé for making the French Treebank available to us.

8. References

Abeillé, A., Clément, L., and Toussenel F. (2003). Building a Treebank for French. In A. Abeillé (Ed.), *Building and Using Parsed Corpora*, *Text, Speech and Language Technology*, 20, Kluwer, Dordrecht, pp. 165-187.

Blanc, O., Constant, M., Watrin, P. (2007). "Segmentation in super-chunks with a finite-state approach", *Proceedings of the Workshop on Finite State Methods for Natural Language Processing*, Potsdam.

Downing, P. (1977). On the Creation and Use of English Compound Nouns. *Language* 53(4), pp. 810-842.

Girju, R. et al. (2005). On the semantics of noun compounds. *Computer Speech and Language*, 19, pp. 479-496.

Gross, M. (1986). Lexicon-Grammar. The representation of compound words. In *Proceedings of the Eleventh International Conference on Computational Linguistics*, Bonn, West Germany, pp. 1-6.

Marcus, M., Santorini, B., Marcinkiewicz, M.A. (1993). "Building a large annotated corpus of English: the Penn Treebank", *Computational Linguistics* 19(2), pp. 313-330.

Merlo, P. (2003). Generalised PP-attachment Disambiguation using Corpus-based Linguistic Diagnostics. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, pp. 251-258.

Mota, C. Carvalho, P. Ranchhod, E. (2004). Multiword Lexical Acquisition and Dictionary Formalization. In Michael Zock (ed.), *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, COLING, Geneva, pp. 73-76.

Paumier, S. (2006). *Unitex Manual*. Université Paris-Est. http://igm.univ-mlv.fr/~unitex/manuel.html.

Rajman, M., Lecomte, J., Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Rapport GRACE GTR-3-2.1.

Ranchhod, E. (2005). "Using Corpora to Increase Portuguese MWE Dictionaries. Tagging MWE in a Portuguese Corpus". In: *Proceedings from The Corpus Linguistics Conference Series, Vol. 1, no. 1, Corpus Linguistics 2005*.

Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In Gelbuk (ed.), *Computational Linguistics and Intelligent Text Processing: Third International Conference CICLing 2002*, Springer-Verlag, Heidelberg/Berlin, pp. 1--15.

Silberztein, M. (1990). "Le dictionnaire électronique des mots composés". *Langue Française* 87, pp. 71-83, Paris : Larousse.

Silberztein, M. (1993). "Les groupes nominaux productifs et les noms composés lexicalisés", *Lingvisticae Investigationes* 17:2, Amsterdam/Philadelphia: Benjamins, pp. 405-426.

Subirats, C., Sato, H. (2004). Spanish FrameNet and FrameSQL. *4th International Conference on Language Resources and Evaluation. Workshop on Building Lexical Resources from Semantically Annotated Corpora, Lisbon (Portugal), May 2004*.

Vadas, D., Curran, J.R. (2007). " Adding Noun Phrase Structure to the Penn Treebank", In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 240–24.

An Electronic Dictionary of French Multiword Adverbs

Éric Laporte, Stavroula Voyatzi
Université Paris-Est
IGM-Labinfo
5, Boulevard Descartes, Champs-sur-Marne
77454 Marne-la-Vallée Cedex 2 (France)
E-mail: eric.laporte@univ-paris-est.fr, voyatzi@univ-mlv.fr

Abstract

We present an electronic dictionary of French multiword adverbs. This dictionary is designed for investigation on information retrieval and extraction, automatic lexical acquisition, as well as on deep and shallow syntactic parsing. We delimit the scope of the dictionary in terms of lexical coverage and of grammatical coverage, we outline the formal description of entries, and we give an overview of the syntactic and semantic features which are associated to the 6,800 adverbial entries of the lexicon. This electronic dictionary is freely available on the web.

1. Introduction

Recognising multiword adverbs such as *à long terme* 'in the long run' in texts is likely to be useful for information retrieval and extraction because of the information that some of these adverbials convey. In addition, it is likely to help resolving prepositional attachment during shallow or deep parsing: most multiword adverbs have the superficial syntax of prepositional phrases; in many cases, recognising them rules out attachments where they are analysed as arguments or noun modifiers.

In the current practices of natural language processing, the handling of multiword expressions (MWEs) in general is in its infancy. Much research effort towards MWE recognition is devoted to algorithms, but results depend also on resources. We describe an electronic dictionary of multiword adverbs of French with syntactic-semantic information. This dictionary is freely available on the web under LGPLLR license. In this article, we survey related work, we define the scope of the dictionary, we present the syntactic and semantic features assigned to entries and we describe their representation.

2. Related research

A considerable amount of research has been conducted in the area of MWEs, e.g. general studies (Sag *et al*., 2002) and efforts towards standardization (Calzolari *et al*., 2002), but they seldom rely on large-coverage lexical resources[1]. Michiels and Dufour, (1998) exploit conventional dictionaries (i.e. written for human readers), but such resources have well-known inherent limitations. However, there do exist NLP-oriented lexicons with a large coverage in MWEs, including multiword adverbs, e.g. WordNet (Miller, 1995). Lexicological research focusing on multi-word adverbs has been devoted to French (Gross, 1990), German (Seelbach, 1990), Spanish (Blanco & Català, 1998/1999), Italian (De Gioia, 2001), Portuguese (Baptista, 2003), Korean (Jung, 2005) and Modern Greek (Voyatzi, 2006) with the Lexicon-Grammar methods of NLP-oriented lexicon design (Gross, 1986; 1994), on the basis of conventional dictionaries, grammars, corpora and introspection[2]. Català and Baptista (2007) show that multiword adverbs are recognized in Spanish text with 77% precision through the use of a Lexicon-Grammar.

In parallel, research on automatic lexical acquisition was targeted both at terminology (Daille, 2000) and general-language MWEs. Such techniques use both statistical approaches and linguistic information, such as parts of speech and inflectional categories, and require large corpora that contain significant numbers of occurrences of MWEs. However, even with corpora of millions of words, frequencies of MWEs are usually too low for statistical extraction (Mota *et al*., 2004). Gross (1986) reports that the number of MWEs in the lexicon of a language is larger than the number of single words (cf. also Jackendoff, 1997), therefore any extraction method must be able to handle extremely sparse data. In addition, adverbs or more generally non-object complements have not been the focus of attention, and their relations to simple sentences are far from being understood[3].

[1] In practice, paradoxically, even investigation in semi-automatic extension of MWE lexicons pays little attention to the structure and contents of existing large-coverage lexicons (e.g. Navigli, 2005). Copestake *et al.* (2002) contains interesting thoughts, but they are not validated against an available large-coverage lexicon and it does not deal with adverbials.

[2] The resulting resources on French, enclosed in the Intex system (Silberztein, 1994), have helped to annotate the French Treebank (Abeillé *et al.*, 2003), in which prepositional phrases and adverbs are annotated with a binary feature ('compound') which indicates whether they are multiword units; the distinction between whether prepositional phrases are verb modifiers, noun modifiers or objects appears only in the function-annotated part of the Treebank (350,000 words).

[3] Several reasons explain this lack of interest. Firstly, adverbials are usually felt as less useful than nouns for information retrieval and extraction. Secondly, many multiword adverbs are difficult to distinguish from prepositional phrases assuming other syntactic functions, such as arguments or noun modifiers: the distinction is hardly correlated to any material markers in texts

The availability of large-coverage lexicons of multiword adverbs is essential to gaining insight on their recognition, including the dual problems of variability and ambiguity. The resource described in this paper is the Lexicon-Grammar of French multiword adverbs (Gross, 1990), in which previously implicit features have been made explicit for more convenient use in NLP.

3. Scope of lexicon

The scope of the lexicon is delimited by the intersection of two criteria: (i) multiword expressions and (ii) adverbial function. In this section, we define both criteria in more detail and we present the features provided in the lexicon.

3.1 The multiword unit criterion

For this work, a phrase composed of several words is considered to be a multiword expression if some or all of its elements are frozen together, that is, if their combination does not obey productive rules of syntactic and semantic compositionality. In the following example, *de nos jours* ('nowadays', lit. 'of our days') is a multiword unit assuming an adverbial function:

(1) *Il est facile* ***de nos jours*** *de s'informer*
'It is easy to get informed **nowadays**'

This criterion ensures a complementarity between lexicon and grammar. In other words, it tends to ensure that any combination of linguistic elements which is licit in the language, but is not represented in syntactic-semantic grammars, will be stored in lexicons.

Syntactic-semantic compositionality is usually defined as follows: a combination of linguistic elements is compositional if and only if its meaning can be computed from its elements. This is also our conception. However, in this definition, we consider that the possibility of computing the meaning of phrases from their elements is of any interest only if it is a better solution than storing the same phrases in lexicons, i.e. if they rely on grammatical rules with sufficient generality. In other words, we consider a combination of linguistic elements to be compositional if and only if its meaning can be computed from its elements **by a grammar**. In example (1) above, the lack of compositionality is apparent from distributional restrictions[4] such as:

* *Il est facile de nos* ***semaines*** *de s'informer*
* 'It is easy to get informed nowa**weeks**'

and by the impossibility of inserting modifiers that are a priori plausible, syntactically and semantically:

**de nos jours (de repos + de fête)*
literally 'of our days (of rest + of feast)'

pendant nos jours (de repos + de fête)
literally 'during our days (of rest + of feast)'

MWEs include many different subtypes, varying from entirely fixed expressions to syntactically more flexible expressions (Sag *et al*., 2002). In (2), the possessive adjective agrees obligatorily in person and number with the subject of the sentence:

(2) ***De (ses + *mes) propres mains****, il a construit une maison en torchis*
'**With (his + *my) own hands**, he built a house in cob'

The lexicon also takes into account expressions which comprise a frozen part and a free part, e.g. *au moyen de ce bouton* 'with the aid of this switch'. The frozen part *au moyen de* 'with the aid of' is encoded in the lexicon, and the syntactic category of the free part, here *NP*, is encoded as a feature[5]. Open classes of multiword adverbs such as named entities (NEs) of date or duration are not included in the dictionary, since they follow quite specific syntactical rules and use a closed lexicon. They can be identified with FST methods (Martineau *et al*., 2007).

3.2 The adverbial function criterion

The dictionary deals only with MWEs which can assume an adverbial role, i.e. circumstantial complements, or complements which are not objects of the predicate of the clause in which they appear. They are identified through criteria (Gross, 1986; 1990) involving the fact that they are optional, they combine freely with a wide variety of predicates and some of them pronominalize with specific forms. Phrases with adverbial function are often called 'circumstantial complements', 'adverbials', 'adjuncts', or 'generalised adverbs'. They assume several morphosyntactic forms: underived (*demain* 'tomorrow') or derived adverbs (*prochainement* 'soon'), prepositional phrases (*à la dernière minute* 'at the last minute') or circumstantial clauses (*jusqu'à ce que mort s'ensuive* 'until death comes'), and special structures in the case of NEs of time (*lundi 20* 'on Monday 20') (cf. section 3.1).

3.3 The features

French multiword adverbs have been assigned a feature describing their internal morphosyntactic structure. The definition of the morphosyntactic structures is based on the number, category and position of the frozen and free components of the adverbial. They are described as a sequence of parts of speech and syntactic categories. For example, *à la nuit tombante* 'at nightfall' is assigned a structure identified by the mnemonic acronym *PCA*, and defined as *Prép Dét C (MPA) Adj*, where *C* stands for a noun frozen with the rest of the adverbial, *Adj* for a post-posed noun modifier (e.g. an adjectival phrase or a relative clause), and *MPA* for a pre-adjectival modifier, empty in this lexical item. The 15 structures, together with an illustrative example and the corresponding number of entries are listed in Table 1.

and lies in complex linguistic notions (Villavicencio, 2002; Merlo, 2003).

[4] The point is that this blocking of distributional variation (as well as other syntactic constraints) cannot be predicted on the basis of general grammar rules and independently needed lexical entries. Therefore, the acceptable combinations are meaning units and have to be included in lexicons as multiword lexical items.

[5] In case of a limited set of possibilities, all of them are listed in independent entries, as in *au sens propre (du mot + du terme + de l'expression)* 'in the proper sense of the (word + term + phrase)'.

Struct.	Example	English equivalent	Size
PC	*par exemple*	for example	664
PDETC	*de nos jours*	nowadays	848
PAC	*à la dernière minute*	at the last minute	776
PCA	*à la nuit tombante*	at nightfall	840
PCDC	*dans la limite du possible*	as far as possible	750
PCPC	*à cent pour cent*	one hundred percent	287
PCONJ	*tôt ou tard*	sooner or later	333
PCDN	*à l'insu de NP*	unbeknowst to NP	555
PCPN	*en comparaison avec NP*	in comparison with NP	151
PV	*à dire vrai*	to tell the truth	285
PF	*jusqu'à ce que mort s'ensuive*	until death comes	396
PECO	*<fidèle> comme un chien*	as <faithful> as a dog	305
PVCO	*<travailler> comme un chien*	<work> as much as a dog	338
PPCO	*<disparaître> comme par enchantement*	<vanish> as by enchantement	50
PJC	*mais aussi et surtout*	but also and foremost	185
		Total	6,763

Table 1: Morphosyntactic structures of multiword adverbs

Examples of other syntactic-semantic features provided in the lexicon are (i) the conjunctive function of the adverbial in discourse, (ii) the omission of the pre-adjectival modifier *MPA* without loss of information, or (iii) the constraint that the adverbial obligatorily occurs in a negative clause (cf. section 4 and table 2).

4. The Electronic Dictionary

The electronic dictionary of French multiword adverbs has 6,800 entries. It is freely available [6] for research and business under the LGPLLR license. It takes the form of a set of Lexicon-Grammar tables (or binary matrices) such as that of Table 2, which displays a sample of the lexical items with the *PCA* morphosyntactic structure.

	N0 = Nhum	N0 = N-hum	Nég obl		Prép	Dét	C	Modif pré-adj	Adj	Prép Dét C	Prép Dét MPA Adj C	Prép MPA Adj C	Conjonction
170	+	-	-	<E>	agir	dans	les	délais	les plus	brefs	-	-	+	-
171	+	-	-	<E>	agir	dans	les	délais	les plus	courts	-	-	+	-
172	+	-	-	<E>	agir	dans	les	délais	les	meilleurs	-	-	+	-
173	+	-	-	<E>	rire	<E>	toutes	dents	<E>	dehors	-	-	-	-
174	-	+	-	se	produire	à	cette	époque-	<E>	ci	+	-	-	-
175	-	+	-	se	produire	à	cette	époque-	<E>	là	+	-	-	+

Table 2: Sample of the table of entries with the *PCA* morphosyntactic structure

In this table, each row describes a lexical item, and each column corresponds:
- either to one of the elements in the morphosyntactic structure of the items (columns with identifiers 'Prép', 'Dét', 'C', 'Modif pré-adj' and 'Adj');
- or to a syntactic-semantic feature (cf. 3.3); these columns hold binary values: 'Conjonction', 'Prép Dét C', 'Nég obl';
- or to illustrative information provided as an aid for the human reader to find examples of sentences containing the adverbial (e.g. columns D and E giving an example of a verb compatible with the adverb).

There are 15 such tables, one for each of the morphosyntactic structures.

4.1 The General Table

A lexicon is not a static resource: it has to be updated with the evolution of language. In order to facilitate the manual maintenance of the lexicon by linguists, the following organization has been adopted. When the values of a syntactic or semantic feature are the same over all entries in a class, it is not displayed in the corresponding class table. We stored it in a General Table (Figure 3).

Table	Prép2	Dét2	C2	Conj	Ppv	V	ConjS	Dét0	C0	comme	de Dimd façon C-a	C-a-ment	C-a	Conjonction	Prép Dét C	Prép Dét MPA Adj C	Prép MPA Adj C	Prép1 Dét1 C1	N2=:Nhum	N2=:N-hum	C1 de N2=Poss2 C1
PC	-	-	-	-	-	-	-	-	-	-	o	o	o	o	-	-	-	-	-	-	-
PDETC	-	-	-	-	-	-	-	-	-	-	o	o	o	o	+	-	-	-	-	-	-
PAC	-	-	-	-	-	-	-	-	-	-	-	-	-	o	o	-	+	-	-	-	-
PCA	-	-	-	-	-	-	-	-	-	-	-	-	-	o	o	o	o	-	-	-	-
PCDC	de	o	o	-	-	-	-	-	-	-	-	-	-	-	-	-	-	o	-	-	-
PCPC	o	o	o	-	-	-	-	-	-	-	-	-	-	-	-	-	-	o	-	-	-
PCONJ	o	o	o	o	-	-	-	-	-	-	-	-	-	-	-	-	-	o	-	-	-
PCDN	de	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	O	o	o	o
PCPN	o	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	o	o	o	-
PV	-	-	-	-	o	o	-	-	-	-	-	-	-	-	-	-	-	-	O	O	-
PF	-	-	-	-	o	o	o	o	o	-	-	-	-	-	-	-	-	-	O	O	-
PECO	-	-	-	-	o	-	-	-	-	comme	-	-	-	-	-	-	-	-	-	-	-
PVCO	-	-	-	-	o	o	-	-	-	comme	-	-	-	-	-	-	-	-	-	-	-
PPCO	-	-	-	-	o	o	-	-	-	comme	-	-	-	-	-	-	-	-	-	-	-
PJC	-	-	o	o	-	-	-	-	-	-	-	-	-	O	-	-	-	-	-	-	-

Figure 3: Sample of General Table of multiword adverbs

The rows correspond to the morphosyntactic structures of multiword adverbs. All the 29 features described in any of the 15 tables are represented in the columns of the general table. Moreover, it also takes into account 12 features that had not been encoded in any of the 15 tables (for example, features connected with the morphosyntactic structures), totaling 41 features, all described in our documentation available with the lexicon. Values used at the intersection of rows and columns indicate that the feature in the column:
- is encoded in the table associated to the row; its value is variable (noted 'o');
- is encoded in the table associated to the row; its value is constant (noted <value>, e.g. Prép2='*de*');
- is not encoded in the table associated to the row; if it were encoded, its value would be constant (noted <value>, e.g. '+' or '-');
- is not encoded in the table associated to the row; if it were encoded, its value would be variable (noted 'O').

5. Conclusion

This paper described the design of an electronic lexicon of

[6] http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Lexiques-Grammaires/View.html.

French multiword adverbs which comprise 6,800 fixed, semi-flexible and flexible combinations, all of them associated with appropriate morphosyntactic and semantic features. This electronic dictionary is freely available on the web for research on information retrieval and extraction, automatic lexical acquisition, as well as on deep and shallow syntactic parsing.

6. Acknowledgements

This task has been partially financed by CNRS and by regional business cluster Cap Digital.

7. References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a Treebank for French. In A. Abeillé (Ed.), *Building and Using Parsed Corpora*, *Text, Speech and Language Technology*, 20, Kluwer, pp. 165--187.

Baptista, J. (2003). Some Families of Compound Temporal Adverbs in Portuguese. In *Proceedings of the Workshop on Finite-State Methods for Natural Language Processing*, *EACL*, Budapest, pp. 97--104.

Blanco, X., Català, D. (1998/1999). Quelques remarques sur un dictionnaire électronique d'adverbes composés en espagnol. *Lingvisticae Investigationes* 22, pp. 213--232.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C. and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, *LREC 2002*, Las Palmas, pp. 1934--1940.

Català, D., Baptista, J. (2007). Spanish Adverbial Frozen Expressions. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, ACL 2007*, Prague, Czech Republic, pp. 33--40.

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I., Flickinger, D. (2002). Multiword Expressions: Linguistic Precision and Reusability, In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, Las Palmas, pp. 1941--1947.

Daille, B. (2000). Morphological rule induction for terminology acquisition. In *Proceedings of the 18th International Conference on Computational Linguistics*, *COLING'00*, Saarbrucken, Germany, pp. 215--221.

De Gioia, M. (2001). *Avverbi idiomatici dell'italiano. Analisi lessico-grammaticale, prefazione di Maurice Gross*, Torino, L'Harmattan.

Gross, M. (1986). Lexicon-Grammar. The representation of compound words. In *Proceedings of the 11th International Conference on Computational Linguistics*, *COLING'86*, Bonn, West Germany, pp. 1--6.

Gross, M. (1990). *Grammaire transformationnelle du français: 3. Syntaxe de l'adverbe*. Paris, ASSTRIL.

Gross, M. (1994). Constructing Lexicon-Grammars. In Atkins & Zampolli (Eds.), *Computational Approaches to the Lexicon*, Oxford University Press, pp. 213--263.

Jackendoff, R. (1997). *The architecture of the Language Faculty*. Cambridge, MA, MIT Press.

Jung, E. J. (2005). *Grammaire des adverbes de durée et de date en coréen*. Thèse de doctorat en Informatique Linguistique. Université Paris-Est Marne-la-Vallée.

Martineau, C., Tolone, E., Voyatzi, S. (2007). Les Entités Nommées : usage et degrés de précision et de désambiguïsation. In *Proceedings of the 26th International Conference on Lexis and Grammar*, Bonifacio, pp. 105--112.

Merlo, P. (2003). Generalised PP-attachment Disambiguisation using Corpus-based Linguistic Diagnostics. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, *EACL*, Budapest, pp. 251--258.

Michiels, A. and Dufour, N. (1998). DEFI, a tool for automatic multi-word unit recognition, meaning assignment and translation selection. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, *LREC*, Granada, pp. 1179--1186.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38:11, pp. 39--41.

Mota, C. Carvalho, P. Ranchhod, E. (2004). Multiword Lexical Acquisition and Dictionary Formalization. In Michael Zock (ed.), In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, *COLING 04*, Geneva, pp. 73--76.

Navigli, R. (2005). Semi-Automatic Extension of Large-Scale Linguistic Knowledge Bases, In *Proceedings of the Florida AI Research Society Conference*, pp.548--553.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbuk (Ed.), *Computational Linguistics and Intelligent Text Processing: Proceedings of the Third International Conference*, *CICLing 2002*, Springer-Verlag, Heidelberg/Berlin, pp. 1--15.

Seelbach, D. (1990). Zur Entwicklung von bilingualen Mehrwortlexica Französisch-Deutsch-Stützverbkon struktionen und adverbiale Ausdrücke. *Lexicon und Lexikographie* 11, pp. 179--207.

Silberztein, M. (1994). INTEX: a corpus processing system. In *Proceedings of the 15th International Conference on Computational Linguistics*, *COLING 94*, Kyoto, Japan, pp. 579--583.

Villavicencio, A. (2002). Learning to distinguish PP arguments from adjuncts. In *Proceedings of the 6th Conference on Natural Language Learning*, *CoNLL 2002*, Taipei, Taiwan, pp. 84--90.

Voyatzi, S. (2006). *Description morphosyntaxique et sémantique des adverbes figés en vue d'un système d'analyse automatique des textes grecs*. Thèse de doctorat en Informatique Linguistique. Université Paris-Est Marne-la-Vallée.

Cranberry Expressions in English and in German

Beata Trawiński[*], Manfred Sailer[#], Jan-Philipp Soehn[*], Lothar Lemnitzer[†] and Frank Richter[†]

[*]University of Tübingen
SFB 441
Nauklerstraße 35
D-72074 Tübingen
trawinski@sfs.uni-tuebingen.de
jp.soehn@uni-tuebingen.de

[#]University of Göttingen
Department of English Studies
Käte-Hamburger-Weg 3
D-37073 Göttingen
manfred.sailer@phil.uni-goettingen.de

[†]University of Tübingen
Department of Linguistics (CL)
Wilhelmstraße 19
D-72074 Tübingen
lothar@sfs.uni-tuebingen.de
fr@sfs.uni-tuebingen.de

Abstract

We describe two data sets submitted to the database of MWE evaluation resources: (1) cranberry expressions in English and (2) cranberry expressions in German. The first package contains a collection of 444 cranberry words in German (`CWde.txt`) and a collection of the corresponding cranberry expressions (`CCde.txt`). The second package consists of a collection of 77 cranberry words in English (`CWen.txt`) and a collection of the corresponding cranberry expressions (`CCen.txt`). The data included in these packages was extracted from the Collection of Distributionally Idiosyncratic Items (CoDII), an electronic linguistic resource of lexical items with idiosyncratic occurrence patterns. Each package contains a `readme` file, and can be downloaded from `multiword.wiki.sourceforge.net/Resources`.

1. Background and Motivation[0]

The original impetus for compiling the present data[1] came from research into the relationship between the regular syntactic and semantic combinatorial system of grammar and irregular, or exceptional, lexical items. Some expressions, such as reflexive pronouns and negative polarity items, appear quite freely in sentences as long as certain occurrence requirements are fulfilled – there must be an appropriate antecedent or a negation, respectively. While those items specify their occurrence requirements in terms of grammatical notions, there is another group of items, cranberry words, that require the presence of a specific lexeme. A typical cranberry word is *sandboy*, which can only occur as part of the expression *happy as a sandboy*. These items are of particular interest in our research on distributional idiosyncrasies, but they are also of interest for the study of multiword expressions in general, as we will explain below. In Section 2, important properties of cranberry expressions and their position between idioms and collocations will be discussed. In Section 3, the linguistic resource will be presented from which the sets of cranberry expressions were extracted. Section 4 will provide a few statistical details on the collected expressions. In Section 5, the potential of the described data sets for computational lexicography and information extraction will be outlined. Section 6 will summarize the discussion.

2. Cranberry Expressions

Cranberry Expressions (CE) are multiword expressions which contain an item that is not found in the language outside this expression. This item is called a *Cranberry Word* (CW) in (Aronoff, 1976), in analogy to "cranberry morph". Alternatively CWs are also called *(phraseologically) bound words* or *unique words* (German: *Unikalia*).

The repertoire of CEs in German and English is well documented in the literature on idioms. (Dobrovol'skij, 1988) contains the most exhaustive list of CEs in German, English and Dutch. Emphasizing the difference between bound and free words, (Dobrovol'skij, 1988) and (Dobrovol'skij and Piirainen, 1994) provide criteria for classifying CEs and the expressions in which they occur. In fact, it is not always clear whether an item should count as a CW or not. For example, the noun *Abstellgleis* (*holding track*) is in our list of CWs because it usually occurs in the CE *jn. aufs Abstellgleis stellen/schieben* (literally: *put so. on the holding track*). This expression receives the metaphorical interpretation *to put so. on inactive reserve* or *to deprive so. of his/her influence*. In contrast to the constituents of typical idioms, the word *Abstellgleis* stems from a technical domain (railway systems) and is not used in everyday language outside the CE.

Dobrovol'skij and Piirainen estimate the number of CEs in German at 600. They classify 180 as belonging to the common vocabulary of native speakers. At present, we have included 444 potential CEs in our collection. For English, (Dobrovol'skij, 1988) lists about 100 items, 77 of which are included. The leading criterion for recording an item was whether it was discussed as a candidate of containing a CW within the phraseological literature. In the CoDII resource, sketched in Section 3 below, we document the linguistic classifications and properties of the CEs.

CEs take a middle position between idioms (such as *spill the beans*) and collocations (such as *take a shower*). Due to their restricted occurrence CWs fulfill the criterion of lexical fixedness typically found with idioms (*spill the peas* is not a variant of *spill the beans*). However, in contrast to typical idioms, there is no (synchronically used) literal meaning. CEs share with collocations a linguistically significant co-occurrence of the CW with the other components of the

[0]We would like to thank Janina Radó for comments and suggestions concerning the content and style of this paper.

[1]The data packages originate from (i) project A5, *Distributional Idiosyncrasies* (2002–2008), of the Collaborative Research Center SFB 441 (*Linguistic Datastructures*) at the University of Tübingen, funded by the German Research Foundation (DFG), `www.sfb441.uni-tuebingen.de/a5/index-engl.html`, and (ii) the linguistics section of the English Department of the University of Göttingen.

CE. However, in the case of CWs this is not a question of preference but a hard restriction.

These differences notwithstanding, some CEs should be grouped with idioms, others with collocations. The idiom-like CEs show an idiomatic interpretation of their non-CW components. They also manifest a small range of possible modifications, and the expression as a whole can be assigned one non-decomposable meaning. This meaning can be indicated by a synonym, an antonym or a paraphrase, cf. *Schiffbruch erleiden*, synonym: *scheitern* ('to fail'); *die Spendierhosen anhaben*, synonym: *großzügig sein* ('to be generous'), antonym: *geizig sein* ('to be thrifty'). The collocation-like CEs have a literal interpretation of the non-CW components. They are also structurally parallel to collocations (CE: *make headway*, *happy as a sandboy*; collocation: *make progress*, *dark as night*). Sometimes the CW is interchangeable. A typical example is *Tacheles/Klartext/Fraktur reden* ('to state sth. clearly and with some force'). The range of interchangeable words is always rather small.

CEs comprise a wide variety of syntactic categories (VP: *make headway*; PP: *on tenterhooks*; AP: *happy as a sandboy*; NP: *the whole caboodle*). Similarly, CWs are of all major syntactic categories (V: *wend one's way*; A: *spick and span*; N: *run the gamut*). They also cover different frequency classes:[2] The German CW *Anhieb* (in *auf Anhieb* ('*right away*')) is of frequency class 12 (i.e. the most frequent German word is 2^{12} times more frequent), the CW *Kattun* (in *jm. Kattun geben* ('*to reprimand so.*')) is of frequency class 21.

The reported properties indicate that, at least in German and English, while defined on the distributional properties of one component, CEs comprise instances of a great number of types of the multiword expressions in the language.

3. The Collection of Distributionally Idiosyncratic Items (CoDII)

The data packages we present here were extracted from the Collection of Distributionally Idiosyncratic Items. CoDII is an electronic multilingual resource for lexical items with idiosyncratic occurrence patterns. It was originally designed to provide an empirical basis for linguistic investigations of these items. CoDII compiles and lists items of interest, providing linguistic documentation and corpus evidence, and specifying possibilities for extracting more context data for the items in the collection. When we created CoDII, we were concerned with two kinds of expressions: (i) negative and positive polarity items as expressions whose distribution is grammatically restricted, and (ii) cranberry expressions as expressions whose distribution is restricted by lexical co-occurrence patterns. Design and data structure of CoDII have been conceived in such a way that subcollections of various types of distributionally idiosyncratic items can be modeled (such as anaphora, negative and positive polarity items, and cranberry words), and collections of distributionally idiosyncratic items from various languages can be integrated.

[2]The frequencies are taken from the data of the project *Deutscher Wortschatz* at the University of Leipzig, `wortschatz.uni-leipzig.de`.

Five collections of distributionally idiosyncratic items are currently available in CoDII: CWs in German, CWs in English, Negative Polarity Items in Romanian, Negative Polarity Items in German, and Positive Polarity Items in German. The collections of cranberry words are based on (Dobrovol'skij, 1988; Dobrovol'skij, 1989) and (Dobrovol'skij and Piirainen, 1994), and are described in (Sailer and Trawiński, 2006). The resources for polarity items are described in (Trawiński and Soehn, To appear).

Each CoDII entry contains the following information blocks: General Information (including glosses and translations, if appropriate, as well as the expression in which the item occurs together with a set of possible paraphrases of this expression), Classification, Syntactic Information (including syntactic variations) and, optionally, search patterns. For the syntactic annotation of German and English items, the *Stuttgart-Tübingen Tagset* (STTS) and the syntactic annotation scheme from the Syntactically Annotated Idiom Database (SAID) were used, respectively. For each context, appropriate examples are provided from various corpora, the Internet and the linguistic literature.

CoDII is encoded in XML and is freely accessible on the Internet at `www.sfb441.uni-tuebingen.de/a5/codii`. A fragment of the XML encoding of the English CW *sandboy* in the CoDII format is presented in Figure 1. The elements `dii` and `dii-expression`, `dii-classification`, `dii-syntax` and `dii-queries` model the information blocks specified above.

```
<dii-entry id="sandboy">
 <dii><ol>sandboy</ol></dii>
 <dii-expression>
   <ol>happy as a sandboy</ol>
   <ol-paraphrase>very happy</ol-paraphrase>
 </dii-expression>
 <dii-classification>
   <dii-class category="bw"
              class="dekompo"
              type="A5">
   <bibliography bib-item="A5"/>
   </dii-class>
    [...]
   <dii-class category="bw"
              type="Dobro88"
              class="gebWB">
   <bibliography bib-item="Dobrovolskij88"/>
   </dii-class>
 <dii-syntax cat="NN"
             hits="sandboy01 [...] sandboy02">
   <dii-expression-syntax cat="AdjP">
      [AP[AP[Ahappy]][COMPas][NP[DETa][NP[Nsandboy]]]]
   </dii-expression-syntax>
 </dii-syntax>
 <dii-queries>
    <query type="google" hits="sandboy01">
      <query-text>
          "happy as a sandboy"
      </query-text>
    </query>
 </dii-queries>
</dii-entry>
```

Figure 1: The CoDII-XML-encoding of *sandboy*

CoDII not only compiles, documents and (alphabetically) lists distributionally idiosyncratic items, it also offers dynamic and flexible access. Taking advantage of the theoretically grounded internal data structure and an annotation scheme which involves syntactic and (partial) semantic information, a comfortable interface for querying the

database was created with the Open Source XML database eXist (`exist.sourceforge.net/`). At present, possible search criteria comprise lemmas, syntactic properties, and classifications. Searching for expressions with particular licensing contexts is also possible. With these tools, the two data sets which are presented here as pure (alphabetically ordered) lists of expressions can be modified and enriched if this is necessary for a particular task.

Several other projects have constructed resources for idiomatic expressions. These projects differ from CoDII by the corpora used, the kind of data and the applied methods. The project *Usuelle Wortverbindungen* (Conventionalized Word Combinations, URL: `www.ids-mannheim.de/ll/uwv/`) of the Institut für Deutsche Sprache (IDS) (Steyer, 2004) starts from statistically highly frequent words which undergo a co-occurrence analysis. It only uses the corpora of the IDS. In contrast to this collection, CoDII is based on linguistic intuitions and theoretical considerations and includes data from different sources. The project *Kollokationen im Wörterbuch* (Collocations in the Lexicon, URL: `kollokationen.bbaw.de`) of the Berlin-Brandenburgische Akademie der Wissenschaft (Fellbaum et al., 2005) is based on the corpus *Das digitale Wörterbuch der deutschen Sprache*. Like CoDII, the project starts with idioms from phraseological literature, but focuses exclusively on German VP idioms. For English, the *Syntactically Annotated Idioms Database* (SAID) encodes the syntactic structure of a large number of idioms (Kuiper et al., 2003), but it contains no other information about the expressions.

4. Some Details on the Collected CEs

As noted above, CWs are of all major syntactic categories. However, the overwhelming majority of German CWs are nouns (80%, e. g. *jn. beim Schlafittchen packen*, 'to take so. by the scruff of the neck'), followed by predicative adjectives (7%, e. g. *sattsam bekannt*, 'widely known'), proper names (5%, e. g. *Büchse der Pandora*, 'Pandora's box'), and verbs (3%, e. g. *alles, was da kreucht und fleucht*, 'everything that crawls and flies'). VPs (83%) are the most common syntactic environment for (the typically nominal) CWs in German CEs. In 87 cases (20%) a CW is the complement of a specific preposition. These "unique nominal complements" form an important subclass of CWs (e. g. ***auf** Anhieb*, 'right away' or ***on** tenterhooks*, cf. (Soehn and Sailer, 2003)). From a theoretical point of view, these data provide excellent evidence that non-heads, including complements, can impose restrictions on the heads they combine with.

English CWs reveal a different pattern. Although the most common category is again nouns (67%, e. g. *at first blush*), the second most common one is attributive adjectives (21%, e. g. *curule chair*). Predicative adjectives and verbs play only a minor role with 7% and 4%, respectively. The leading syntactic category of CEs is not VP (31%) but NP (41%). This is a consequence of the fact that free nouns form compounds with bound nouns. Compounding is a morpho-lexical process which works differently in English and in German: English compounds consist of several orthographic words which are categorized as multi-word expressions (NPs). In German compounds form one orthographic unit. The difference leads to many English NPs with bound nouns; additional bound adjectives in NPs further increase their frequency.

5. Our Data Sets and Other Resources

Our cranberry expression data sets for English and German are a valuable resource for the documentation of a special aspect of these languages as well as an empirical base for investigations into multi-word expressions. However, we believe that one should think beyond these applications and explore how these data can a) inform the development of other lexical resources and b) be useful for data-driven information extraction experiments.

The information provided in our data sets goes well beyond the mere listing of the CEs and includes semantic glosses which contain synonyms, antonyms and examples. Linking those CEs which behave like non-decomposable idioms to semantically related lexical items, i. e. to their synonyms, antonyms, and hypernyms, would make it possible to enrich other lexical resources.[3] Many CEs in our collection contain links to semantically related words through their paraphrases or glosses. Admittedly, there is a wide variety of glosses and not all of them consist of a single lexically related word. Nevertheless, they are a good starting point for creating more explicit lexical-semantic relations.

Moreover, systematically connecting the CEs with the English and German wordnets would benefit both resources: a) Wordnet users gain access to an interesting set of multi-word expressions. These are currently underrepresented in wordnets, and in particular in GermaNet; b) on the other hand, the CEs of our collections would be embedded into broader semantic fields, e. g. the CE *aufs Abstellgleis schieben* would be related to the verb *abschieben* and its hypernym, other hypernyms of this verb etc. (Lüngen et al., 2008) present an approach of linking general language wordnets with specialized lexical resources which can also be applied to our resources.

Regarding the use of our data for information extraction purposes, the CEs and their lexical-semantic neighbors – either in the glosses or through the links to wordnets – become an interesting resource for the training of methods for extracting semantically related lexical items from corpora, which is an active research area in the field of lexical acquisition, cf. (Hendrickx et al., 2007; Snow et al., 2006). It is comparatively easy to collect a set of contexts, in the form of concordances, for CWs because most of them are unambiguous or at least occur most often in (semi-)fixed contexts. From the concordance, context vectors can be derived and abstracted which represent the distributional characteristics or "fingerprint" of these lexical items. This can be the basis for a comparison with other, semantically related, lexical items. Currently, there are corpus citations for only a few CEs in the collection. For some of them, however, we provide search patterns which can be applied to a corpus to extract a larger set of examples.

Our collection of idiom-like CEs is suitable as training material in yet another lexical acquisition task. It is well-

[3] One might also consider including those collocation-like CEs as well whose meaning can be mapped to one single concept.

known that many idioms may undergo a range of mainly syntactic variations and internal modifications (cf. (Nunberg et al., 1994), and for German (Lemnitzer and Kunze, 2007, chapter 11) and (Soehn, 2006)). Rules and methods for the automatic detection, annotation, and extraction of idioms must take this variability into account. As said before, it is relatively easy to build a collection of examples for our CEs. As they represent several types of multiword expressions, the data can be used to capture variations and modifications in idiomatic expressions and help to acquire and / or fine tune these rules.

6. Summary and Outlook

Cranberry expressions are multiword expressions with special properties that make them interesting for the theoretically oriented linguist as well as for use as an electronic resource for lexical acquisition, and for evaluation and extraction tasks. In this paper we presented two resources, a list of 444 cranberry words in German and a list of 77 cranberry words in English, accompanied by corresponding lists of cranberry collocations in which the cranberry words occur.

What makes CEs interesting is their middle position between idiomatic expressions and collocations. Their special property is the obligatory occurrence of a unique cranberry word in each CE. Once a cranberry word has been identified, the obligatoriness of this lemma and its categorial and robust lexical occurrence restriction makes the exhaustive retrieval of its CEs from corpora and the Internet much easier and much more reliable than the retrieval of idiomatic expressions in general. It follows that well-documented CEs with particular properties may be good candidates for a gold standard in a retrieval task for otherwise similar multiword expressions without CWs. It is at this point that the middle position of CEs between idioms and collocations becomes particularly interesting, since it opens the door for using appropriate subclasses of CEs in both research contexts. The additional documentation of our CEs in CoDII, comprising linguistic classification and access by means of various search categories, further enhances the usefulness of the resource.

7. References

Mark Aronoff. 1976. *W*ord Formation in Generative Grammar. MIT Press, Cambridge, Massachusetts and London, England.

Dmitrij Dobrovol'skij and Elisabeth Piirainen. 1994. Sprachliche Unikalia im Deutschen: Zum Phänomen phraseologisch gebundener Formative. *F*olia Linguistica, 27(3–4):449–473.

Dmitrij Dobrovol'skij. 1988. *Phraseologie als Objekt der Universallinguistik*. Verlag Enzyklopädie, Leipzig.

Dmitrij Dobrovol'skij. 1989. Formal gebundene phraseologische Konstituenten: Klassifikationsgrundlagen und theoretische Analyse. In Wolfgang Fleischer, Rudolf Große, and Gotthard Lerchner, editors, *Beiträge zur Erforschung der deutschen Sprache*, volume 9, pages 57–78. Leipzig, Bibliographisches Institut.

Christiane Fellbaum, Undine Kramer, and Gerald Neumann. 2005. Corpusbasierte lexikographische Erfassung und linguistische Analyse deutscher Idiome. In Annelies Häcki Buhofer and Harald Burger, editors, *Phraseology in Motion I. Methoden und Kritik. Akten der Internationalen Tagung zur Phraseologie (Basel, 2004)*, pages 183–199. Schneider Verlag, Hohengehren.

Iris Hendrickx, Roser Morante, Caroline Sporleder, and Antal van den Bosch. 2007. ILK: Machine learning of semantic relations with shallow features and almost no data. In *P*roceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 187–190, Prague, Czech Republic, June. Association for Computational Linguistics.

Koenraad Kuiper, Heather McCann, Heidi Quinn, Therese Aitchison, and Kees van der Veer. 2003. Syntactically annotated idiom database (SAID) v.1. Documentation to a LDC resource.

Lothar Lemnitzer and Claudia Kunze. 2007. *C*omputerlexikographie. Gunter Narr Verlag, Tübingen.

Harald Lüngen, Claudia Kunze, Lothar Lemnitzer, and Storrer Angelika. 2008. Towards an Integrated OWL Model for Domain-Specific and General Language WordNets. In *P*roceedings of the Fourth Global WordNet Conference, pages 281–296. University of Szeged, Hungary, Department of Informatics.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *L*anguage, 70:109–132.

Manfred Sailer and Beata Trawiński. 2006. The Collection of Distributionally Idiosyncratic Items: A Multilingual Resource for Linguistic Research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 471–474, Genoa, Italy.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *A*CL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pages 801–808, Morristown, NJ, USA. Association for Computational Linguistics.

Jan-Philipp Soehn and Manfred Sailer. 2003. At First Blush on Tenterhooks. About Selectional Restrictions Imposed by Nonheads. In Gerhard Jäger, Paola Monachesi, Gerald Penn, and Shuly Wintner, editors, *P*roceedings of Formal Grammar 2003, pages 149–161.

Jan-Philipp Soehn. 2006. ¨Uber Bärendienste und erstaunte Bauklötze – Idiome ohne freie Lesart in der HPSG. Phd dissertation (2005), Friedrich-Schiller-Universität Jena.

Kathrin Steyer. 2004. Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikographische Perspektiven. In Kathrin Steyer, editor, *Wortverbindungen – mehr oder weniger fest*, pages 87–116. de Gruyter, Berlin and New York.

Beata Trawiński and Jan-Philipp Soehn. To appear. A Multilingual Database of Polarity Items. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.

Standardised Evaluation of English Noun Compound Interpretation

Su Nam Kim[◇♠], **Timothy Baldwin**[♠]

National University of Singapore[◇]
Department of Computer Science, School of Computing, Singapore
kimsn@comp.nus.edu.sg
and
University of Melbourne[♠]
Department of CSSE, Carlton, Victoria, Melbourne
{snkim,tim}@csse.unimelb.edu.au

Abstract

We present a resource for English noun compound interpretation and describe the method used to generate them. In order to collect noun compounds, we extracted binary noun compounds (i.e. noun-noun pairs) by looking for sequences of two nouns in the POS tag data of the Wall Street Journal component of the Penn Treebank 2.0. We then manually filtered out all noun compounds which were incorrectly tagged or included proper nouns. This left us with a data set of 2,169 noun compounds, which we annotated using a set of 20 semantic relations defined by Barker and Szpakowicz (1998) allowing the annotators to assign multiple semantic relations if necessary. The initial agreement was 52.31%. The final data set contains 1,081 test noun compounds and 1,088 training noun compounds.

1. Introduction

Noun compounds (or NCs), such as *computer science* and *paper submission*, have received significant attention in the linguistic and computational linguistic literature over the past couple of decades, particularly in the area of interpreting the semantic relations between a head noun and its modifier(s). We define **noun compounds** as sequences of nouns contained in a single NP, where the rightmost noun is the NP head.

Semantic relations (or SRs) are directed binary predicates which represent the nature of the semantic link between the component nouns of an NC. For example, the semantic relation for *orange juice* is MATERIAL, indicating that the modifier, *orange* is the material from which the *juice* (head noun) is made. On the other hand, *morning juice* is of type TIME, indicating that the *juice* has some temporal significance (i.e. *morning*). Since NCs are both highly productive and semantically underspecified (Lapata, 2002), the task of interpreting them is not easy. Moreover, the interpretation can differ depending on contextual and pragmatic factors.

Interpreting SRs has been undertaken in the past from both linguistic and computational perspectives. Defining possible semantic relations in noun compounds has been studied and proposed from different perspectives (Levi, 1979; Finin, 1980; Vanderwende, 1994; Saghdha and Copestake, 2007). One approach has been to propose a predefined number of SRs to interpret NCs (Levi, 1979; Sparck Jones, 1983), while a second has been to suggest that there is an unbounded set of SRs, and propose a context-sensitive means of interpretation (Finin, 1980). The main issues here are the granularity and coverage of the SRs, and distribution of different SRs in a given dataset (Saghdha and Copestake, 2007). Indeed, Saghdha (2007) recently devised a set of semantic relations based on the core notation of ease/reproducibility of annotation, attempting to produce a set of discrete semantic relations with a relative uniform token distribution. Since the set of SRs directly influences the automatic interpretation task, it is necessary to agree on a standard set of SRs. Unfortunately, however, this is still under active debate.

On the other hand, recent approaches to automatic interpretation have achieved success to a certain degree based on supervised learning approaches (Moldovan et al., 2004; Kim and Baldwin, 2005; Kim and Baldwin, 2006; Nastase et al., 2006; Nakov and Hearst, 2006; Girju, 2007). The two main basic supervised approaches to the interpretation of NCs have been semantic similarity (Moldovan et al., 2004; Kim and Baldwin, 2005; Girju, 2007) ellipsed predicate recovery (Kim and Baldwin, 2006; Nakov and Hearst, 2006). In addition, the workshop on semantic evaluation 2007 (SemEval-2007) provided a standardised (sub)set of SRs and NC instances over which to compare approaches to NC interpretation. Although the problem was restricted to binary classification (i.e. is a given NC compatible with a given SR or not), it provided a great opportunity to gather many ideas and clearly showcased promising approaches for future study. Moreover, it allowed researchers to understand problems such as the impact of label and training data bias on interpretation (Girju et al., 2007).

Our goal in this work is to outline a standardised data set for noun compound interpretation, which we hope will complement the SemEval-2007 dataset in furthering empirical research on NC interpretation.

In the following sections, we describe where and how we obtained the NCs in the dataset (Section 2.), describe semantic relations used in the data set (Section 3.), describe the annotation procedure (Section 4.), and summarise our contribution (Section 5.).

2. Data

In this section, we describe the data collection process. We will also look at the statistics of the data set and data format, including examples.

2.1. Data Collection

First, we selected a sample set of NCs based on the gold-standard part-of-speech (POS) tags in the Wall Street Jour-

	Total	Test	Train
Total no. of NCs	2169	1081	1088
No. of (NC,SR) pairs	2347	1163	1184
No. of NCs with multiple SRs	178	82	96

Table 1: Composition of the data set

nal component of the Penn Treebank 2.0, by retrieving all binary noun compounds (i.e. noun-noun pairs). That is, if two nouns appeared contiguously at least once, we tagged that combination as an NC candidate. Second, we excluded NCs that contained proper nouns such as country names, or names of people/companies from our data set (e.g. *Apple computer, Melbourne Cup*). We then manually filtered out any false positives from the remaining data. For example, in the sentence .. *was a library students used* .. we initially retrieved *library students* based on the simple POS information, which we later excluded on the basis of not occurring in a single NP.[1]

2.2. Data Split and Annotation

The total number of NCs in our data set is 2,169. We split the data into approximately 50% for test and the remainder for training in the interests of providing a standardised means of experimenting over the dataset. The number of test and training NCs are 1,081 and 1,088, respectively.
In order to annotate the data set, we hired two human annotators and trained them over 200 held-out NCs to familiarise them with the annotation task and set of SRs (see Section 4.). The SRs used for the annotation were taken from Barker and Szpakowicz (1998), as detailed in Section 3..
The annotators were instructed to tag with a unique SR where possible, but also that multiple SRs were allowed in instances of genuine ambiguity. For example, *cable operator* can be interpreted as corresponding to the SR TOPIC (as in *operator is concerned with cable(s)*) or alternatively OBJECT (as in *cable is acted on by operator*). On completion of the annotation, the two annotators were instructed to come together and resolve any disputes in annotation. In the final dataset, about 8.2% of the NCs were tagged with multiple semantic relations.
We present a breakdown of the final dataset in Table 1, in terms of the total number of NCs, the number of discrete pairings of NC and SR, and the number of noun compounds which have multiple SRs.

2.3. Data Format

The data format is simple, consisting of a single NC and its SR(s) per line. The nouns are space delimited, and the SRs are tab-delimited from the NC. In the files, the NCs are sorted in ascending order. An excerpt of the file is listed in Table 2.

[1]Note that an alternative approach would have been to use the parse trees in the Penn Treebank to determine if a phrase boundary occurs between the words. We chose the manual approach so as to have a data collection procedure which has maximal applicability, i.e. makes as few assumptions about the original data source as possible.

```
chest pain   source
computer expert topic
printer tray cause
student loan object
student protest agent
```

Table 2: Data sample

3. Semantic Relations

To annotate our NCs, rather than attempting to define a new set of SRs, we used the set defined by Barker and Szpakowicz (1998). In detail, the authors defined 20 SRs for NCs and provided definitions and examples for each. Later, Nastase et al. (2006) classified these into 5 super-classes for their own usage. The SRs are detailed in Table 3., along with the number of test and training instances containing in the data set. Note that the SRs were developed for a more general class of data than our NCs, including adjective-noun compounds . Hence, some of examples contain adjective as modifiers (e.g. *charitable compound* for BENEFICIARY and *late supper* for TIME).
In Table 3., the final column details the number of NCs tagged with each SR in the test and training data sets, respectively. The numbers in parentheses indicate the number of instances for each subset of the data that are tagged with multiple relations.

4. Annotation

We describe the annotation methodology in this section. First, we briefly describe our human annotators. Second, we outline the annotator training procedure. Finally, we describe the annotation procedure in detail.

4.1. Human Annotators

Our human annotators were two PhD students. One is an English native speaker with some experience in computational linguistics, and the other (the first author) is a non-native English speaker with wide experience in computational linguistics and various annotation tasks.

4.2. Training the Annotators

To train our human annotators, we collected 200 noun compounds not contained in our final data set. The training procedure started with an introduction to the set of semantic relations in Barker and Szpakowicz (1998) that enabled them to understand the semantic relations and differences between them. We then made the annotators independently tag the 200 held-out NCs, and had them come together to compare their annotations. In the case of disagreements, they discussed their respective annotations and tried to agreed upon a unique SR. However, when they could not come up with a unique mutually-agreeable SR, they were allowed to assign both SRs. We also allowed them to individually assign multiple SRs in instances of genuine ambiguity, such as *cotton bag*, which can be interpreted as either MATERIAL (*bag made of cotton*) or PURPOSE (*bag for cotton*). Note that in practice, context can disambiguate the

Relation	*Definition*	*Example*	*Test/training instances*
AGENT	N_2 is performed by N_1	*student protest, band concert, military assault*	10(1)/5(0)
BENEFICIARY	N_1 benefits from N_2	*student price, charitable compound*	10(1)/7(1)
CAUSE	N_1 causes N_2	*printer tray, flood water, film music, story idea*	54(5)/74(3)
CONTAINER	N_1 contains N_2	*exam anxiety, overdue fine*	13(4)/19(3)
CONTENT	N_1 is contained in N_2	*paper tray, eviction notice, oil pan*	40(2)/34(2)
DESTINATION	N_1 is destination of N_2	*game bus, exit route, entrance stairs*	1(0)/2(0)
EQUATIVE	N_1 is also head	*composer arranger, player coach*	9(0)/17(1)
INSTRUMENT	N_1 is used in N_2	*electron microscope, diesel engine, laser printer*	6(0)/11(0)
LOCATED	N_1 is located at N_2	*building site, home town, solar system*	12(1)/16(2)
LOCATION	N_1 is the location of N_2	*lab printer, desert storm, internal combustion*	29(9)/24(4)
MATERIAL	N_2 is made of N_1	*carbon deposit, gingerbread man, water vapour*	12(0)/14(1)
OBJECT	N_1 is acted on by N_2	*engine repair, horse doctor*	88(6)/88(5)
POSSESSOR	N_1 has N_2	*student loan, company car, national debt*	33(1)/22(1)
PRODUCT	N_1 is a product of N_2	*automobile factory, light bulb, color printer*	27(0)/32(6)
PROPERTY	N_2 is N_1	*elephant seal, fairy penguin*	76(3)/85(3)
PURPOSE	N_2 is meant for N_1	*concert hall, soup pot, grinding abrasive*	159(13)/161(9)
RESULT	N_1 is a result of N_2	*storm cloud, cold virus, death penalty*	7(0)/8(0)
SOURCE	N_1 is the source of N_2	*chest pain, north wind, foreign capital*	86(11)/99(15)
TIME	N_1 is the time of N_2	*winter semester, morning class, late supper*	26(1)/19(0)
TOPIC	N_2 is concerned with N_1	*computer expert, safety standard, horror novel*	465(24)/447(39)

Table 3: The set of semantic relations (N_1 = modifier, N_2 = head noun)

semantic relation in such cases, but the annotators were not provided with this extra information source.

4.3. Annotation Procedure

The process for annotating the data set was similar to that described for training the annotators. The final annotation was performed over the 2,169 noun compounds, allowing multiple SRs. For all cases of disagreement, the two annotators came together to discuss their respective annotations and agree on a finalise set of SRs.

The initial agreement for the two annotators was 52.31%, with instances of the annotators agreeing on at least one SR being classified as agreement. Common confusion pairs amongst the initial disagreements were SOURCE and CAUSE, PURPOSE and TOPIC, and OBJECT and TOPIC.

5. Summary

In this paper, we have presented a dataset for English noun compound interpretation. We collected 2,169 English noun compounds from the POS-tagged Wall Street Journal, and annotated each NC type according to the 20 SRs defined by Barker and Szpakowicz (1998). Finally, we split the overall dataset into 1,081 and 1,088 instances for test and training, respectively.

During the annotation task, we confirmed that the agreement between human annotators for the NC interpretation task is low (Moldovan et al., 2004; Saghdha and Copestake, 2007). We also noticed that some NCs can be interpreted with multiple SRs, according to context (Downing, 1977). Finally, we reaffirm that defining and annotating SRs for NCs is a non-trivial task, and we hope that our data provides a reliable resource for further research.

6. References

Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998)*, pages 96–102, Montreal, Canada.

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

Timothy Wilking Finin. 1980. *The semantic interpretation of compound nominals*. Ph.D. thesis, University of Illinois, Urbana, Illinois, USA.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th Semantic Evaluation Workshop(SemEval-2007)*, pages 13–18, Prague, Czech Republic.

Roxana Girju. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 568–575, Prague, Czech Republic.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of compound nouns using wordnet::similarity. In *Proceedings of 2nd International Joint Conference on Natual Language Processing (IJCNLP-2005)*, pages 945–956, Jeju, Korea.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics (COLING/ACL-2006)*, pages 491–498, Syd-

ney, Australia.

Maria Lapata. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Judith Levi. 1979. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, New York, USA.

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

Preslav Nakov and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA)*, pages 233–244, Bularia.

Vivi Nastase, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and wordnet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 781–787, Boston, Massachusetts, USA.

Karen Sparck Jones. 1983. *Compound noun interpretation problems*. Prentice-Hall, Englewood Cliffes, NJ, USA.

Diarmuid Saghdha and Ann Copestake. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, pages 57–64, Prague, Czech Republic.

Diarmuid Saghdha. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the Corpus Linguistics*.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational linguistics*, pages 782–788, Kyoto, Japan.

Interpreting Compound Nominalisations

Jeremy Nicholson and Timothy Baldwin

Department of Computer Science and Software Engineering
University of Melbourne
Victoria 3010 Australia

{jeremymn,tim}@csse.unimelb.edu.au

Abstract

This paper describes a dataset intended to provide a common platform for evaluating research on the identification and interpretation of compound nominalisations in English.

1. Introduction

Compound nouns are a notable type of multiword expression whose underspecified semantics are notoriously difficult to recover (Levi, 1978; Copestake, 2003; Kim and Baldwin, 2006). A compound noun is a sequence of two or more nouns comprising an N, for example *cat house* "a house for a cat" and *house cat* "a cat which lives in a house".

Compound nominalisations are one important subclass of compound noun, and occur when the head noun is deverbal. The scope for interpretation of compound nominalisations is reduced considerably over the more general class of compound nouns, and can be defined relative to the argument structure of the verb from which the deverbal head noun was derived (Levi, 1978; Lapata, 2002; Grover et al., 2005; Nicholson and Baldwin, 2006). An example is *product replacement*, which can be interpreted relative to the underlying verb *replace* as "(the act of) replacing the product".

In this paper, we describe a dataset for use in identifying and interpreting compound nominalisations.

2. Data

The dataset is based on a random sample of 1000 sentences from the British National Corpus (BNC: Burnard (2000)). Being entirely random, this lead to "sentences" such as (1) which do not contain compounds, as well as those such as (2) which contain more than one compound.

(1) 3.

(2) Vibration to the platform caused the power supply to be disrupted when the generators stopped, creating a temporary disruption to production and affecting the drilling operation.

3. Annotation

Three non-specialist annotators were instructed to: (1) identify binary compound nouns (i.e. sequences of two nouns comprising an N) within the raw text; (2) tag all binary compound nouns including one or more proper nouns (PN), and exclude them from further annotation; (3) provide the underlying verb for all identified compound nouns where the head noun is deverbal, and the compound noun is interpretable using the underlying verb; and (4) determine the semantic relation of all compound nominalisations. As the annotators were non-specialist and untrained, an adjudicator resolved all disagreements between the annotators to form the gold standard. For example, in (2), both *power supply* and *drilling operation* would first be identified as binary compound nouns. Additionally, the two compound nouns would be analysed as being made up exclusively of common nouns and having the underlying verbal forms *supply* and *operate*, respectively. Finally, they would both be annotated as having the direct object interpretation: "[someone] supplies power" and "[someone] operates the drilling", respectively.

32% of the sentences were found to contain at least one compound noun, with 464 compounds in total. About a quarter (119) of these were identified as containing one or more proper nouns.

In annotating the underlying verb of the head noun, three categories were used: (1) the head of the compound is not deverbal (NV); (2) the head of the compound noun has a verbal form, but it does not occur in a productive semantic relation with the modifier (NA); or (3) the head has a verbal form which forms the basis of a semantic relation with the modifier. In the latter case, three semantic relations are considered: (1) the head noun is deverbal, and the modifier corresponds to the subject of the base verb (SUB); (2) the head noun is deverbal, and the modifier corresponds to the direct object of the base verb (DOB); and (3) the head noun is deverbal, and the modifier corresponds to a prepositional object of the base verb (POB). In the final case of the modifier being analysed as a prepositional object of the base verb, annotators were additionally asked to provide the preposition, and indicate whether the prepositional object is an argument or adjunct.

In the case of ambiguity, the annotators were instructed to choose the default interpretation, and to break ties in favour of verb–argument readings. The hierarchical nature of the class set is depicted in Figure 1. Note that our subclassification of compound nominalisations is syntactic in nature, in line with the research of Grover et al. (2005) and Nicholson and Baldwin (2006). We deliberately avoid classifying other compound noun types beyond flagging the head noun for deverbalisation, due to the lack of agreement on a set of semantic relations.

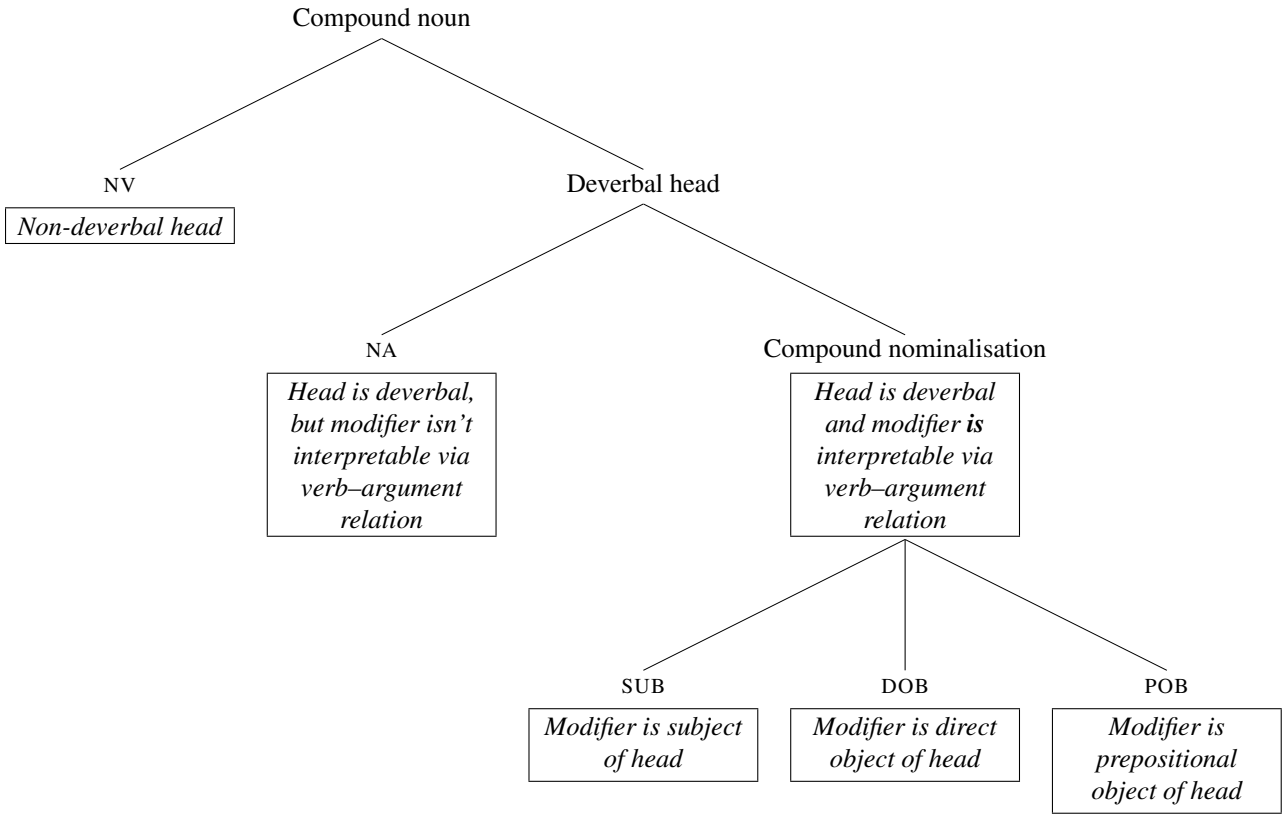


Figure 1: The 5 categories of compound noun used in the dataset, and 3 categories of compound nominalisation

```
<doc>
Demand for the new car is strongest in large urban areas like New <cn rel="PN" hvf="">York
city</cn>, Los Angeles and Miami , where bombings , riots and car-jackings fill the
<cn rel="NA" hvf="bulletin">news bulletins</cn> .
</doc>

<doc>
During my first attack I experienced some very inaccurate <cn rel="POB" prep="in" com="ADJ"
hvf="fire">return fire</cn> which ceased just before I broke away .
</doc>
```

Table 1: Sample data

Class	Example	Frequency
SUB	eyewitness report	22 (6.4%)
DOB	eye irritation	63 (18.2%)
POB	side show	44 (12.8%)
NV	scout hut	58 (16.8%)
NA	memory size	158 (45.8%)

Table 2: Composition of the dataset

4. Analysis

The five classes occurred with the frequencies indicated in Table 2. 129 of the 345 analysed compounds were given a verb–argument relation.

In detecting the 345 compounds, the three annotators had a mean precision of 92.5% and a mean recall of 84.8% relative to the gold standard. Most incorrectly-tagged compounds contained adjectives (e.g. *green beret* or *phonemic association*). Their raw inter-annotator agreement over unigrams was 98.4%. The kappa coefficient (Carletta, 1996) relative to the gold standard is 0.83, which indicates good (> 0.8) but not perfect agreement, attesting to the difficulty of the task.

5. Data Format

Annotated examples are provided in Table 1. In the first example, both *York city* and *news bulletins* have been identified as binary compound nouns. *York city* has been tagged as incorporating a proper noun (PN), while *news bulletins* has been analysed as having a deverbal head (base verb *bulletin*) but also as the base verb having no bearing on the semantic relation (NA). In the second example, *return fire* is identified as a compound noun, *fire* as the base verb of the head noun, and the modifier (*return*) as a prepositional

object (POB) of the head noun, where the preposition is *in* (i.e. the interpretation is of the form *fire in return*) and the prepositional object is an adjunct.

6. Summary

This paper has outlined a dataset which is intended to provide a platform for standardised evaluation of the identification and interpretation of English compound nominalisations.

7. References

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Ann Copestake. 2003. Compounds revisited. In *Proc. of the 2nd International Workshop on Generative Approaches to the Lexicon*, Geneva, Switzerland.

Claire Grover, Mirella Lapata, and Alex Lascarides. 2005. A comparison of parsing technologies for the biomedical domain. *Journal of Natural Language Engineering*, 11(01):27–65.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proc. of COLING/ACL 2006*, pages 491–8, Sydney, Australia.

Maria Lapata. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–88.

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

Jeremy Nicholson and Timothy Baldwin. 2006. Interpretation of compound nominalisations using corpus and web statistics. In *Proc. of the COLING/ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 54–61, Sydney, Australia.

Paraphrasing Verbs for Noun Compound Interpretation

Preslav Nakov

Linguistic Modeling Department, Institute for Parallel Processing
Bulgarian Academy of Sciences
25A, Acad. G. Bonchev St., 1113 Sofia, Bulgaria
nakov@lml.bas.bg

Abstract

An important challenge for the automatic analysis of English written text is the abundance of noun compounds: sequences of nouns acting as a single noun. In our view, their semantics is best characterized by the set of all possible paraphrasing verbs, with associated weights, e.g., *malaria mosquito* is *carry (23)*, *spread (16)*, *cause (12)*, *transmit (9)*, etc. Using Amazon's Mechanical Turk, we collect paraphrasing verbs for 250 noun-noun compounds previously proposed in the linguistic literature, thus creating a valuable resource for noun compound interpretation. Using these verbs, we further construct a dataset of pairs of sentences representing a special kind of textual entailment task, where a binary decision is to be made about whether an expression involving a verb and two nouns can be transformed into a noun compound, while preserving the sentence meaning.

1. Introduction

An important challenge for the automatic analysis of English written text is posed by noun compounds – sequences of nouns acting as a single noun[1], e.g., *colon cancer tumor suppressor protein* – which are abundant in English: Baldwin and Tanaka (2004) calculated that noun compounds comprise 3.9% and 2.6% of all tokens in the *Reuters corpus* and the *British National Corpus*[2], respectively.

Understanding noun compounds' syntax and semantics is difficult but important for many natural language applications (NLP) including question answering, machine translation, information retrieval, and information extraction. For example, a question answering system might need to determine whether '*protein acting as a tumor suppressor*' is a good paraphrase for *tumor suppressor protein*, and an information extraction system might need to decide whether *neck vein thrombosis* and *neck thrombosis* could possibly co-refer when used in the same document. Similarly, a machine translation system facing the unknown noun compound *WTO Geneva headquarters* might benefit from being able to paraphrase it as *Geneva headquarters of the WTO* or as *WTO headquarters located in Geneva*. Given a query like *migraine treatment*, an information retrieval system could use suitable paraphrasing verbs like *relieve* and *prevent* for page ranking and query refinement.

2. Noun Compound Interpretation

The dominant view in theoretical linguistics is that noun compound semantics can be expressed by a small set of abstract relations. For example, in the theory of Levi (1978), complex nominals (a more general notion than noun compounds) can be derived by two processes – predicate deletion (e.g., *pie made of apples* → *apple pie*) and predicate nominalization (e.g., *the President refused general MacArthur's request* → *presidential refusal*). The former can only delete the 12 abstract recoverably deletable predicates (RDPs) shown in Table 1.

RDP	Example	Subj/obj	Traditional Name
CAUSE$_1$	*tear gas*	object	causative
CAUSE$_2$	*drug deaths*	subject	causative
HAVE$_1$	*apple cake*	object	possessive/dative
HAVE$_2$	*lemon peel*	subject	possessive/dative
MAKE$_1$	*silkworm*	object	productive/composit.
MAKE$_2$	*snowball*	subject	productive/composit.
USE	*steam iron*	object	instrumental
BE	*soldier ant*	object	essive/appositional
IN	*field mouse*	object	locative
FOR	*horse doctor*	object	purposive/benefactive
FROM	*olive oil*	object	source/ablative
ABOUT	*price war*	object	topic

Table 1: **Levi's recoverably deletable predicates (RDPs).** Column 3 shows modifier's function in the relative clause.

Similarly, in the theory of Warren (1978), noun compounds can express six major types of semantic relations (which are further divided into finer sub-relations): Constitute, Possession, Location, Purpose, Activity–Actor, and Resemblance.

A similar view is dominant in computational linguistics. For example, Nastase and Szpakowicz (2003) use 30 fine-grained relations (e.g., Cause, Effect, Purpose, Frequency, Direction, Location), grouped into 5 coarse-grained super-relations: QUALITY, SPATIAL, TEMPORALITY, CAUSALITY, and PARTICIPANT. Similarly, Girju et al. (2005) propose a set of 21 abstract relations (e.g., CAUSE, INSTRUMENT, PURPOSE, RESULT), and Rosario and Hearst (2001) use 18 abstract domain-specific biomedical relations (e.g., Defect, Material, Person Afflicted).

An alternative view is held by Lauer (1995), who defines the problem of noun compound interpretation as predicting which among the following eight prepositions best paraphrases the target noun compound: *of*, *for*, *in*, *at*, *on*, *from*, *with*, and *about*. For example, *olive oil* is *oil from olives*.

Lauer's approach is attractive since it is simple and yields prepositions representing paraphrases directly usable in NLP applications. However, it is also problematic since

[1] This is Downing (1977)'s definition of noun compounds.

[2] There are 256K distinct noun compounds out of the 939K distinct wordforms in the 100M-word *British National Corpus*.

mapping between prepositions and abstract relations is hard (Girju et al., 2005), e.g., *in*, *on*, and *at*, all can refer to both LOCATION and TIME.

Using abstract relations like CAUSE is problematic as well. First, it is unclear which relation inventory is the best one. Second, being both abstract and limited, such relations capture only part of the semantics, e.g., classifying *malaria mosquito* as CAUSE obscures the fact that mosquitos do not directly cause malaria, but just transmit it. Third, in many cases, multiple relations are possible, e.g., in Levi's theory, *sand dune* can be interpreted as both HAVE and BE.

Some of these issues are addressed by Finin (1980), who proposes to use a specific verb, e.g., *salt water* is interpreted as *dissolved in*. In a number of publications (Nakov and Hearst, 2006; Nakov, 2007; Nakov and Hearst, 2008), we introduced and advocated an extension of this idea, where noun compounds are characterized by the set of all possible paraphrasing verbs, with associated weights, e.g., *malaria mosquito* is *carry (23)*, *spread (16)*, *cause (12)*, *transmit (9)*, *etc*. These verbs are fine-grained, directly usable as paraphrases, and using multiple of them for a given noun compound approximates its semantics better.

Following this line of research, below we present two noun compound interpretation datasets which use human-derived paraphrasing verbs and are consistent with the view of an infinite inventory of relations. By making these resources publicly available, we hope to inspire further research in paraphrase-based noun compound interpretation.

3. Manual Annotations

We used a subset of the 387 examples listed in the appendix of (Levi, 1978). As we mentioned above, Levi's theory targets complex nominals, which include not only nominal compounds (e.g., *peanut butter*, *snowball*), but also nominalizations (e.g., *dream analysis*), and nonpredicate noun phrases (e.g., *electric shock*). We kept the former two categories since they are composed of nouns only and thus are noun compounds under our definition, but we removed the nonpredicate noun phrases, which have an adjectival modifier. We further excluded all concatenations (e.g., *silkworm*), thus ending up with 250 noun-noun compounds.

We then defined a paraphrasing task which asks human subjects to produce verbs, possibly followed by prepositions, that could be used in a paraphrase involving *that*. For example, *come from*, *be obtained from*, and *be from* are good paraphrasing verbs for *olive oil* since they can be used in paraphrases like '*oil that comes from olives*', '*oil that is obtained from olives*' or '*oil that is from olives*'. Note that this task definition implicitly allows for prepositional paraphrases when the verb is to *be* and is followed by a preposition. For example, the last paraphrase above is equivalent to '*oil from olives*'.

In an attempt to make the task as clear as possible and to ensure high quality of the results, we provided detailed instructions, we stated explicit restrictions, and we gave several example paraphrases. We instructed the participants to propose at least three paraphrasing verbs per noun-noun compound, if possible. We used the *Amazon Mechanical Turk* Web service[3], which represents a cheap and easy way to recruit subjects for various tasks that require human intelligence; it provides an API allowing a computer program to ask a human to perform a task and return the results.

We randomly distributed the noun-noun compounds into groups of 5 and we requested 25 different human subjects per group. We had to reject some of the submissions, which were empty or were not following the instructions, in which cases we requested additional workers in order to obtain about 25 good submissions per HIT (Human Intelligence Task). Each human subject was allowed to work on any number of groups, but was not permitted to do the same one twice, which is controlled by the *Amazon Mechanical Turk* Web Service. A total of 174 different human subjects produced 19,018 verbs. After removing the empty and the bad submissions, and after normalizing the verbs, we ended up with 17,821 verb annotations for the 250 examples. See Nakov (2007) for further details on the process of extraction and cleansing.

[3] http://www.mturk.com

4. Lexicons of Paraphrasing Verbs

We make freely available three lexicons of paraphrasing verbs for noun compound interpretation: two generated by human subjects recruited with *Amazon Mechanical Turk*, and a third one automatically extracted from the Web, as described in (Nakov and Hearst, 2008).

4.1. Human-Proposed: All

The dataset is provided as a text file containing a separate line for each of the 250 noun-noun compounds, ordered lexicographically. Each line begins with an example number (e.g., 94), followed by a noun compound (e.g., *flu virus*), the original Levi's RDP (e.g., $\texttt{CAUSE}_1$; see Table 1), and a list of paraphrasing verbs. The verbs are separated by a semicolon and each one is followed in parentheses by a count indicating the total number of distinct human annotators that proposed it. Here is an example line:

```
94 flu virus CAUSE1 cause(19); spread(4); give(4);
result in(3); create(3); infect with(3); contain(3);
be(2); carry(2); induce(1); produce(1); look like(1);
make(1); incubate into(1); exacerbate(1); turn into(1);
happen from(1); transmit(1); be made of(1); involve(1);
generate(1); breed(1); be related to(1); sicken with(1);
lead to(1); intensify be(1); disseminate(1); come
from(1); be implicated in(1); appear(1); instigate(1);
be conceived by(1); bring about(1)
```

4.2. Human-Proposed: First Only

As we mentioned above, the human subjects recruited to work on *Amazon Mechanical Turk* (workers) were instructed to provide at least three paraphrasing verbs per noun-noun compound. Sometimes this was hard, and many introduced some bad verbs in order to fulfill this requirement. Assuming that the very first verb is the most likely one to be correct, we created a second dataset in the same format, where only the first verb from each worker is considered. For example, line 94 in that new text file becomes:

```
94 flu virus CAUSE1 cause(17), be(1), carry(1),
involve(1), come from(1)
```

4.3. Automatically Extracted from the Web

Finally, we provide a text file in the same format, where the verbs are automatically extracted from the Web using the method described in (Nakov and Hearst, 2008). This dataset might be found useful by other researchers for comparison purposes. The corresponding line 94 in that file starts as follows (here truncated due to a very long tail):

```
94 flu virus CAUSE1 cause(906); produce(21);
give(20); differentiate(17); be(16); have(13);
include(11); spread(7); mimic(7); trigger(6); induce(5);
develop from(4); be like(4); be concealed by(3); be
characterized by(3); bring(3); carry(3); become(3); be
associated with(3); ...
```

4.4. Comparing the Human-Proposed and the Program-Generated Paraphrasing Verbs

In this section, we describe a comparison of the human- and the program-generated verbs aggregated by Levi's RDP (see Table 1). Given an RDP like $\texttt{HAVE}_1$, we collected all verbs belonging to noun-noun compounds from that RDP together with their frequencies. From a vector-space model point of view, we summed their corresponding frequency vectors. We did this separately for the human- and the program-generated verbs, and we compared them for each RDP. Figure 4.4. shows the cosine correlations (in %s) between the human- and the program-generated verbs by Levi's RDP: using all human-proposed verbs vs. using the first verb from each worker only. As we can see, there is a very-high correlation (mid 70s to mid 90s) for RDPs like $\texttt{CAUSE}_1$, $\texttt{MAKE}_1$, and BE, but low correlation 11-30% for reverse RDPs like $\texttt{HAVE}_2$ and $\texttt{MAKE}_2$, and for rare RDPs like ABOUT. Interestingly, using the first verb only improves the results for RDPs with high cosine correlation, but damages low-correlated ones. This suggests that when the RDP is more homogeneous, the first verbs proposed by the workers are good enough and the following ones only introduce noise, but when it is more heterogeneous, the additional verbs are more likely to be useful.

We also performed an experiment using the verbs as features in a nearest-neighbor classifier, trying to predict the Levi's RDP the noun compound belongs to. We first filtered out all nominalizations, thus obtaining 214 noun compounds, each annotated with one of the 12 RDPs shown in Table 1, and we then used this dataset in a leave-one-out cross-validation. Using all human-proposed verbs, we achieved 73.71%±6.29% accuracy (here we also show the confidence interval). For comparison, using Web-derived verbs and prepositions only yields 50.47%±6.68% accuracy. Therefore, we can conclude that the performance with human-proposed verbs is an upper bound on what can be achieved with Web-derived ones. See (Nakov and Hearst, 2008) for additional details.

5. A Dataset for Textual Entailment

Collecting this dataset was motivated by the Pascal Recognizing Textual Entailment (RTE) Challenge,[4] which addresses a generic semantic inference task arguably needed by many NLP applications, including question answering,

[4] www.pascal-network.org/Challenges/RTE2/

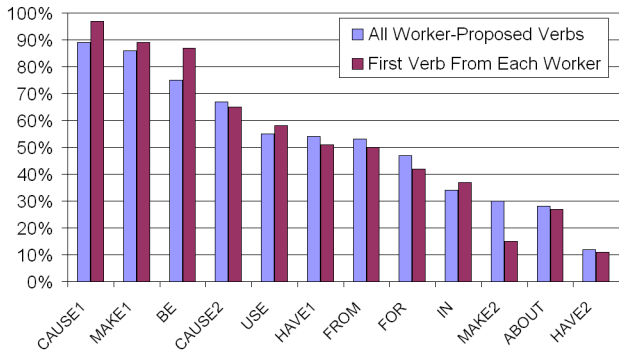


Figure 1: **Cosine correlation (in %s) between the human- and the program- generated verbs by Levi's RDP:** using all human-proposed verbs vs. using the first verb from each worker only.

information retrieval, information extraction, and multi-document summarization. Given two textual fragments, a text T and a hypothesis H, the goal is to recognize whether the meaning of H is entailed (can be inferred) from the meaning of T. Or, as the RTE2 task definition puts it:

> "*We say that T entails H if, typically, a human reading T would infer that H is most likely true. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge.*"

In many cases, solving such entailment problems requires deciding whether a noun compound can be paraphrased in a particular way.

The sentences in our Textual Entailment dataset are collected from the Web and involve some of the above-described human-derived paraphrasing verbs. These sentences are further manually annotated and provided in format that is similar to that used by RTE. Each example consists of three lines, all starting with the example number. The first line continues with T: (the text), followed by a sentence where the target nouns involved in a paraphrase are marked. The second line continues with H: (the hypothesis), followed by the same sentence but with the paraphrase re-written as a noun compound. The third line continues with A: (the answer), followed by either YES or NO, depending on whether T implies H.

The following example is positive since *professors that are women* is an acceptable paraphrase of the noun compound *women professors*:

```
17 T: I have friends that are organizing
to get more <e2>professors</e2> that are
<e1>women</e> and educate women to make
specific choices on where to get jobs.
17 H: I have friends that are organizing to
get more <e1>women</e1> <e2>professors</e2>
and educate women to make specific choices
on where to get jobs.
17 A: YES
```

The example below however is negative since a bad paraphrasing verb is used in the first sentence:

```
18 T: As McMillan collected, she also
quietly gave, donating millions of
dollars to create scholarships and
fellowships for black Harvard Medical
School students, African filmmakers,
and MIT <e2>professors</e2> who study
<e1>women</e1> in the developing world.
18 H: As McMillan collected, she also
quietly gave, donating millions of dollars
to create scholarships and fellowships
for black Harvard Medical School students,
African filmmakers, and MIT <e1>women</e1>
<e2>professors</e2> in the developing
world.
18 A: NO
```

Here is another kind of negative example, where the semantics is different due to a different phrase attachment. The first sentence refers to the action of giving, while the second one refers to the process of transfusion:

```
19 T: Rarely, the disease is transmitted
via transfusion of blood products from
a <e2>donor</e2> who gave <e1>blood</e1>
during the viral incubation period.
19 H: Rarely, the disease is transmitted
via transfusion of blood products from a
<e1>blood</e1> <e2>donor</e2> during the
viral incubation period.
19 A: NO
```

6. Conclusion

We have presented several novel resources consistent with the idea of characterizing noun compound semantics by the set of all possible paraphrasing verbs. These verbs are fine-grained, directly usable as paraphrases, and using multiple of them for a given noun compound approximates its semantics better. By making these resources publicly available, we hope to inspire further research in the direction of paraphrase-based noun compound interpretation, which opens the door to practical applications in a number of NLP tasks including but not limited to machine translation, text summarization, question answering, information retrieval, textual entailment, relational similarity, etc.

Unfortunately, the present situation with noun compound interpretation is similar to the situation with word sense disambiguation: in both cases, there is a general agreement that the research is important and much needed, there is a growing interest in performing further research, and a number of competitions are being organized, e.g., as part of SemEval (Girju et al., 2007). Still, despite that research interest, there is a lack of actual NLP applications using noun compound interpretation, with the notable exceptions of Tatu and Moldovan (2005) and Nakov (2008), who demonstrated improvements on textual entailment and machine translation, respectively. We believe that demonstrating more successful applications in real NLP problems is key for the advancement of the field, and we hope that other researchers will find the resources we release here helpful in this respect.

7. License

All datasets are released under the *Creative Commons License*[5].

8. References

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pages 24–31.

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

Timothy Finin. 1980. *The Semantic Interpretation of Compound Nominals*. Ph.d. dissertation, University of Illinois, Urbana, Illinois.

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

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of SemEval*, pages 13–18, Prague, Czech Republic.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Dept. of Computing, Macquarie University, Australia.

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

Preslav Nakov and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In *AIMSA*, volume 4183 of *Lecture Notes in Computer Science*, pages 233–244. Springer.

Preslav Nakov and Marti Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of ACL'08: HLT*, Columbus, OH.

Preslav Nakov. 2007. *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. Ph.D. thesis, EECS Department, University of California, Berkeley, UCB/EECS-2007-173.

Preslav Nakov. 2008. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, OH.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 285–301, Tilburg, The Netherlands.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of EMNLP*, pages 82–90.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of HLT*, pages 371–378.

Beatrice Warren. 1978. Semantic patterns of noun-noun compounds. In *Gothenburg Studies in English 41, Goteburg, Acta Universtatis Gothoburgensis*.

[5] http://creativecommons.org/

An Evaluation of Methods for the Extraction of Multiword Expressions

Carlos Ramisch[♣♢], Paulo Schreiner[♣], Marco Idiart[♡] and Aline Villavicencio[♣♠]

[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♢]GETALP Laboratory, Joseph Fourier University - Grenoble INP (France)
[♡]Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
[♠]Department of Computer Sciences, Bath University (UK)
{ceramisch,pschreiner}@inf.ufrgs.br, idiart@if.ufrgs.br, avillavicencio@inf.ufrgs.br

Abstract

This paper focuses on the evaluation of some methods for the automatic acquisition of Multiword Expressions (MWEs). First we investigate the hypothesis that MWEs can be detected solely by the distinct statistical properties of their component words, regardless of their type, comparing 3 statistical measures: Mutual Information, χ^2 and Permutation Entropy. Moreover, we also look at the impact that the addition of type-specific linguistic information has on the performance of these methods.

1. Introduction

The task of automatically identifying Multiword Expressions (MWEs) like phrasal verbs (*sell out*) and compound nouns (*science fiction*) using statistical measures has been the focus of considerable investigative effort, (e.g. Pearce (2002), Evert and Krenn (2005) and Zhang et al. (2006)). Among these, some research has focused on the development of methods for dealing with specific types of MWEs (e.g. Pearce (2002) on collocations and Villavicencio (2005) on verb-particle constructions), and some work on dealing with MWEs in general (e.g. Zhang et al. (2006)). These works tend to focus on one language (e.g. Pearce (2002) and Zhang et al. (2006) for English and Evert and Krenn (2005) for German).

As basis for helping to determine whether a given sequence of words is in fact an MWE some of them employ (language and/or MWE-type dependent) linguistic knowledge for the task, while others employ (language and type independent) statistical methods, such as Mutual Information and Log-likelihood (e.g. Pearce (2002) and Zhang et al. (2006)), or a combination of both (e.g. Baldwin (2005) and Sharoff (2004)). Given the heterogeneousness of the different phenomena that are considered to be MWEs, there is no consensus about which method is best suited for which type of MWE, and if there is a single method that can be successfully used for any kind of MWE. Therefore, it would be of great value to know if a given MWE extraction approach could be successfully applied to other MWE types and/or languages (or families of languages), and if so, how good their performance would be.

In this paper we use three distinct MWE types from two different languages to evaluate some association measures: Mutual Information (MI), χ^2 and Permutation Entropy (PE) (Zhang et al., 2006). We also investigate the effect of adding some language and MWE-type specific information to the identification task, proposing a new measure, Entropy of Permutation and Insertion (EPI).

This paper starts with a brief description of the data sets (§ 2.). We then present the two different approaches used for identifying MWEs: a language and MWE-type independent set of association measures (§ 3.), and a language and type dependent set (§ 4.). We finish with a discussion of the overall results (§ 5.).

2. The Data

The evaluation of these association measures was performed over three distinct data sets: a list of 3,078 English Verb-Particle Constructions (VPCs) with manual annotation of idiomatic verb-particle pairs (which we refer to as EN-VPC); a manually annotated list of 1,252 German adjective-noun pairs (DE-AN); and a manually annotated list of 21,796 German combinations of prepositional phrase and governing verb (DE-PNV).[1]

These data sets are used as gold standard for the evaluation, as they are annotated with information about positive and negative instances of each of these MWE types. In addition the two German sets also contain frequency information, based on which the association measures are computed. The only pre-processing done in the data sets was that for DE-PNV we filtered out all candidates that appear less than 30 times in the Frankfurter Rundschau (FR) German corpus, to obtain a cleaner data set. The frequencies for the English set were collected from two different sources: the Web, using Yahoo APIs, which return the number of pages indexed for each search (henceforth referred to as *Yahoo*) and a fragment of the British National Corpus (BNC - Burnard (2000)) of 1.8M sentences (the same employed by Zhang et al. (2006), henceforth BNC_f).

3. A Language and Type Independent Approach

For each data set we compute three type independent statistical measures for MWE identification: MI, χ^2 and PE. The first two are typical measures of association while PE is a measure of order association. PE was proposed by Zhang et al. (2006) as a possible measure to detect MWEs, under the hypothesis that MWEs are more rigid to permutations and therefore present smaller PEs. Even though it is quite

[1]The data sets were provided by: Timothy Baldwin for EN-VPC; Dictionary editors of Langenscheidt KG and Stefan Evert for DE-AN; Brigitte Krenn and Stefan Evert for DE-PNV. All data sets are available from multiword.sf.net/mwe2008/shared_task.html.

different from MI and χ^2, PE can also be thought as an indirect measure of statistical independence, since the more independent the words are the closer PE is to its maximal value ($\ln 2$, for bigrams).

For a bigram with words w_1w_2, χ^2 and MI are calculated respectively as:

$$\chi^2 = \sum_{a,b} \frac{[\, n(ab) - n_\emptyset(ab) \,]^2}{n_\emptyset(ab)}$$

$$\text{MI} = \sum_{a,b} \frac{n(ab)}{N} \log_2 \left[\frac{n(ab)}{n_\emptyset(ab)} \right]$$

where a corresponds either to the word w_1 or to $\neg w_1$ (all but the word w_1) and so on. $n(ab)$ is the number of bigrams ab in the corpus, $n_\emptyset(ab) = n(a)n(b)/N^2$ is the predicted number from the *null* hypothesis, $n(a)$ is the number of unigrams a, and N the number of bigrams in the corpus. For these two measures we only use the FR and BNC_f corpora, since for them the size of the corpus is known (the value of N). PE is calculated as:

$$\text{PE} = -\sum_{(i,j)} p(w_iw_j) \ln [\, p(w_iw_j) \,]$$

where the sum runs over all the permutations of the indices and, therefore, over all possible positions of the selected words in the bigram. The probabilities are estimated from the number of occurrences of each permutation (e.g. *computer science* and *science computer*) as:

$$p(w_1w_2) = \frac{n(w_1w_2)}{\sum_{(i,j)} n(w_iw_j)}$$

For calculating PE we used the Yahoo corpus and for each of the data sets we restricted the search to return only pages in that language (English or German). The Yahoo corpus can be used for PE, since, unlike MI and χ_2, PE is calculated independently of the size of the corpus, and the use of Yahoo as a corpus can minimize the problem of data sparseness.

The results of these three evaluations can be seen in figures 1 to 3 and in table 1. In all these cases the statistical measures perform better than the baseline, with the expected trade-off between precision and recall. The exception is PE. When this measure is calculated on the basis of varying the order of the words, it provides a stronger contribution when there is no underlying grammatical constraint preventing the combination of the constituents in the permuted orders. If, as in the case of English VPCs, the particle is only expected after the verb (but not before), PE does not add much information, since due to grammatical constraints the permuted orders are not going to be often found.

For both EN-VPC and DE-AN, MI and χ^2 have very similar performances. However, for DE-PNV MI seems to have a much better predictive power than χ^2 and any of the other measures.

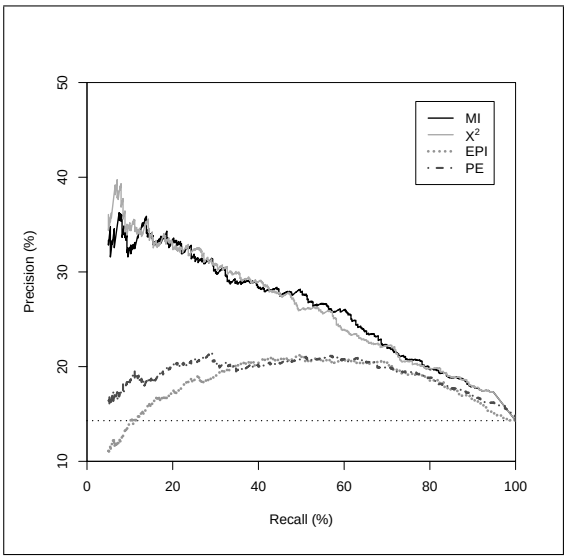


Figure 1: Precision-recall graphic for EN-VPC data set

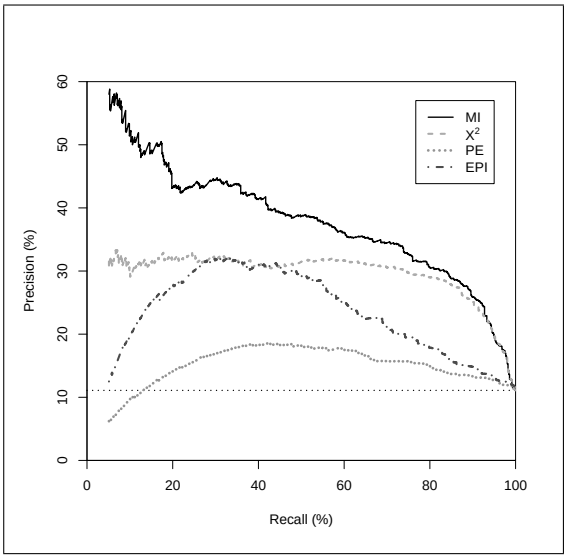


Figure 2: Precision-recall graphic for DE-PNV data set

4. A Language and Type Dependent Approach

In order to evaluate whether the addition of linguistic information can further improve MWE identification compared to the use of purely frequency-based measures, we performer further tests with two of the data sets: the German DE-PNV and the English EN-VPC. For that we introduce an entropy measure, the Entropy of Permutation and Insertion (EPI), that takes into account linguistic information about the MWE type. EPI is calculated as follows:

$$\text{EPI} = -\sum_{a=0}^{m} p(ngram_a) \ln [\, p(ngram_a) \,]$$

where $ngram_0$ is the original expression, and $ngram_a$ for $a = 1...m$, are m syntactic variants of the original expression. As before we calculate the probability of occurrences

Measure	Corpus	Data set EN-VPC	DE-PNV	DE-AN	EN-VPC-DICT
MI	BNC_f–FR	26.09%	39.05%	56.09%	39.59%
χ^2	BNC_f–FR	26.41%	29.85%	56.91%	41.46%
PE	Yahoo	17.96%	14.64%	40.35%	35.74%
EPI	Yahoo	19.33%	22.74%	–	39.23%
Baseline	–	14.29%	11.09%	41.53%	30.15%

Table 1: Average precisions for the studied measures

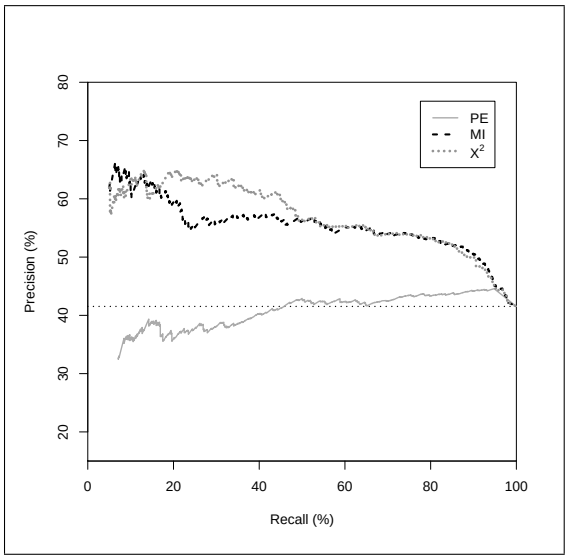


Figure 3: Precision-recall graphic for DE-AN data set

of any one of the variants as:

$$p(ngram_a) = \frac{n(ngram_a)}{\sum_{a=0}^{m} n(ngram_a)}$$

EPI is an extension to PE based on the idea that not all types of MWEs have the same behaviour. Therefore, if we know what kinds of modification an MWE type accepts or refuses in a particular language, we may be able to obtain more accurate entropies that may improve the identification task.

VPCs in English, for instance, have a very strict word order in that the verb comes before the particle, and accept little intervening material between them, but appear in a number of different syntactic configurations (e.g. the split and joint configuration for transitive VPCs). Thus, to identify VPCs contained in the EN-VPC data set we performed further Yahoo searches to account for some linguistic features that distinguish them from prepositional verbs or free verb-preposition combinations (e.g. *walk up the hill*).[2] The following patterns were used:

- Intransitive VPC: VERB + PARTICLE + DELIMITER
- Split Transitive VPC: VERB + NP + PARTICLE + DELIMITER
- Joint Transitive VPC: VERB + PARTICLE + NP + DELIMITER

We searched for exact matches of these patterns, where VERB corresponds to the verb element of a VPC candidate, PARTICLE, to the particle of the VPC, NP is either '*this*' or '*the* *' (with the Yahoo wildcard standing for one word), and DELIMITER is the preposition *with*. The delimiter is used to avoid retrieving pages where the particle is followed by an NP, which would also be ambiguous with prepositional verbs and free verb-preposition combinations, following Villavicencio (2005). Two distinct transitive VPC configurations were used: the *split* for when the verb is separated from the particle by an NP complement, and the *joint*, for when the verb and particle are adjacent to each other. Note that in the joint configuration pattern, there may be some false positive cases (e.g. prepositional verbs) since the delimiter is not immediatly following the particle anymore, which will introduce some noise in the frequencies obtained. However, since this is one of three configurations that are combined in EPI, even if there is some noise, it will be counterbalanced by the other configurations.

For VPCs an EPI closer to 1 indicates a VPC since variations are more characteristic of a genuine VPCs while non-VPCs will show a peak for the canonical form (*verb particle/preposition*). Figure 1, also shows the results for EPI, which has a higher precision-recall rate than PE, and therefore a higher average precision (19.33% vs 17.96%), but still lower than MI and χ_2.

The original EN-VPC data set was manually marked for true positives for all VPC candidates that are idiomatic. A closer look at the data, however, revealed that many of the unmarked candidates are nonetheless present in machine-readable dictionaries. Therefore, in order to evaluate the measures in terms of their effectiveness in detecting VPCs, regardless of their idiomaticity we used a list of 3,156 VPCs contained in either the Alvey Natural Language Tools (ANLT) lexicon (Carroll and Grover, 1989), the Comlex lexicon (Macleod and Grishman, 1998), and the LinGO English Resource Grammar (ERG) (Copestake and Flickinger, 2000)[3]. Using this as a gold standard, we obtained a new baseline of 30.15%, and a considerable improvement in performance. Average precision of χ^2, for example, improves to 41.46% (vs 26.41% with manual annotation). These results suggest that these measures seem more adequate to detect VPCs in general rather than to detect id-

[2]These features are based on those used by Baldwin (2005).

[3]Version of November 2001.

iomaticity in them.

For the DE-PNV data set, the first attempt to include linguistic information in the identification task is done by means of capturing inflectional patterns of German prepositions, which in the data set are marked with a "+" symbol if they inflect (e.g. *in+:Bett* as *ins Bett* or *im Bett*). To account for this variability we use the boolean operators available in Yahoo and a search for a combination like *in+:Bett liegen* originates the exact search term *(in OR im OR ins) Bett liegen* that has the potential to return either of those three prepositional forms occurring as the first word.[4]

Besides prepositional inflection, the other source of language-dependent information for the identification of DE-PNV is based on the assumption that fixed and semi-fixed MWEs do not accept determiners being inserted into the expression. This behaviour is essentially different from English VPCs, where genuine candidates do accept some syntactic variation. In German, a verb may appear before or after the indirect complement, depending on the context (e.g. both *in Kontakt treten* and –the less frequent but possible– *treten in Kontakt* might occur). However, true MWEs accept less well the addition of a determiner (except eventually for an article) placed between the preposition and the noun (e.g. *in Kontakt treten* but not *in großen Kontakt treten* nor *in den Kontakt treten*). To capture that we searched the Web for four different combinations (the Yahoo wildcard stands for a word like a pronoun, an article, an adjective, etc.): (1) *in Kontakt treten*, (2) *treten in Kontakt*, (3) *in * Kontakt treten* and (4) *treten in * Kontakt*. For DE-PNVs a high EPI indicates a more homogeneous distribution (i.e. not an MWE), while a low EPI suggests that there is a peak with only one acceptable form (i.e. indicating an MWE). This change in EPI interpretation shows that the measure can be easily adapted from one language and/or MWE type to another with the addition of some linguistic information and the appropriate interpretation. These patterns can be easily obtained, for instance with a linguist, and verified in a corpus (or in the Web), independently of expensive resources like dictionaries, huge corpora and thesauri and easily refined online.

Although the new measure is fairly superior than conventional PE for DE-PNV (figure 2), the result is far from being optimal, and we believe that some additional variation tests should be performed in order to reach higher quality levels. In terms of average precision, we go from 14.64% with PE to 22.74% with EPI.

The addition of linguistic information to both EN-VPC and DE-PNV had indeed an effect when compared to the standard PE. However, both MI and χ^2 still perform better.

5. Conclusions

One of the important challenges for robust natural language processing systems is to be able to successfully deal with Multiword Expressions and related constructions. In this paper we presented a first step towards investigating whether MWE identification methods can be robustly and successfully applied to different types of MWEs and different languages. The results suggest that although statistical measures on their own can detect trends and preferences in the co-ocurrences and combinations of words, for different languages and MWE types, they also have limited success in capturing some specific linguistic features, such as compositionality (in the EN-VPC data), which would require more sophisticated measures. Moreover, even if measures like MI and χ^2 seem to often agree on their rankings (Villavicencio et al., 2007), they may also have different performances for different MWE-types (e.g. for the DE-PNV). Finally, the individual performances of these measures may well be improved if they are combined together, offering different insights into the problem, and this is planned for future work.

[4]Some noun-verb combinations exclude some prepositional forms (like the impossible **ins Bett liegen* and **in Bett liegen*, and these will be reflected in the frequencies obtained, with any occasional noise being automatically corrected by the size of the Web.

6. References

Timothy Baldwin. 2005. Deep lexical acquisition of verb-particle constructions. *Computer Speech and Language*, 19(4):398–414.

Lou Burnard. 2000. User reference guide for the British National Corpus. Technical report, Oxford University Computing Services.

John Carroll and Claire Grover. 1989. The derivation of a large computational lexicon of English from LDOCE. In B. Boguraev and E. Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman.

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Catherine Macleod and Ralph Grishman. 1998. Comlex syntax reference manual, Proteus Project.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

Serge Sharoff. 2004. What is at stake: a case study of russian expressions starting with a preposition. pages 17–23, Barcelona, Spain.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043.

Aline Villavicencio. 2005. The availability of verb-particle constructions in lexical resources: How much is enough? *Journal of Computer Speech and Language Processing*, 19(4):415–432.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.

A Machine Learning Approach to Multiword Expression Extraction

Pavel Pecina

Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic
pecina@ufal.mff.cuni.cz

Abstract

This paper describes our participation in the MWE 2008 evaluation campaign focused on ranking MWE candidates. Our ranking system employed 55 association measures combined by standard statistical-classification methods modified to provide scores for ranking. Our results were crossvalidated and compared by Mean Average Precision. In most of the experiments we observed significant performance improvement achieved by methods combining multiple association measures.

1. Introduction

Four gold standard data sets were provided for the MWE 2008 shared task. The goal was to re-rank each list such that the "best" candidates are concentrated at the top of the list[1]. Our experiments were carried out only on three data sets – those provided with corpus frequency data by the shared task organizers: German Adj-N collocation candidates, German PP-Verb collocation candidates, and Czech dependency bigrams from the Prague Dependency Treebank. For each set of experiments we present the best performing association measure (AM) and results of our own system based on combination of multiple association measures (AMs).

2. System Overview

In our system which was already described in (Pecina and Schlesinger, 2006) and (Pecina, 2005), each collocation candidate x^i is described by the *feature vector* $\mathbf{x}^i = (x^i_1, \ldots, x^i_{55})^T$ consisting of 55 association scores from Table 1 computed from the corpus frequency data (provided by the shared task organizers) and assigned a label $y^i \in \{0,1\}$ which indicates whether the bigram is considered as true positive ($y = 1$) or not ($y = 0$). A part of the data is then used to train standard statistical-classification models to predict the labels. These methods are modified so they do not produce 0–1 classification but rather a score that can be used (similarly as for association measures) for ranking the collocation candidates (Pecina and Schlesinger, 2006). The following statistical-classification methods were used in experiments described in this paper: Linear Logistic Regression (GLM), Linear Discriminant Analysis (LDA), Neural Networks with 1 and 5 units in the hidden layer (NNet.1, NNet.5).

For evaluation we followed a similar procedure as in our previous work (Pecina and Schlesinger, 2006). Before each set of experiments every data set was split into seven stratified folds each containing the same ratio of true positives. Average precision (corresponding to the area under the precision-recall curve) was estimated for each data fold and its mean was used as the main evaluation measure (Mean Average Precision - MAP). The methods combining multiple association measures used 6 data folds for training and one for testing (7-fold crossvalidation).

[1] http://multiword.sf.net/mwe2008/

3. German Adj-N Collocation Candidates

3.1. Data Description

This data set consits of 1 252 German collocation candidates randomly sampled from the 8 546 different adjective-noun pairs (attributive prenominal adjectives only) occurring at least 20 times in the Frankfurter Rundschau corpus (FR, 1994). The collocation candidates were lemmatized with the IMSLex morphology (Lezius et al., 2000), preprocessed with the partial parser YAC (Kermes, 2003) for data extraction, and annotated by professional lexicographers with the following categories:

1. true lexical collocations, other multiword expressions
2. customary and frequent combination, often part of collocational pattern
3. common expression, but no idiomatic properties
4. unclear / boundary cases
5. not collocational, free combinations
6. lemmatization errors corpus-specific combinations

3.2. Experiments and Results

Frequency counts were provided for 1 213 collocation candidates from this data set. We performed two sets of experiments on them. First, only the categories 1–2 were considered true positives. There was a total of 511 such cases and thus the baseline precision was quite high (42.12%). The highest MAP of 62.88% achieved by *Piatersky–Shapiro coefficient* (51) was not outperformed by any of the combination method.

In the second set of experiments, the true positives comprised categories 1–2–3 (total of 628 items). The baseline precision was as high as 51.78%. The best association measure was again *Piatersky–Shapiro coefficient* (51) but it was slightly outperformed by most of the combination methods. The best one was based on LDA and achieved MAP of 70.77%. See detailed results in Table 2.

	1–2		*1–2–3*	
Baseline	42.12		51.78	
Best AM	**62.88**	(51)	69.14	(51)
GLM	60.88		70.62	
LDA	61.30		**70.77**	
NNet.1	60.52		70.38	
NNet.5	59.87		70.16	

Table 2: MAP results of ranking German Adj-N collocation candidates

#	Name	Formula
1.	**Joint probability**	$P(xy)$
2.	**Conditional probability**	$P(y\|x)$
3.	**Reverse conditional prob.**	$P(x\|y)$
4.	**Pointwise mutual inform.**	$\log \frac{P(xy)}{P(x*)P(*y)}$
5.	**Mutual dependency (*MD*)**	$\log \frac{P(xy)^2}{P(x*)P(*y)}$
6.	**Log frequency biased *MD***	$\log \frac{P(xy)^2}{P(x*)P(*y)} \quad \log P(xy)$
7.	**Normalized expectation**	$\frac{2f(xy)}{f(x*) \quad f(*y)}$
8.	**Mutual expectation**	$\frac{2f(xy)}{f(x*) \quad f(*y)} \cdot P(xy)$
9.	**Salience**	$\log \frac{P(xy)^2}{P(x*)P(*y)} \cdot \log f(xy)$
10.	**Pearson's χ^2 test**	$\sum_{i,j} \frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$
11.	**Fisher's exact test**	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$
12.	**t test**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
13.	**z score**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$
14.	**Poisson significance measure**	$\frac{\hat{f}(xy)-f(xy)\log \hat{f}(xy) \quad \log f(xy)!}{\log N}$
15.	**Log likelihood ratio**	$-2\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$
16.	**Squared log likelihood ratio**	$-2\sum_{i,j} \frac{\log f_{ij}^2}{\hat{f}_{ij}}$
17.	**Russel-Rao**	$\frac{a}{a \quad b \quad c \quad d}$
18.	**Sokal-Michiner**	$\frac{a \quad d}{a \quad b \quad c \quad d}$
19.	**Rogers-Tanimoto**	$\frac{a \quad d}{a \quad 2b \quad 2c \quad d}$
20.	**Hamann**	$\frac{(a \quad d)-(b \quad c)}{a \quad b \quad c \quad d}$
21.	**Third Sokal-Sneath**	$\frac{b \quad c}{a \quad d}$
22.	**Jaccard**	$\frac{a}{a \quad b \quad c}$
23.	**First Kulczynsky**	$\frac{a}{b \quad c}$
24.	**Second Sokal-Sneath**	$\frac{a}{a \quad 2(b \quad c)}$
25.	**Second Kulczynski**	$\frac{1}{2}(\frac{a}{a \quad b} \quad \frac{a}{a \quad c})$
26.	**Fourth Sokal-Sneath**	$\frac{1}{4}(\frac{a}{a \quad b} \quad \frac{a}{a \quad c} \quad \frac{d}{d \quad b} \quad \frac{d}{d \quad c})$
27.	**Odds ratio**	$\frac{ad}{bc}$
28.	**Yulle's ω**	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad} \quad \sqrt{bc}}$
29.	**Yulle's Q**	$\frac{ad-bc}{ad \quad bc}$
30.	**Driver-Kroeber**	$\frac{a}{\sqrt{(a \quad b)(a \quad c)}}$
31.	**Fifth Sokal-Sneath**	$\frac{ad}{\sqrt{(a \quad b)(a \quad c)(d \quad b)(d \quad c)}}$
32.	**Pearson**	$\frac{ad-bc}{\sqrt{(a \quad b)(a \quad c)(d \quad b)(d \quad c)}}$
33.	**Baroni-Urbani**	$\frac{a \quad \sqrt{ad}}{a \quad b \quad c \quad \sqrt{ad}}$
34.	**Braun-Blanquet**	$\frac{a}{\max(a \quad b, a \quad c)}$
35.	**Simpson**	$\frac{a}{\min(a \quad b, a \quad c)}$
36.	**Michael**	$\frac{4(ad-bc)}{(a \quad d)^2 \quad (b \quad c)^2}$
37.	**Mountford**	$\frac{2a}{2bc \quad ab \quad ac}$
38.	**Fager**	$\frac{a}{\sqrt{(a \quad b)(a \quad c)}} - \frac{1}{2}\max(b,c)$
39.	**Unigram subtuples**	$\log \frac{ad}{bc} - 3.29\sqrt{\frac{1}{a} \quad \frac{1}{b} \quad \frac{1}{c} \quad \frac{1}{d}}$
40.	***U* cost**	$\log(1 \quad \frac{\min(b,c)}{\max(b,c)} \quad \frac{a}{a})$
41.	***S* cost**	$\log(1 \quad \frac{\min(b,c)}{a \quad 1})^{-\frac{1}{2}}$
42.	***R* cost**	$\log(1 \quad \frac{a}{a \quad b}) \cdot \log(1 \quad \frac{a}{a \quad c})$
43.	***T* combined cost**	$\sqrt{U \times S \times R}$
44.	**Phi**	$\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$
45.	**Kappa**	$\frac{P(xy) \quad P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
46.	***J* measure**	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)} \quad P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)} \quad P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$
47.	**Gini index**	$\max[P(x*)(P(y\|x)^2 \quad P(\bar{y}\|x)^2) - P(*y)^2$ $P(\bar{x}*)(P(y\|\bar{x})^2 \quad P(\bar{y}\|\bar{x})^2) - P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2 \quad P(\bar{x}\|y)^2) - P(x*)^2$ $P(*\bar{y})(P(x\|\bar{y})^2 \quad P(\bar{x}\|\bar{y})^2) - P(\bar{x}*)^2]$
48.	**Confidence**	$\max[P(y\|x), P(x\|y)]$
49.	**Laplace**	$\max[\frac{NP(xy) \quad 1}{NP(x*) \quad 2}, \frac{NP(xy) \quad 1}{NP(*y) \quad 2}]$
50.	**Conviction**	$\max[\frac{P(x*)P(*y)}{P(x\bar{y})}, \frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$
51.	**Piatersky-Shapiro**	$P(xy) - P(x*)P(*y)$
52.	**Certainity factor**	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)}, \frac{P(x\|y)-P(x*)}{1-P(x*)}]$
53.	**Added value (*AV*)**	$\max[P(y\|x) - P(*y), P(x\|y) - P(x*)]$
54.	**Collective strength**	$\frac{P(xy) \quad P(\bar{x}\bar{y})}{P(x*)P(y) \quad P(\bar{x}*)P(*y)} \cdot$ $\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$
55.	**Klosgen**	$\sqrt{P(xy)} \cdot AV$

$a=f(xy)$	$b=f(x\bar{y})$	$f(x*)$
$c=f(\bar{x}y)$	$d=f(\bar{x}\bar{y})$	$f(\bar{x}*)$
$f(*y)$	$f(*\bar{y})$	N

A contingency table contains observed frequencies and marginal frequencies for a bigram xy; $\bar{w}$ stands for any word except w; $*$ stands for any word; N is a total number of bigrams. The table cells are sometimes referred to as f_{ij}. Statistical tests of independence work with contingency tables of expected frequencies $\hat{f}(xy) = f(x*)f(*y)/N$.

Table 1: Lexical association measures used for ranking MWE candidates.

4. German PP-Verb Collocation Candidates

4.1. Data Description

This data set comprises 21 796 German combinations of a prepositional phrase (PP) and a governing verb extracted from the Frankfurter Rundschau corpus (FR, 1994) and used in a number of experiments, e.g. (Krenn, 2000). PPs are represented by combination of a preposition and a nominal head. Both the nominal head and the verb were lemmatized using the IMSLex morphology (Lezius et al., 2000) and processed by the partial parser YAC (Kermes, 2003). See (Evert, 2004) for details of the extraction procedure. The data were manually annotated as lexical collocations or non-collocational by Brigitte Krenn (Krenn, 2000). In addition, distinction was made between two subtypes of lexical collocations: support-verb constructions (FVG), and figurative expressions (Figur).

4.2. Experiments and Results

On this data we carried out several series of experiments. First, we focused on the support-verb constructions and figurative expressions separately, then we attempted to extract all of them without making this distinction. Frequency data were provided for a total of 18 649 collocation candidates. The main experiments were performed on all of them. Further, as suggested by the shared task organizers, we restricted ourselves to a subset of 4 908 candidate pairs that occur at least 30 times in the Frankfurter Rundschau corpus (*in.fr30*). Similarly, additional experiments were restricted to candidate pairs containing one of 16 typical *light verbs*. This was motivated by assumption that filtering based on this condition should significantly improve the performance of association measures. After applying this filter the resulting set contained 6 272 collocation candidates.

Support-Verb Constructions

The baseline precision for ranking only the support-verb constructions in all the data is as low as 2.91%, the best MAP (18.26%) was achieved by *Confidence* measure. Additional substantial improvement was achieved by all combination methods. The best score (30.77%) was obtained by Neural Network (1 unit). When focused on the candidates occurring at least 30 times (baseline precision 5.75%), the best individual association measure appeared to be again *Confidence* measure with MAP 28.48%. The best combination method was then Neural Network with 5 units: MAP 43.40%. The best performing individual association measure on light verb data was *Poisson significance measure* (14) with MAP as high as 43.97% (baseline 7.25%). The performance gain achieved by the best combination method was not, however, so significant (45.08%, LDA). Details are shown in Table 3.

	all		*in.fr30*		*light.v*	
Baseline	2.91		5.75		7.25	
Best AM	18.26	(48)	28.48	(48)	43.97	(14)
GLM	28.40		26.59		41.25	
LDA	28.38		40.44		**45.08**	
NNet.1	**30.77**		42.42		44.98	
NNet.5	30.49		**43.40**		44.23	

Table 3: MAP results of ranking German PP-Verb support-verb construction candidates.

Figurative Expressions

Ranking figurative expressions seems more difficult. The best individual association measure on all data is again *Confidence* measure with MAP of only 14.98%, although the baseline precision is a little bit higher then in the case of support-verb constructions (3.16%). The best combination of multiple AMs is obtained by Logistic Regression (GLM) with MAP equal to 19.22%. Results for the candidates occurring at least 30 times (baseline precision 5.70%) are higher: the best AM (*Piatersky-Shapiro coefficient*) with MAP 21.04% and LDA with MAP 23.32%. In case of PP combinations with light verbs, the winning individual AM is *t test* (12) with MAP of 23.65% and the best combination method is Neural Network (5 units) with 25.86%. Details are depicted in Table 4.

	all		*in.fr30*		*light.v*	
Baseline	3.16		5.70		4.56	
Best AM	14.98	(48)	21.04	(51)	23.65	(12)
GLM	**19.22**		15.28		10.46	
LDA	18.34		**23.32**		24.88	
NNet.1	19.05		22.01		24.30	
NNet.5	18.26		22.73		**25.86**	

Table 4: MAP results of ranking German PP-Verb figurative expression candidates.

Support-Verb Constructions and Figurative Expressions

The last set of experiments performed on the German PP-Verb data aimed at ranking both support-verb constructions and figurative expressions without making any distinction between these two types of collocations. The results are shown in Table 5 and are not very surprising. The best individual AM on all the candidates as well as on the subset of the frequent ones was *Piatersky-Shapiro coefficient* with MAP 31.17% and 43.85%, respectively. *Poisson significance measure* (14) performed best on the candidates containing light verbs (63.59%). The best combination method were Neural Networks with 1 or 5 units. The most substantial performance improvement obtained by combining multiple AMs was observed on the set of all candidates (no filtering applied).

	all		*in.fr30*		*light.v*	
Baseline	6.07		11.45		11.81	
Best AM	31.17	(48)	43.85	(48)	63.59	(14)
GLM	44.66		47.81		65.37	
LDA	41.20		57.77		65.54	
NNet.1	44.71		**60.59**		65.10	
NNet.5	**44.77**		59.59		**66.06**	

Table 5: MAP results of ranking German PP-Verb candidates of both support-verb constructions and figurative expressions.

5. Czech PDT Bigrams

5.1. Data Description

The PDT data consist of notated list of 12 233 normalized dependency bigrams occurring in the manually annotated Prague Dependency Treebank (2.0, 2006) more than five times and having part-of-speech patterns that can possibly

form a collocation. Every bigram is assigned to one of the six categories described below by three annotators. Only the bigrams that all annotators agreed to be collocations (of any type, categories 1–5) are considered true positives. The entire set contains 2 572 such items. See (Pecina and Schlesinger, 2006) for details.

0. non-collocations
1. stock phrases, frequent unpredictable usages
2. names of persons, organizations, geographical locations, and other entities
3. support verb constructions
4. technical terms
5. idiomatic expressions

5.2. Experiments and Results

The baseline precision on this data is 21.02%. In our experiments, the best performing individual association measure was *Unigram subtuple measure* (39) with MAP of 65.63%. The best method combining all AMs was Neural Network (5 units) with MAP equal to 70.31%. After introducing a new (categorial) variable indicating POS patterns of the collocation candidates and adding it to the combination methods, the performance increased up to 79.51% (in case of the best method – Neural Network with 5 units) .

	AMs	*AMs+POS*
Baseline	21.01	
Best AM	65.63 (39)	
GLM	67.21	77.27
LDA	67.23	75.83
NNet.1	67.34	77.76
NNet.5	**70.31**	**79.51**

Table 6: MAP results of ranking Czech PDT collocation candidates. The second column refers to experiments using combination of association measures and information about POS patterns.

6. Conclusions

The overview of the best results achieved by individual AMs and by combination methods on all the data sets (and their variants) is shown in Table 7. With only one exception the combination methods significantly improved ranking of collocation candidates on all data sets. Our results showed that different measures give different results for different tasks (data). It is not possible to recommend "the best general association measure" for ranking collocation candidates. Instead, we suggest to use the proposed machine learning approach and let the classification methods do the job. Although it seems that Neural Network is probably the most suitable method for this task, we treat all the combination methods as equally good. We only recommend to use models that are fitted properly. Further, we also suggest to reduce the number of AMs employed in the combination methods by removing those that are redundant or do not help the prediction (see Pecina and Schlesinger (2006) for details.

Acknowledgments

This work has been supported by the Ministry of Education of the Czech Republic, projects MSM 0021620838.

Data Set	*Var*	*Baseline*	*Best AM*	*Best CM*	*+%*
GR Adj-N	1-2	42.40	62.88	61.30	-2.51
	1-2-3	51.74	69.14	70.77	2.36
GR PP-V FVG	all	2.89	18.26	30.77	68.51
	in.fr30	5.71	28.48	43.40	52.39
	light.v	7.26	43.97	45.08	2.52
GR PP-V Figur	all	3.15	14.98	19.22	28.30
	in.fr30	5.71	21.04	23.32	10.84
	light.v	4.47	23.65	25.86	9.34
GR PP-V	all	6.05	31.17	44.77	43.63
	in.fr30	11.43	43.85	60.59	38.18
	light.v	11.73	63.59	66.06	3.88
CZ PDT Bigram		21.01	65.63	70.31	7.13
	+POS	21.01	65.63	79.51	21.15

Table 7: Summary of the results obtained on all data sets and their variants. The last two columns refer to the best method combining multiple association measures and the corresponding relative improvement compared to the best individual association measure. The last row refers to the experiment using combination of association measures and information about POS patterns.

7. References

PDT 2.0. 2006. http://ufal.mff.cuni.cz/pdt2.0/.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Univ. of Stuttgart.

1994. The FR corpus is part of the ECI Multilingual Corpus I distributed by ELSNET. See http://www.elsnet.org/eci.html for more information and licensing conditions.

Hannah Kermes. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, IMS, University of Stuttgart.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. Ph.D. thesis, Saarland University.

Wolfgang Lezius, Stefanie Dipper, and Arne Fitschen. 2000. IMSLex - representing morphological and syntactical information in a relational database. In *U. Heid, S. Evert, E. Lehmann, and C. Rohrer (eds.), Proceedings of the 9th EURALEX International Congress*, Stuttgart, Germany.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, Australia.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL 2005 Student Research Workshop*, Ann Arbor, USA.