Automatic Compound Processing: Compound Splitting and Semantic Analysis for Afrikaans and Dutch

Ben Verhoeven
CLiPS - Computational Linguistics
University of Antwerp
Antwerp, Belgium
ben.verhoeven@uantwerp.be

Menno van Zaanen
TiCC, School of Humanities
Tilburg University
Tilburg, the Netherlands
mvzaanen@uvt.nl

Walter Daelemans
CLiPS - Computational Linguistics
University of Antwerp
Antwerp, Belgium
walter.daelemans@uantwerp.be

Gerhard van Huyssteen
Centre for Text Technology (CTexT)
North-West University
Potchefstroom, South Africa
gerhard.vanhuyssteen@nwu.ac.za

Abstract

Compounding, the process of combining several simplex words into a complex whole, is a productive process in a wide range of languages. In particular, concatenative compounding, in which the components are "glued" together, leads to problems, for instance, in computational tools that rely on a predefined lexicon. Here we present the AuCoPro project, which focuses on compounding in the closely related languages Afrikaans and Dutch. The project consists of subprojects focusing on compound splitting (identifying the boundaries of the components) and compound semantics (identifying semantic relations between the components). We describe the developed datasets as well as results showing the effectiveness of the developed datasets.

1 Introduction

In many human language technology applications (e.g. machine translators and spelling checkers), many concatenatively written compounds are processed incorrectly. One of the reasons for this is that these applications rely on a predefined lexicon and the productive nature of the process of compound formation automatically results in incomplete lexicons. For example, consider the novel Afrikaans (Afr.) compound *ministerskatkis* 'treasury of a minister' that should be segmented as *minister+skatkis* **minister+treasury**. Should it be incorrectly segmented as *minister_s+kat+kis* **minister_LINK+cat+coffin**[1] (where LINK refers to a linking morpheme), one would get the (possible but improbable) interpretation 'coffin of a minister's cat'. From a technological perspective, deficiencies related to automatic compound splitting (also known as compound segmentation) are particularly problematic, since many other technologies (such as morphological analyzers, or semantic parsers) might rely on highly accurate compound splitting.

For more advanced natural language processing applications like information extraction, question answering and machine translation systems, proper semantic analysis of compounds might also be required. With semantic analysis of compounds we refer to the task of determining that the Dutch (Du.) compound *keuken+tafel* **kitchen+table** construes 'table in kitchen', while Du. *baby+tafel* **baby+table** means 'table for a baby' (and not, fatally so, *'table in a baby'). Internationally, research on automatic compound analysis has focused almost exclusively on English; very little work in this regard has been done for other languages (see section 4.1).

Concatenative compounding is a highly productive process in many languages of the world, such as West-Germanic languages (Afrikaans, Dutch, Frisian, German, and to a far lesser extent English), Nordic

[1] Note that compound boundaries are marked using a "+" sign and the start of a linking morpheme is indicated by an "_" sign.

Proceedings of the First Workshop on Computational Approaches to Compound Analysis, pages 20–30,
Dublin, Ireland, August 24 2014.

languages (Danish, Icelandic, Norwegian, and Swedish) and Modern Greek; our focus in this research is only on Afrikaans and Dutch. Next to derivation, the process of right-headed, recursive compounding is the most productive word-formation process in these two languages. While almost all parts-of-speech categories can be found as components of compounds, noun+noun compounds are by far the most frequent type, while noun+verb compounding is generally considered to be non-productive in Germanic languages (Don, 2009, p. 378). Components of a compound sometimes need to be "glued" together using linking morphemes. The occurrence of linking morphemes in Afrikaans and Dutch compounds is well-known (Neijt et al., 2010), like Afr. *besigheid_s+besluit* **business_LINK+decision** 'business decision'.

Besides regular compounding, one also finds, amongst others, phrasal compounds (e.g. Afr. *help-my-fris-lyk-hemp* **help-me-strong-look-shirt** 'gym vest'), (neo)classical compounds (e.g. Afr. *neuro+wetenskap* **neuro+science** 'neuroscience', or Du. *bio+logie* **bio+logy** 'biology'), separable verbal compounds (e.g. Du. *op+bellen* **up+call** 'to phone'), reduplicative compounds (e.g. Afr. *speel_+speel* **play_LINK+play** 'easily'), and compounding compounds (i.e. where the two left constituents are normally a phrase, but joined in a compound through the right-most constituent, e.g. Du. *onder+water+camera* **under+water+camera** 'under-water camera'). Except for the latter, none of these marginal types of compounds were considered as data for any of the systems developed in this research project.

In section 2 we provide an overview of the automatic compound processing (AuCoPro) project, which forms the background of this research. Sections 3 and 4 provide details of each of the subprojects relating to compound splitting and semantic analysis, with details about related research, the development of datasets, and our experiments. We conclude with a discussion of results and future work in section 5.

2 Overview: The AuCoPro Project

Running from 2012 to 2013, the AuCoPro project was funded by the Dutch Language Union and the Department of Arts and Culture of the South African Government in a programme to support collaborative research in human language technology between Belgium, The Netherlands and South Africa. Additional funding was provided by the South African National Research Foundation, and the European Network on Word Structure (NetWordS). The partners involved in the project were the University of Antwerp (Belgium), Tilburg University (The Netherlands), and North-West University (South Africa).

The primary aim of the project was to develop resources (including annotation protocols, and training and testing data) for the development of robust compound splitters (subproject 1), and first-generation compound analyzers (subproject 2) for Afrikaans and Dutch, through a combination of cross-language transfer (allowing technology recycling), data pooling, and various machine learning approaches. In a subpart of subproject 2 we also aimed to gain insight in compound semantics by unifying perspectives from computational semantics (Ó Séaghdha, 2008), typological studies (Scalise and Bisetto, 2009), and construction-based approaches to word-formation (specifically cognitive grammar (Langacker, 2008) and construction morphology (Booij, 2010)); the results of which can be found in Van Huyssteen (2014) and Van Huyssteen and Verhoeven (2014).

Deliverables included eight peer-reviewed publications, a technical report on annotation guidelines for compound processing, and six datasets. All deliverables are available in the open-source domain at `https://sourceforge.net/projects/aucopro`, while more information about the project is available at `http://tinyurl.com/aucopro`.

3 Compound Splitting

The aim of subproject 1 was to develop datasets that can be used to build robust compound splitters for Afrikaans and Dutch, or for a cross-lingual analysis of the use of compounds in the closely related languages Afrikaans and Dutch. Based on existing datasets containing words that are morphologically analyzed, we extracted (potential) compounds, removed unwanted morphological information, and reanalysed and corrected them.

In the AuCoPro datasets, compounds are analyzed in a shallow manner: no deep hierarchical ordering of components is performed. Compounds consisting of more than two elements are annotated by indicating the location of the boundaries, so for instance, Du. *bloem+boll_en+veld* **flower+bulb_LINK+field** 'bulb field' consists of four components, viz. *bloem*, *boll-*, *-en-*, and *veld*, without any indication of their syntagmatic relations. The parts *bloem*, *boll-* and *veld* are all simplex words, which we will call constituents. Constituents are the meaningful parts of a compound. These constituents are prototypically independent words, but in some cases affixoids (i.e. forms that are somewhere between a word and an affix in its development) can also occur in compounds (e.g. *boer* in Du. *krant_en+boer* **newspaper_LINK+farmer** 'newspaper seller' does not have the literal meaning of farmer; see Booij (2010)). In some cases a word may undergo morphophonological changes in the context of a compound. For instance, in the *bloembollenveld* example, *boll-* is an allomorph (or allograph) of *bol* 'bulb'.

As mentioned above, some compounds require linking morphemes (indicated by LINK in the examples above) to "glue" components together. Besides ordinary linking morphemes like *-e-*, *-en-*, and *-s-* (in both languages), we also defined hyphens as linking morphemes. In the orthographies of Afrikaans and Dutch in general a hyphen is used in cases of vowel collision, i.e. between compound constituents when the left-hand constituent ends on a vowel, and the right-hand constituent begins with the same vowel, for example Afr. *see_-+eend* **sea_LINK+duck** 'seaduck'.

We also mentioned above that marginal compound types such as phrasal compounds, reduplicative compounds, separable verbal compounds, etc. were not considered as part of the datasets. Similarly, we excluded synthetic compounds from the datasets when the right-hand element of a synthetic compound is a non-word (e.g. in Du. *blauw+ogig* **blue+eye-ADJR**[2] 'blue-eyed', **ogig* is not a valid independent word in Dutch). However, for this subproject we accepted and annotated compounding compounds, since they can generally be split quite easily (e.g. Afr. *drie+vlak+regering* **three+level+government** 'three-level government').

To demonstrate the effectiveness of the developed datasets, we started building and evaluating compound splitters for both Dutch and Afrikaans based on the data only. A compound splitter takes a word as input, and provides as output the input string divided into valid compound components. Note that these results are only to illustrate that these datasets can be used successfully as training data for such systems. The actual results can potentially be improved, as the systems are not optimized.

3.1 Related Research

In general, the problem of splitting compounds is found in a wide range of languages. Some of these languages show non-concatenative compound formation (i.e. compounds are written with whitespaces between constituents), such as English. Compounds in these languages fall under the umbrella term multiword expressions (MWEs), which also includes idioms and collocations. Ramisch et al. (2013) show that this is a quite active research field.

Focusing on concatenative compounding (i.e. where constituents are written conjunctively so that a compound is always written as a single string without any whitespaces), previous work on Afrikaans has been performed in the context of the development of spelling checkers (Van Zaanen and Van Huyssteen, 2002; Van Huyssteen and Van Zaanen, 2004). Van Huyssteen and Van Zaanen (2004) describe a compound splitter for Afrikaans. To our knowledge, no stand-alone compound splitter for Dutch is available. Research done in this field is over ten years old (e.g. Pohlmann and Kraaij (1996)), uses expensive resources (e.g. Ordelman et al. (2003)), does complete morphological analysis (e.g. De Pauw et al. (2004)), and/or has not been released for re-use in the open-source domain.

3.2 Dataset Development

The datasets developed during this subproject are based on compounds taken from existing (morphologically annotated) datasets. For Dutch, a few morphologically annotated datasets exist, although none focus on compounds specifically. The development of the Dutch dataset is based on the e-Lex dataset.[3]

[2] Adjectiviser.

[3] This dataset was extended with a compound dataset extracted from CELEX by Lieve Macken (LT3, UGent).

The e-Lex dataset contains words annotated with more morphological information than required for our dataset, but it also contains morphologically annotated non-compound words. After removing non-compound words (and removing duplicates), 71,274 potential Dutch compounds remained.

For Afrikaans, the situation is more difficult. No dataset containing compound boundary and linking morpheme boundary information is freely available. The Afrikaans AuCoPro dataset is based on the PUK-Protea corpus as well as the CTexT Afrikaans spelling checking lexicon (CTexT, 2005; Pilon et al., 2008). Both corpora do not describe any morphological information. To identify potential compounds, a longest string matching algorithm (Van Huyssteen and Van Zaanen, 2004) is applied. This algorithm identifies compounds by searching for known (simplex) words from the left and right ends of the potential compound, taking the possibility of the occurrence of linking morphemes into account. This algorithm seems to identify most compounds as well as some non-compounds, which resulted in a list of 77,651 potential Afrikaans compounds.

After this automatic collection and cleanup (for Dutch) and automatic identification and annotation (for Afrikaans), annotators checked each compound for correct linking morpheme and compound boundaries. For Afrikaans, seven annotators together checked 25,266 compounds. For Dutch, two annotators checked 26,000 potential compounds. In the end, this resulted in 18,497 and 21,997 true compounds for Afrikaans and Dutch respectively.

To be able to calculate inter-annotator agreement, subsets of approximately 1,000 words were annotated by pairs of annotators. For Dutch in total 6,000 words were used to calculate inter-annotator agreement and for Afrikaans 12,818 words. This leads to an average Cohen's Kappa of 98.6 and 97.6 for Afrikaans and Dutch respectively.

The annotators had access to an annotation manual (Verhoeven et al., 2014), which was developed specifically for this project. The manual is based on the annotation guidelines that were developed during the CKarma project (CTexT, 2005; Pilon et al., 2008). These initial guidelines only apply to Afrikaans, and was hence extended to handle Dutch compounds as well as more complicated cases not foreseen in the original CKarma guidelines. During the annotation process, regular discussions between the annotators took place, which resulted in changes in the data and (minor) modifications to the annotation guidelines.

3.3 Experiments

One of the reasons for creating the compound splitting datasets is to show their usefulness in the development of automatic compound splitting systems. These systems search for compound boundaries, effectively identifying the simplex words in compounds. This information is essential, for instance, when developing spelling correction systems or machine translation systems for languages that have productive compound formation processes.

As a classifier, we used the algorithm developed by Liang (1983). This system, which is used as the hyphenation method in the LaTeX typesetting system, identifies letter combinations that either allow or disallow boundary breaks. Even though the task of compound boundary detection is different from hyphenation (or syllabification), the tasks are similar enough to use the same method. Since the system is trainable, instead of hyphenation breaks, compound boundaries are provided.

Since no separate annotated gold standard test set is available, we performed leave-one-out evaluation (using all but one instance for training and the remaining instance for testing; all instances are evaluated once) using the full dataset. This approach is preferred over, for instance, 10-fold cross validation, which each time removes 10% of the training data for testing. Additionally, it does not depend on a "lucky" selection of test data from the training data, as all compounds are tested.

Evaluating the datasets using this system (which does not have any additional tuning parameters) results in classification accuracies of 88.28% and 91.48% on the word level for Afrikaans and Dutch respectively. We assume that further improvements are possible with alternative systems and parameter optimization.

4 Compound Semantics

The automatic processing of the semantics of compounds (or other complex nominals) is a topic in computational linguistics that, although it has been studied regularly in the past, cannot be considered a solved problem. Although previous research was often promising, it also had an almost exclusive focus on English noun-noun (NN) compounds. In recent years, more languages have been studied (e.g. German (Hinrichs et al., 2013) and Italian (Celli and Nissim, 2009)), and this project added Dutch and Afrikaans to the list.

It is worth noting that a number of different operationalizations of compound interpretation have been studied. The most notable are semantic classification of the constituent relation according to a limited set of semantic categories (e.g. Ó Séaghdha (2008)), and the generation of possible paraphrases for the compound that express its meaning more explicitly (Hendrickx et al., 2013). Our study adopts the classification model, in which the set of semantic relations to be predicted (the classification scheme) is crucial.

4.1 Related Research

Several attempts have been made in the past to postulate appropriate classification schemes for noun-noun compound semantics. These schemes are mainly inventory-based in that they present a limited list of predefined possible classes of semantic relations a compound can manifest.

In some cases, proposed classes are abstractly represented by a paraphrasing preposition (Lauer, 1995; Girju et al., 2005; Lapata and Keller, 2004). For example, all compounds that can be paraphrased by putting the preposition "of" between the constituents belong to the class OF, e.g. a *car door* is the 'door of a car'. Another possibility is using predicate-based classes where the relations between the constituents are not merely described by a preposition, but by definitions or paraphrasing predicates for each class. The class AGENT would contain compounds that could be paraphrased as 'X is performed by Y' (Kim and Baldwin, 2005), e.g. *enemy activity* can be paraphrased as 'activity is performed by the enemy'. Different schemes vary from 9 to 43 classes with Cohen's Kappa scores for inter-annotator agreement ranging from 52% to 62% (Barker and Szpakowicz, 1998; Girju et al., 2005; Moldovan et al., 2004; Nakov, 2008; Ó Séaghdha, 2008).

With regard to the information used by the classifier to assign the classes to the compounds (the features of a compound to be analyzed), two main approaches are available, viz. taxonomy-based methods, or corpus-based methods.

Taxonomy-based methods (also called semantic network similarity (Ó Séaghdha, 2009)) base their features on a word's location in a taxonomy or hierarchy of terms. Most of the taxonomy-based techniques use WordNet (Miller, 1995) for these purposes; especially the hyponym information in the hierarchy is used. A bag of words is created of all hyponyms and the instance vector contains binary values for each feature (the feature being whether the considered word from the bag of words is a hyponym of the constituent or not). Kim and Baldwin (2005) reached an accuracy of 53.3% using only WordNet. Other research was based on Wikipedia as a semantic network (Strube and Ponzetto, 2006).

Corpus-based methods use co-occurrence information of the constituents of the selected compounds in a corpus. The underlying idea (the distributional hypothesis) is that the set of contexts in which a word occurs, is an implicit representation of the semantics of this word (Harris, 1968). The lexical similarity measure assumes that compounds have a similar semantic interpretation when their respective constituents are semantically similar. Two compounds, for example *flour can* and *corn bag* will be considered similar if they have similar modifying constituents (*flour* and *corn*) and similar head constituents (*can* and *bag*). The co-occurrences of both constituents will be combined to calculate a measure of similarity for the entire compound. This approach implicitly uses the lexical semantic knowledge also used in taxonomy-based methods but without the need for a taxonomy. Performances of up to 64% F-score have been reached (Ó Séaghdha and Copestake, 2013).

Corpus-based and taxonomy-based methods have also been combined by several researchers. Accuracies of 58.35% (Ó Séaghdha, 2007), 73.9% (Tratz and Hovy, 2010) and even 82.47% (Nastase et al., 2006) were reported.

4.2 Dataset Development

For this project, we developed datasets of semantically annotated compounds for Afrikaans and Dutch. This section describes these new resources.

The annotation scheme and guidelines that we used as basis, were developed by Ó Séaghdha (2008) for semantic annotation of English NN compounds. For purposes of our project, some adaptations were in order, while Dutch and Afrikaans examples were added (Verhoeven et al., 2014). Ó Séaghdha (2008) describes eleven classes of compounds; six of these classes are semantically specific (see Table 1).

Class	Definition	Example
BE	The compound can be rewritten as 'N2 which is (like) (a) N1' with N1 and N2 being the two constituents nouns.	*woman doctor*
HAVE	The compound denotes some sort of possession. Part-whole compounds, typical one-to-many possession, compounds expressing conditions or properties and meronymic compounds belong here.	*car door*
IN	The compound denotes a location in time or place.	*garden party*
ACTOR	The compound denotes a characteristic event or situation and one of the constituents is a salient entity.	*enemy activity*
INST	The compound denotes a characteristic event and there is no salient entity present.	*cheese knife*
ABOUT	The compound describes a topical relation between its constituents.	*film character*

Table 1: Overview of semantically specific categories in the semantics annotation scheme.

The other five categories are less specific. The MISTAG and NONCOMPOUND categories serve to classify compounds that do not belong in the dataset. The REL class describes compounds with a clear meaning that does not belong to any of the other classes, but of which the relation between the constituents seems productive (e.g. *sodium chloride*). The LEX category is almost the same as REL, but the relation does not seem to be productive (e.g. *monkey business*). The UNKNOWN category is for correct NN compounds of which the meaning is not clear enough to annotate.

As a subpart of this subproject, we also developed an annotation protocol for nominal compounds that do not have a noun as first constituent (XN) (Verhoeven and Van Huyssteen, 2013). Such XN compounds had thus far mostly been neglected, despite the fact that they are fairly productive in some Germanic languages (although far less frequent than NN compounds). Our annotation guidelines followed the general approach of Ó Séaghdha (2008).

In the course of the project, several datasets were developed. For both Dutch and Afrikaans there were two annotation rounds for NN compounds and one smaller annotation experiment for XN compounds. An overview of the semantics data can be found in Table 2, including the average Cohen's Kappa scores.

The Dutch NN compounds were taken from the same raw compound list of 71,274 compounds described in section 3.2 above. Subsequent annotations were performed by students in linguistics at the University of Antwerp, all native speakers of Dutch. The first dataset was annotated by one student, and a subset of 500 compounds by one of the authors in order to calculate inter-annotator agreement. The second round of data was annotated by three students, with the data divided between them in such a way that we had two annotations for each compound. For the XN compound dataset, only 600 compounds were annotated.

The NN compounds for the Afrikaans dataset were taken from the CKarma list of split compounds (see section 3.2 above). The complete Afrikaans dataset was annotated by three undergraduate linguistics students, all native speakers of Afrikaans. This resulted in three annotations for each compound. With regard to the XN compound subpart, a large dataset of 4,553 compounds was annotated.

4.3 Experiments

The data from the first annotation rounds were used for semantic classification experiments that were based on those conducted by Ó Séaghdha (2008). We used the annotations made by the main annotator

language	annotation type	# items	# annotators	avg. Kappa score
Afrikaans	NN-Round1	1,449	3	53.4
Afrikaans	NN-Round2	2,328	3	37.6
Afrikaans	XN	4,553	3	33.5
Dutch	NN-Round1	1,766	2	60.0
Dutch	NN-Round2	2,000	3	51.0
Dutch	XN	600	2	48.6

Table 2: Overview of semantics data.

for each language in order to maintain his or her consistency of annotation. What follows is a description of our own experimental setup. In our classification experiment, classifiers trained by machine learning methods use feature vectors arising from a combination of the distributional hypothesis (as proposed above) with the idea of analogical reasoning. It is assumed that the semantic category of a compound can be predicted by comparing compounds with similar meanings (Ó Séaghdha, 2008).

4.3.1 Vector Creation

For every compound constituent, the co-occurrence context was calculated. For this purpose, for each instance of the constituents in the corpus, the surrounding *n* words (that belong to the 10,000 most frequent words of the corpus) were held in memory. The relative frequencies of these context words (the number of times the word appeared in the context of the constituent, divided by the frequency of the constituent in the corpus) for each constituent were stored.

For Dutch, the Twente News Corpus (Ordelman et al., 2007) was used. This is a 340 million word corpus of newspaper articles. For Afrikaans, we used the Taalkommissie corpus (Taalkommissie, 2011), a 60 million word corpus that consists of a variety of text genres.

A concatenation of the constituent data was used to create the instance vector. This is a new but very simple technique of composition whereby each instance vector thus contains the relative frequencies for the 1,000 most frequent words for each constituent (hence 2,000 per compound). Compounds of which one or both of the constituents did not appear in the corpus were excluded from the data.

The classification experiment dealt with those compounds that were annotated with a semantically specific category. This means that only compounds with the category tags BE, HAVE, IN, INST, ACTOR and ABOUT were used for the experiments. The final vector set for Afrikaans contained 1,439 compounds, while the final vector set for Dutch had 1,447 compounds.

4.3.2 Results

As machine learning method, we used the SMO algorithm, which is WEKA's (Witten et al., 2011) support vector machines (SVM) implementation, in a 10-fold cross-validation setup.

Since this was the first research on both Dutch and Afrikaans (Verhoeven et al., 2012), we assumed a majority baseline which represents the accuracy that can be obtained by always guessing the most frequent class as the output class. For Dutch, this baseline is 29.5% (428 instances of class IN on a total of 1,447 compounds) (Verhoeven, 2012). For Afrikaans, this baseline is 28.2% (407 instances of class ABOUT on a total of 1,439 instances).

The outcome of these experiments showed that the semantic relation between compound constituents in Dutch and Afrikaans can be learned using our simple new composition method of concatenating the constituent vectors into a compound vector. F-scores of 47.8 (Dutch) and 51.1 (Afrikaans) were achieved using the counts of three context words left and right of the constituent for computing their semantic representation. The approach turned out to be robust for varying sizes of context (different numbers of context words), as well as for the way corpus counts were done: on either lemmas or word forms (Verhoeven, 2012; Verhoeven and Daelemans, 2013). Our results are a good improvement of our baselines, and provide a baseline for future research.

4.3.3 WordNet-based method for Afrikaans

In another subpart of this subproject, we experimented with an alternative approach, namely to use the Afrikaans WordNet (CTexT, 2011) to infer compound semantics of Afrikaans compounds (Botha et al., 2013). We followed the same approach as Kim and Baldwin (2005), and achieved precision results similar to the general approach described above. i.e. 50.49% using the Afrikaans WordNet, vs. 50.80% reported by Verhoeven et al. (2012). However, recall was much worse: 29.27% in this approach, vs. 51.60% using the other approach. This poor recall can be attributed to the small size of the Afrikaans WordNet, which only contains 10,045 synsets, compared to 115,424 synsets in the Princeton WordNet (Miller, 1995). We therefore conclude that a WordNet-approach holds much promise, on the premise that the WordNet is large enough to ensure good coverage.

5 Discussion

We described machine learning approaches to the segmentation and semantic interpretation of compounds in Dutch and Afrikaans, two related languages where concatenative compounding is a highly productive morphological process. Success of machine learning approaches to any natural language processing task is based on the presence of sufficient high quality training data and relevant information sources allowing the classification problem to be solved.

For compound splitting, high annotator agreement in the annotation of the training data and high generalization accuracy could be obtained for both languages using a statistical pattern induction method working on the orthography of the input compounds, without need for other information sources. Further improvement can be achieved here with more and richer training data. Other methods for sequence learning could lead to further improvements as well, although Liang's method (1983) turns out to be a strong algorithm for this task.

The task of compound interpretation is much more difficult, both for people (who reached relatively low annotation agreement for both languages) and for machine learners, suggesting that crucial information is missing in the semantic representations we used for our compound constituents. Nevertheless, also for this task, we were able to set a standard, well above baseline, for future work in compound interpretation for Dutch and Afrikaans. Further improvement can potentially be found in many directions: more fine-grained and more learnable semantic relation types, more consistently annotated training data (and much more of it from different domains), and better semantic representations for the constituents, for example using deep learning (Mikolov et al., 2013).

Acknowledgments

The AuCoPro project was funded through a research grant from the Nederlandse Taalunie (Dutch Language Union) and the South African Department of Arts and Culture (DAC), as well as grants from the South African National Research Foundation (NRF) (grant number 81794), and the European Network on Word Structure (NetWordS) (European Science Foundation) (Grant number: 5570). Views expressed in this publication cannot be ascribed to any of these funding organizations.

We would also like to acknowledge the contributions of numerous students and colleagues, including Zandré Botha, Roald Eiselen, Joanie Liversage, Benito Trollip, Nanette van den Bergh (North-West University); Natasja Loyens, Maxim Baetens, Frederik Vaassen (University of Antwerp); Chris Emmery, Suzanne Aussems (Tilburg University).

References

Ken Barker and Stan Szpakowicz. 1998. Semi-Automatic Recognition of Non- Modifier Relationships . *Proceedings of the 17th International Conference on Computational Linguistics*, pages 96–102.

Geert Booij. 2010. *Construction Morphology*. Oxford University Press, Oxford.

Zandré Botha, Roald Eiselen, and Gerhard van Huyssteen. 2013. Automatic compound semantic analysis using wordnets. In *Proceedings of the Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa*, Johannesburg, South Africa.

Fabio Celli and Malvina Nissim. 2009. Automatic Identification of Semantic Relation in Italian complex nominals. In *Proceedings of the Eighth International Conference on Computational Semantics (IWCS-8)*, Tilburg, The Netherlands.

CTexT. 2005. CKarma (C5 KompositumAnaliseerder vir Robuuste Morfologiese Analise). [C5 Compound Analyser for Robust Morphological Analysis]. Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa.

CTexT. 2011. Afrikaans WordNet. Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa.

Guy De Pauw, Tom Laureys, Walter Daelemans, and Hugo Van Hamme. 2004. A Comparison of Two Different Approaches to Morphological Analysis of Dutch. In *Proceedings of the Workshop of the ACL Special Interest Group on Computational Phonology (SIGPHON)*, Barcelona, Spain.

Jan Don. 2009. IE, Germanic: Dutch. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Compounding*, pages 370–385. Oxford University Press, Oxford, UK.

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

Zellig Harris. 1968. *Mathematical structures of language*. Interscience, New York.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 Task 4: Free Paraphrases of Noun Compounds. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Erhard Hinrichs, Verena Henrich, and Reinhild Barkey. 2013. Using Part-Whole Relations for Automatic Deduction of Compound-internal Relations in GermaNet. *Language Resources and Evaluation*, 24(3):363–372.

Su Nam Kim and Timothy Baldwin. 2005. Automatic Interpretation of Noun Compounds Using WordNet Similarity. *Wall Street Journal*, pages 945–956.

Ronald Langacker. 2008. *Cognitive Grammar: A Basic Introduction.* Oxford University Press, New York.

Mirella Lapata and Frank Keller. 2004. The Web as a Baseline: Evaluating the Performance of Unsupervised Web-based Models for a Range of NLP Tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 121–128. Association for Computational Linguistics, Boston.

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

Franklin Mark Liang. 1983. *Word Hy-phen-a-tion by Com-put-er*. Ph.D. thesis, Stanford University, Stanford, USA.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751.

George Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Dan Moldovan, A Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the Semantic Classification of Noun Compounds. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 60–67. MA: Association for Computational Linguistics, Boston.

Preslav Nakov. 2008. Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA08)*.

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*, pages 781–787. MA: American Association for Artificial Intelligence, Boston, aaai-06 edition.

Anneke Neijt, Robert Schreuder, and Carel Jansen. 2010. Van boekenbonnen en feëverhale: De tussenklank e(n) in Nederlands en Afrikaanse samestellingen: vorm of betekenis? [The interfix e(n) in Dutch and Afrikaans compounds: form or meaning?]. *Nederlandse Taalkunde*, 15(2):125–147.

Diarmuid Ó Séaghdha and Ann Copestake. 2013. Interpreting compound nouns with kernel methods. *Journal of Natural Language Engineering, Special Issue on the Semantics of Noun Compounds*, 19:331–356.

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, University of Cambridge, Cambridge, UK.

Roeland Ordelman, Arjan Van Hessen, and Franciska De Jong. 2003. Compound decomposition in Dutch large vocabulary speech recognition. In *Proceedings of Eurospeech 2003*, pages 225–228, Geneva, Switzerland.

Roeland Ordelman, Franciska de Jong, Arjan van Hessen, and Hendri Hondorp. 2007. TwNC: a Multifaceted Dutch News Corpus. *ELRA Newsletter 12*, pages 3–4.

Diarmuid Ó Séaghdha. 2007. Annotating and Learning Compound Noun Semantics. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 73–78. Association for Computational Linguistics, Prague.

Diarmuid Ó Séaghdha. 2009. Semantic classification with WordNet kernels. In *Computational Linguistics*, NAACL-Short '09, pages 237–240. Association for Computational Linguistics.

Sulene Pilon, Martin Puttkammer, and Gerhard Van Huyssteen. 2008. Die ontwikkeling van 'n woordafbreker en kompositumanaliseerder vir Afrikaans. *Literator*, 29(1):21–41.

Renee Pohlmann and Wesley Kraaij. 1996. Improving the precision of a text retrieval system with compound analysis. In *Proceedings of the 7th Computational Linguistics in the Netherlands (CLIN 1996)*, pages 115–129, Eindhoven, The Netherlands.

Carlos Ramisch, Aline Villavicencio, and Valia Kordoni. 2013. Introduction to the special issue on multiword expressions: From theory to practice and use. *ACM Transactions on Speech and Language Processing*, 10(2):1–10.

Sergio Scalise and Antonietta Bisetto. 2009. The classification of compounds. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Compounding*, pages 34–53. Oxford University Press, Oxford.

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA.

Taalkommissie. 2011. Taalkommissiekorpus 1.1. Taalkommissie van die Suid-Afrikaanse Akademie vir Wetenskap en Kuns. Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa.

Stephen Tratz and Ed Hovy. 2010. A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687. Uppsala: Association for Computational Linguistics.

Gerhard Van Huyssteen and Menno Van Zaanen. 2004. Learning Compound Boundaries for Afrikaans Spelling Checking. In *Proceedings of First Workshop on International Proofing Tools and Language Technologies*, pages 101–108, Patras.

Gerhard Van Huyssteen and Ben Verhoeven. 2014. A Taxonomy for Dutch and Afrikaans Compounds. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA)*, Dublin, Ireland.

Gerhard Van Huyssteen. 2014. Morfologie. In Wannie Carstens and Nerina Bosman, editors, *Kontemporêre Afrikaanse Taalkunde*, pages 171–208. Van Schaik Uitgewers, Pretoria, South Africa.

Menno Van Zaanen and Gerhard Van Huyssteen. 2002. Improving a Spelling Checker for Afrikaans. In *Computational Linguistics in the Netherlands 2002-Selected Papers from the Thirteenth CLIN Meeting*, page 143156, Groningen, the Netherlands.

Ben Verhoeven and Walter Daelemans. 2013. Semantic Classification of Dutch Noun-Noun Compounds: A Distributional Semantics Approach. *CLIN Journal*, 3:2–18.

Ben Verhoeven and Gerhard Van Huyssteen. 2013. More Than Only Noun-Noun Compounds: Towards an Annotation Scheme for the Semantic Modelling of Other Noun Compound Types. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, Potsdam, Germany.

Ben Verhoeven, Walter Daelemans, and Gerhard B. Van Huyssteen. 2012. Classification of noun-noun compound semantics in Dutch and Afrikaans. In *Proceedings of the Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2012)*, pages 121–125, Pretoria, South Africa.

Ben Verhoeven, Gerhard Van Huyssteen, Menno Van Zaanen, and Walter Daelemans. 2014. Annotation guidelines for compound analysis. *CLiPS Technical Report Series (CTRS)*, 5.

Ben Verhoeven. 2012. A computational semantic analysis of noun compounds in Dutch. Master's thesis, University of Antwerp, Antwerp, Belgium.

Ian Witten, Eibe Frank, and Mark Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. Elsevier.