ISSN 2033-3544

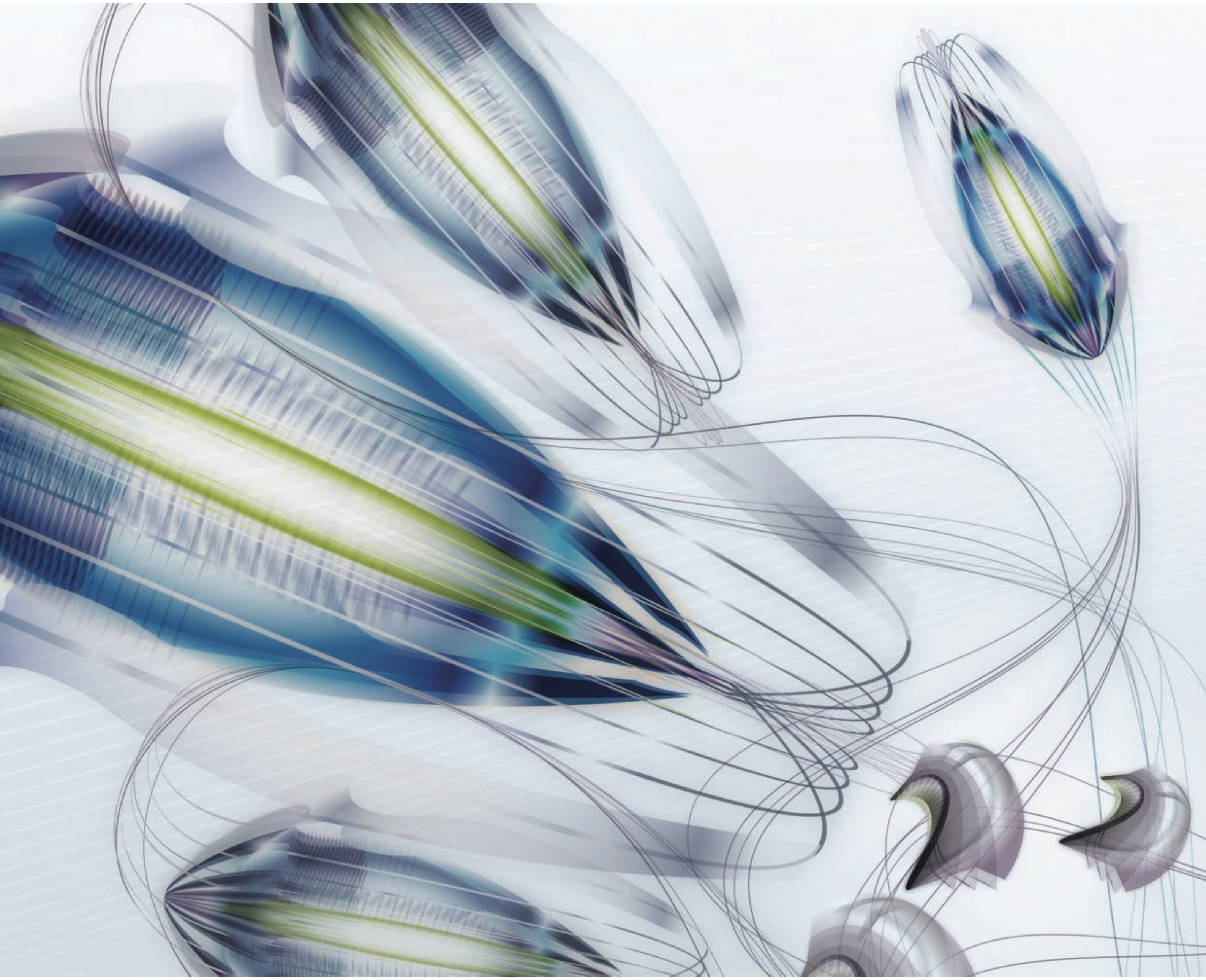
CLiPS 5

COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS
TECHNICAL REPORT SERIES, CTRS-005, JANUARY 2014

WWW.CLIPS.UANTWERPEN.BE/CTRS

Annotation Guidelines for Compound Analysis

Ben Verhoeven, Gerhard B. van Huyssteen,
Menno van Zaanen & Walter Daelemans



CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

Computational Linguistics and Psycholinguistics Research Center
CLiPS Technical Report Series (CTRS)

CTRS-005
January 31, 2014

Annotation Guidelines for Compound Analysis

Ben Verhoeven
CLiPS, University of Antwerp, Belgium
ben.verhoeven@uantwerpen.be

Gerhard van Huyssteen
CTexT, North-West University, South Africa
gerhard.vanhuyssteen@nwu.ac.za

Menno van Zaanen
TiCC, Tilburg University, The Netherlands
menno.vanzaanen@uvt.nl

Walter Daelemans
CLiPS, University of Antwerp, Belgium
walter.daelemans@uantwerpen.be

http://tinyurl.com/aucopro

Lange Winkelstraat 40-42, 2000 Antwerp, Belgium
Phone: (+32) (0)3 265 52 22, Fax: (+32) (0)3 265 58 98

Abstract

This technical report introduces three sets of annotation guidelines for the analysis of compounds in Afrikaans and Dutch. The first protocol serves the annotation of compound boundaries when creating a dataset to use for compound segmentation. The second and third protocol serve the semantic annotation of the relation between the constituents of compounds. Where the second protocol only focuses on noun-noun (NN) compounds, the third protocol deals with other two-part nominal (XN) compounds.

The report further contains a terminology list with definitions of concepts and abbreviations relevant to the analysis of compounds and an overview of the AuCoPro project in the context of which these guidelines were developed.

Contents

Abstract 1

Contents 2

1 Introduction 3
1.1 AuCoPro Project . . . 3
1.1.1 Acknowledgements . . . 4

2 Overview 5
2.1 Compound Splitting . . . 5
2.1.1 Related Research . . . 6
2.1.2 Dataset Development . . . 6
2.1.3 Experiments . . . 7
2.2 Compound Semantics . . . 7
2.2.1 Related Research . . . 8
2.2.2 Dataset Development . . . 9
2.2.3 Experiments . . . 9
2.3 Discussion . . . 11

3 Introducing the Guidelines 12
3.1 Segmentation . . . 12
3.2 Semantic Analysis . . . 12

4 Typographic conventions, terminology and abbreviations 13
4.1 Typographic conventions . . . 13
4.2 Terminology and abbreviations . . . 14

References 21

A Annotation Guidelines: Compound Segmentation 25

B Annotation Guidelines: Semantic Analysis of NN Compounds 33

C Annotation Guidelines: Semantic Analysis of XN Compounds 46

Chapter 1

Introduction

1.1 AuCoPro Project

In many human language technology applications (e.g. machine translation and spelling checking), 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 lexica. 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 analysis, or semantic parsing) might rely on highly accurate compound splitting.

For more advanced HLT applications like information extraction, question answering and machine translation, 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** means '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 2.2.1).

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 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

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

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'), neoclassical compounds (e.g. Afr. *neuro+wetenskap* **neuro+science** 'neuroscience', or Du. *bio+brandstof* **bio+fuel** 'biofuel'), 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 (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.

The primary aim of the Automatic Compound Processing (AuCoPro) 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 analysers (subproject 2) for Afrikaans and Dutch, through a combination of cross-language transfer (i.e. technology recycling), data pooling, and various machine learning approaches.

All datasets are available in the open-source domain and can be retrieved from https://sourceforge.net/projects/aucopro, while more information about the project and all relevant publications are available at http://tinyurl.com/aucopro.

1.1.1 Acknowledgements

We would like to acknowledge the work of many students and collaborators assisting and helping us in this project. Our gratitude goes to: Roald Eiselen, Benito Trollip, Joani Liversage, Zandre Botha, Martin Puttkammer, Martin Schlemmer, Carli de Wet, Nadia Schultz, Nanette Van Den Berg, Sansi Eiselen (North-West University, South Africa); Suzanne Aussems, Chris Emmery, Rick Smetsers, Nanne van Noord, Vincent Lichtenberg, Bas Goris, Sylvie Bruys (Tilburg University, The Netherlands); and Frederik Vaassen, Jana Declercq, Nikolaas Stroobants, Tim Thomaes, Natasja Loyens, and Maxim Baetens (University of Antwerp, Belgium).

We would like to extend this gratitude to the sponsors of this research. The AuCoPro project was co-funded by the Dutch Language Union (Belgium, The Netherlands) and the Department of Arts and Culture (South Africa). Additional funding was obtained from the National Research Foundation (South Africa) (Grant number: 81794) and the European Network on Word Structure (NetWordS) (European Science Foundation) (Grant number: 5570).

None of the views expressed in this report or any other publication can be attributed to any of these funders; all intellectual property and responsibility remain that of the authors.

Chapter 2

Overview

2.1 Compound Splitting

The aim of subproject 1 was to develop datasets that can be used, for instance, 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 analysed, we extracted (potential) compounds, removed unwanted morphological information, and reanalysed and corrected them.

In the AuCoPro datasets, compounds are analysed 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'). 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 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. 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'). We also excluded synthetic compounds from the datasets, since the right-hand element of a synthetic compound is by definition always a non-word (e.g. in Du. *blauw+ogig* **blue+eye-ADJR** 'blue-eyed', **ogig* is not a valid independent word in Dutch.

To demonstrate the effectiveness of the developed datasets, we have built initial 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 in 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 results are not necessarily state-of-the-art, as we do not optimize the systems.

2.1.1 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, such as English. Compounds in these languages fall under the umbrella term multi-word 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, previous work on Afrikaans has been performed in the context of the development of spelling checkers (Van Zaanen & Van Huyssteen, 2002; Van Huyssteen & Van Zaanen, 2004). Van Huyssteen & 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 & 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.

2.1.2 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. For Afrikaans, the situation is more difficult. No dataset containing compound boundary and linking morpheme boundary information is freely available.

The development of the AuCoPro dataset for Dutch is based on the e-Lex dataset.[1] 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 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 compound information. To identify potential compounds, a longest string matching algorithm (Van Huyssteen & 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 link-

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

ing 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 as described in Section 3.1 of this technical report, which was developed specifically for this project. The annotation manual can be found as Appendix A.

2.1.3 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 machine learner, we used the algorithm developed by Liang (1983). This system is used as the hyphenation method in the LaTeX typesetting system. 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 the full dataset. This approach is preferred over, for instance, 10-fold cross validation, as that effectively removes 10% of the training data. 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 (van Zaanen *et al.* , 2014). We assume that further improvements are possible with alternative systems and parameter optimization.

2.2 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, the semantic processing of compounds has been studied in more languages (e.g. German (Hinrichs *et al.* , 2013) and Italian (Celli & 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.

2.2.1 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 (Girju *et al.* , 2005; Lapata & 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 & 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 & 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 analysed), 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 & Baldwin (2005) reached an accuracy of 53.3% using only WordNet. Other research was based on Wikipedia as a semantic network (Strube & 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. Accuracies of 54.98% (Ó Séaghdha & Copestake, 2007) and 61% have been reached (Ó Séaghdha, 2008).

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

2.2.2 Dataset Development

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

For both Dutch and Afrikaans there were two annotation rounds for NN compounds (using the annotation guidelines in Appendix B) and one smaller annotation experiment for XN compounds (using the annotation guidelines in Appendix C). This section is a summary of all the semantics data; compare also Table 2.1 for an overview of the data development, 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 2.1.2 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 splitted compounds (see section 2.1.2 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.

language	annotation type	# items	# annotators	avg. Kappa score
Afrikaans	NN-Round1	1449	3	53.4
Afrikaans	NN-Round2	2328	3	37.6
Afrikaans	XN	4553	3	33.5
Dutch	NN-Round1	1766	2	60.0
Dutch	NN-Round2	2000	3	51.0
Dutch	XN	600	2	48.6

Table 2.1: Overview of semantics data.

2.2.3 Experiments

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

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.

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 1447 compounds) (Verhoeven, 2012). For Afrikaans, this baseline is 28.2% (407 instances of class ABOUT on a total of 1439 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 & Daelemans, 2013). Our results are a good improvement of our baselines, and provide a baseline for future research.

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 & 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.

2.3 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 (i.e. the context in which they appear). 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).

Chapter 3

Introducing the Guidelines

3.1 Segmentation

For the analysis of Afrikaans compound boundaries, an annotation guidelines document already existed, which had been used for the annotation of compound boundaries within the CKarma project (CTexT, 2005; Pilon *et al.* , 2008). This document has been used as the basis for our current annotation guidelines for both Afrikaans and Dutch. Our new proposal of guidelines for the annotation of compound boundaries can be found in Appendix A.

3.2 Semantic Analysis

For the semantic annotation of compounds, we started out by using the annotation guidelines of Ó Séaghdha (2008). These guidelines are developed to describe the semantic relation between the constituents of two-noun compounds (NN). We then adapted these guidelines to include Dutch and Afrikaans examples. We also made some changes to which types of compounds will be annotated. More details on the adaptation of these guidelines can be found in chapter 4 of Verhoeven (2012). To help annotators getting acquainted with the guidelines, we developed a classification scheme with paraphrasing prepositions and predicates as well as a decision tree. These tools can be found on the project website [1]. The annotation guidelines for NN compounds can be found in Appendix B.

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 & 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). These annotation guidelines also follow the general approach of Ó Séaghdha (2008). They can be found in Appendix C.

[1] http://tinyurl.com/aucopro

Chapter 4

Typographic conventions, terminology and abbreviations

Terminology used in this project is by and large compatible with the definitions discussed in Van Huyssteen (2010), unless indicated differently.

4.1 Typographic conventions

Glossing is based by and large on the conventions and abbreviations used in the Leipzig Glossing Rules [1]. Since the hyphen ("-") is often used in Afrikaans and Dutch orthography, we use the plus symbol ("+") to distinguish between independent words in compounds, and the underscore ("_") for boundaries with affixes.

In-text examples

LangAbr. *WordInLanguage* **Gloss** 'meaning/translation'
For example:
Du. *massa+gebed* **mass+prayer** 'collective prayer'
Afr. *geel+kleur_ig* **yellow+colour_ADJR** 'yellow coloured'

Numbered examples

(#) LanguageName
WordInLanguage
Gloss
'meaning/translation'

For example:

(9) Dutch
massa+gebed
mass+prayer
'collective prayer'

[1] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

(10) Afrikaans
geel+kleur_ig
yellow+color_ADJR
'yellow coloured'

Analyses

For linguistic analyses, we use the bracketing convention as used by Booij (2010). In theory-specific publications, we use the conventions consistent with the specific theory.

Du. [[massa]$_N$ [gebed]$_N$]$_N$
Afr. [[geel kleur]$_{NP}$ ig]$_A$

When presenting analyses as they would occur in a dataset (as opposed to linguistic glosses in protocols or documentation), we use the lettertype Courier to indicate that it is an example from a dataset, e.g. "In our dataset Afr. *sit+kamer* **sit+room** 'lounge' is analysed as sit + kamer."

4.2 Terminology and abbreviations

A Adjective/Adverb

ADJ Adjective

ADJR Adjectiviser

ADV Adverb; see also B

ADVR Adverbialiser

AR Adjectiviser or adverbialiser; see also **ADJR** and **ADVR**

affixoid A wordform that also functions as an affix; an affixlike word, or wordlike affix. Typically, when the wordform is used as an affix it has undergone some systematic meaning shift. For instance, *boer* in Du. *melk+boer* **milk+farmer** 'milk delivery man'; *laat* in Afr. *laat+herfs* **late+autumn** 'end of autumn'.

allomorph A semantically independent and phonologically dependent word, which has undergone graphemic alteration due to morphonological processes. For instance, Du. *pann_en+koek* **pan_LINK+cake** 'pancake', where *pann-* is an allomorph of *pan*; also, in Du. *schap_en+staart* **sheep.SG_LINK+tail** 'sheep's tail', where *schap-* is an allomorph of *schaap*.

ATAP compounds "The two ATAP classes are meant to include formations featuring a differently expressed attribution relation" (Scalise & Bisetto, 2009: 51). Head is modified by a non-head expressing a property of the head. (This could also be understood in terms of what Lakoff and Johnson called "similarity-based metaphors".)

- **ATAP: Attributive**

Role of non-head element is to express a property ("quality" - Scalise & Bisetto, 2009: 51) of the head. The non-head matches at least one of the encyclopedic features (and not semantic frame roles) of the head; in other words, it "has the sole function of specifying a trait of the... head" (Scalise & Bisetto, 2009:49), i.e. the non-head IS a property of the head. The non-head is typically an adjectival or adverbial element.

- Prototypical: A+N (e.g. Afr. *wit+wyn* **white+wine** 'white wine'; Eng. *high school*), AP+N (Du. *kant_-+en_-+klaar+maaltijd* **ready_-+and_-+done +meal** 'ready-to-go meal')
- Prototypical: Adv+V (Afr. *heen+gaan* **away+go** 'departure'), Adv+N (Afr. *terug+reis* **back+journey** 'homeward journey'), Adv+A (Du. *niet_-+pro duktief* **not_-+productive** 'non-productive').
- Prototypical: Phrasal compounds where non-head is a phrase with an adjectival function (i.e. CP+N), e.g. Du. *doe_-+het_-+zelf+winkel* **do_-+it_-+self+shop** 'DIY store'.

- **ATAP: Appositive**

The non-head element contains a property of the head in its semantic structure (cf. body of skeleton (Scalise & Bisetto, 2009: 49)). In its most prototypical form, the non-head is a noun, an apposition, acting as an attribute (Scalise & Bisetto, 2009: 51). The noun thus has an 'adjectival' function because it expresses a property, rather than a "thing". E.g. in *snail mail*, the property SLOW in "snail" matches the feature TAKES TIME of "mail".

- Prototypical: N+N (Du. *hoofd+beginsel* **head+principle** 'main principle'; Afr. *gunsteling+boek* **favourite+book** 'favourite book'; Eng. *key word*) [2].

Other appositive compounds can have adjectives or verbs as non-head. The compound is then an intensive form of the head, where the head is closely associated with a property of the non-head.

- V+A compounds[3], e.g. Du. *druip+nat* **drip+wet** 'dripping wet; as wet as something/someone that drips'; the verbal element acts as modifier, but is at the same time distinguished from attributive compounds that generally take adjectives as non-heads.
- N+A compounds, e.g. Du. *ijs+koud* **ice+cold** 'ice cold; cold as ice', Afr. *hond_s+getrou* **dog_LINK+loyal** 'loyal as a dog'.

B Adverb (when distinguishing between adjectives and adverbs)

base(-word) see **simplex**

complex A word consisting of more than one morpheme, such as stem_suffix, or stem +stem. For example, Du. *melk+koe* **milk+cow** 'milking cow'; Afr. *vrees_lik* **fear_ADJR** 'terrible'.

[2] Other examples: Afr. *trefferliedjie, bielieslot, reusemislukking*; Eng. *satellite nation, ape man, ghost writer, swordfish, mushroom cloud*

[3] Other examples: Du. *bouwrijp, fonkelnieuw, hapklaar, kraakhelder, snikheet, spilziek*

component An element (or "building block") in a compound. This can be either a word, allomorph, affixoid, neo-classical stem or a linking morpheme. Also see **constituent**.

compound A complex word consisting of two or more constituents (or words). We distinguish four major syntactic-semantic types of compounds, viz. attributive compounds (see ATAP), appositive compounds (see ATAP), coordinate compounds, and subordinate compounds (Scalise & Bisetto, 2009). Constructs such as synthetic compounds, particle verbs and compounding compounds are considered peripheral phenomena of compounding in general.

compounding compounds A compound with three constituents, where the first and second constituents form a word-group, but when it combines with a third constituent (the head) the whole is considered a compound. E.g. Du. *hoge+hak+schoen* **high+heel+shoe** 'high-heel shoe'; Afr. *intellektuele+goedere+reg* **intellectual +property+law** 'intellectual property law'. Note that *hoge hak* and *intellektuele goedere* are word groups that cannot form compounds on their own. In Afrikaans this is called 'samestellende samestelling', and in Dutch 'samenstellende samenstelling'.

- Prototypical: PP+N (Du. *onder+water+camera* **under+water+camera** 'under-water camera')
- Prototypical: NP+N (Afr. *derde+jaar+student* **third+year+student** 'third-year student')

Note that phrasal compounds (CP+N) are not compounding compounds, because the phrase has more than two units.

constituent An element (or "building block") of a compound that has semantic content; in other words all components, excluding linking morphemes, are constituents. Also see **component**.

coordinate compounds Coordinate compounds show a high level of matching features in the encyclopaedic body of the constituents (i.e. they are semantically similar), as well as matching of grammatical features (e.g. same POS); the constituents are therefore syntactically virtually identical. Both constituents are considered heads of the compound. The compound can usually be paraphrased as: constituent 1 AND constituent 2.

- Prototypical: N+N (Eng. *producer-director*); A+A (Eng. *deaf-mute*); V+V (Eng. *stir-fry*)
- Prototypical in Afrikaans: N+N (Afr. *winkel_-+winkel* **shop_-+shop** 'children's game pretending to work in a shop'); V+V (Afr. *klop_-+klop* **knock_- +knock** 'knocking') (reduplicative compounds)

endocentric vs exocentric compounds A compound is considered either semantically or syntactically exocentric when it does not fulfill the same semantic or syntactic function (e.g. does not refer to same type of entity or POS category) as one of its parts. Exocentric compounds are always lexicalised, while endocentric compounds are mostly non-lexicalised (but might of course also be lexicalised).

extA Adjectiviser or adverbialiser; see also **AR**, **ADJR** and **ADVR**

extN Nominaliser; see also **NR**

extV Verbaliser; see also **VR**

interfix see **linking morpheme**

lexeme see **word**

lexicalised unit Well-established unit with a non-compositional meaning, e.g. Afr. *padda+stoel* **frog+chair** ‘mushroom’. Lexicalised words are generally found as defined or undefined lemmas in dictionaries.

LINK linking morpheme

linking morpheme A morpheme (specifically paramorpheme) that is used between two constituents in a compound. For purposes of this project, we consider the hyphen (“-”) as a linking morpheme. Synonyms include interfix (not infix) and valence morpheme.

Prototypical linking morphemes in Afrikaans and Dutch are ‘s’, ‘e’, ‘en’, and ‘-’.

morpheme A simplex symbolic (i.e. (grammatically) meaningful) unit in the language system. It is simplex in the sense that it does not contain smaller symbolic units as subparts. In other words, a morpheme is the smallest language unit with a form and meaning. For instance, Eng. *book* and Eng. plural ‘-s’ are morphemes.

N Noun

NP Noun phrase

neoclassical stem A semantically partially independent and phonologically dependent morpheme from Greek or Latin origin, e.g. Afr. *fono-* and *-logie* in *fono+logie* **phono+logy** ‘phonology’, or Eng. *bio-* in *bio-diesel*.

non-lexicalised unit Opposite of lexicalised unit, i.e. these units are generally not lemmas in a dictionary; e.g. Du. *scherm+rand* **screen+edge** ‘the edge of the screen’.

NR Nominaliser

paramorpheme A morpheme that extends so far away from the prototypical morpheme, that it almost seems non-morphemic. It is nonetheless still considered to be a morpheme, since we could analyze it as a symbolic unit. Examples include linking morphemes (with phonological content, but highly schematic semantic content), and zero morphemes (with semantic content, but highly schematic phonological content).

particle verbs A compound consisting of a preposition and a verb, where the preposition functions as a particle that could be separated from the verb syntactically or morphologically. It usually has a non-compositional meaning. For example, Du. *op+zettten* **up+set** ‘to set up’ can be separated syntactically (E.g. Du. *We zetten de tent op*. ‘We are setting up the tent.’) or morphologically (*op_ge+zet* **up_PRTC+set** ‘set up.PRTC’). Also called separable compound verbs (Afr. ‘samekoppeling’ and Du. ‘samenkoppeling’).

phrasal compound see **subordinate compound**

PREP Preposition

PL Plural

SG Singular

simplex A semantically and phonologically independent morpheme (i.e. a morpheme without any inflectional or derivational affixes, or other morphemes). In compounds, all non-final constituents are considered to be in singular form, unless it is clear from the context that it is in the plural. Consider for example Afr. *universiteit+span* **university+team** 'team of a university', vs. *universiteite+span* **universities+team** 'team made up of players from various universities'. (Note in the latter case that the '-e' is analysed as PL and not as LINK.) For purposes of this project, note that there are a few cases where non-final constituents could be inflected, e.g. Du. *procureurs_-+generaal* **attorneys_LINK+general** 'attorneys' general'. Synonyms for simplex include **base(-word)**.

sN Neoclassical stem

Subordinate compound The interpretation of subordinate compounds depends on the possibility of associating the different encyclopedic pieces of information in the frame structure (or argument structure; see Plag, 2003: 149) of the two constituents. In other words, constituent A fits in a semantic role (or empty slot in the argument structure) of constituent B, e.g. as an object, subject, instrument, part, location, etc., but not as a feature/charateristic. It is important to note that this relation is one of FULFILLING: if constituent B (the head; e.g. *cake*) requires ingredients, then constituent A (the non-head; e.g. *apple*) must have the property that it could be an ingredient; A therefore fulfils a semantic property of B. It is also important to note that if A is a trait/characteristic of B, then the compound is an ATAP (e.g. Afr. *witwyn* 'white wine' is ATAP, since 'white' describes a trait (the colour) of the wine).

- **Subordinate: Ground**

Prototypical subordinate compounds are usually ground compounds, where the constituents are in their root form - i.e. not derived or inflected.

- Prototypical: N+N (Afr. *tafel+poot* **table+leg** 'leg of a table'; Eng. *chimney sweep, sunrise, boat ride* (Lieber, 2009: 361));
- Prototypical: N+V (Afr. *stof+suig* **dust+suck** 'to vacuum'; Eng. *head-hunt, machine-wash, spoon-feed* (Lieber (2009:361) calls these "verb-containing"));
- Prototypical: V+N (Du. *kook+tijd* **cook+time** 'duration for food to be cooked'; Eng. *kick-ball, attack dog, skate park* (Lieber, 2009: 361)).

- **Subordinate: Verbal-Nexus**

The head of the compound is always a deverbal noun/adjective; note that the derivation from verb to other part-of-speech should be morphologically overt (i.e. nouns derived through conversion are not considered as verbal nexus, but rather as ground - e.g. Eng. *kick-ball, attack dog, skate park*).

  - Prototypical: (N+(V+NR)$_N$)$_N$; e.g. Du. *roman+schrijv_er* **novel+write_NR** 'novelist', *weer_s+voorspell_ing* **weather_LINK+predict_NR** 'weather forecast' These are compounds with derived words as second constituents.
  - Prototypical: ((V+NR)$_N$+N)$_N$; e.g. Afr. *koöpter_ing_s+beleid* **co-opt_NR _LINK+policy** 'co-optation policy', *besprek_ing_s+kantoor* **reserve_NR _LINK+office** 'reservations office'; Eng. *swimming pool, dining room*
  - Prototypical: (N+(V+AR)$_A$)$_A$; e.g. Du. *computer+ge_stuur_d* **computer +AR_control_AR** 'computer controlled', *hand+be_schilder_d* ; Eng. *university-controlled, hair-raising, Washington-based, awe-inspiring* (see Plag, 2003: 153-154)
  - Prototypical: (V+(V+NR)$_N$)$_N$ e.g. Du. *eet+stak_ing* **eat+strike_NR** 'hunger strike'

- **Subordinate: Neoclassical Compounds**

One or more of the components are neoclassical stems.

  - Prototypical: sN+sN (Du. *filan+troop* **philan+tropist** 'philantropist'), sN+N (Afr. *ekso+skelet* **exo+skeleton** 'exoskeleton'), N+sN (Du. *insect_o+loog* **insect_LINK+logist** 'entomologist/insectologist')

- **Subordinate: Phrasal Compounds**

The first component is a phrase, usually consisting of more than two words.

  - Prototypical: NP+N (Du. *dag_-+en_-+nacht+slot* **day_-+and_+night+ lock** 'day and night lock')

synthetic compound A compound that is formed on the basis of a phrase (whether NP, VP or PP) that is concatenated through a derivational process (prototypically suffixation); i.e. compounding (the concatenation) and derivation takes place simultaneously (Booij, 2010: 50). A definitive characteristic of a synthetic compound is that the component with the affix can not exist as an independent word, e.g. Afr. *besluit+nem_ing* **decision+take_NR** 'decisionmaking', where *neming* is not an independent word, or Du. *rood+kleur_ig*red+colour_ADJRred coloured., where *kleurig* is not an independent word. These words are therefore analysed as [[besluit neem]$_{VP}$ ing]$_N$ and [[rood kleur]$_{NP}$ ig]$_A$ respectively. Within this definition, cases like Afr. *bus+bestuurd_er* **bus+drive_NR** 'bus driver' and Du. *bloem+kwek_erij* **flower+grow_NR** 'flower nursery' are not regarded as synthetic compounds, since *bestuurder* 'driver' and *kwekerij* 'nursery' are independent words; hence these two cases are rather analysed as [[bus]$_N$ [[bestuur]$_V$ der]$_N$]$_N$ and [[bloem]$_N$ [[kweek]$_V$ erij]$_N$]$_N$. (i.e. derivation and compounding do not take place simultaneously). As a practical test one can paraphrase Du. *busbestuurder* as 'bestuurder van 'n bus' ('driver of a bus'), but one can not paraphrase Afr. *besluitneming* as 'neming van 'n besluit' ('taking of a decision'). The same test will help to indicate that Afr. *ter+aarde+bestell_ing* **to+earth+deliver_NR** 'burial' is a synthetic compound, since the word can not be paraphrased as 'bestelling van ter aarde' ('delivery of to earth', or something similar), despite the fact that *bestelling* is an independent word in Afrikaans (notably with a different meaning, viz. 'an order').

- Prototypical: $((N+V)_{VP}+NR)_{N}$; e.g. Du. *naam+gev_ing* **name+give_NR** 'name giving'
  Paraphrase as: "[V-PART/INF]N2 of N1" (where N2 is not an independent word in Afrikaans and Dutch)
- Prototypical: $(((P+N)_{PP}+V)_{V}+NR)_{N}$; e.g. Afr. *ten+toon+stell_ing* **for+show +offer_NR** 'exhibition'
- Prototypical: $((A+N)_{NP}+ADJR)_{A}$; e.g. Du. *blauw+og_ig* **blue+eye_ADJR** 'blue-eyed'; Afr. *lang+drad_ig* **long+thread_ADJR** 'tedious; clear-sighted' (see Plag, 2003: 153)

V Verb

valence morpheme see **linking morpheme**

VP Verb phrase

VR Verbaliser

word A simplex or complex symbolic unit in the language system, larger than a morpheme and smaller than a phrase, and consists of a (relatively) stable, integral and promiscuous phonological structure associated with a (relatively) stable semantic structure. Words can be simplex symbolic structures, just like morphemes, or complex in that they could contain smaller symbolic assemblies as subparts. Synonyms for word include lexeme and word-form.

word-form see **word**

References

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

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

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

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.

De Pauw, Guy, Laureys, Tom, Daelemans, Walter, & Van Hamme, Hugo. 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).*

Don, Jan. 2009. IE, Germanic: Dutch. *Pages 370–385 of:* Lieber, Rochelle, & Štekauer, Pavol (eds), *The Oxford Handbook of Compounding*. Oxford, UK: Oxford University Press.

Girju, Roxana, Moldovan, Dan, Tatu, Marta, & Antohe, Daniel. 2005. On the Semantics of Noun Compounds. *Computer Speech and Language*, **19**, 479–496.

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

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

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

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

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

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

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

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

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

Nakov, Preslav. 2008. Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. *In: Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'08)*.

Nastase, Vivi, Sayyad-Shirabad, Jelber, Sokolova, Marina, & Szpakowicz, Stan. 2006. Learning Noun-Modifier Semantic Relations with Corpus-based and WordNet-based Features. *Pages 781–787 of: Proceedings of the 21st National Conference on Artificial Intelligence*, aaai-06 edn. Boston: MA: American Association for Artificial Intelligence.

Neijt, Anneke, Schreuder, Robert, & Jansen, Carel. 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.

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

Ordelman, Roeland, Van Hessen, Arjan, & De Jong, Franciska. 2003. Compound decomposition in Dutch large vocabulary speech recognition. *Pages 225–228 of: Proceedings of Eurospeech 2003*.

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

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

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

Ó Séaghdha, Diarmuid, & Copestake, Ann. 2007. Co-occurrence Contexts for Noun Compound Interpretation . *Pages 57–64 of: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Prague: Association for Computational Linguistics.

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

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

Ramisch, Carlos, Villavicencio, Aline, & Kordoni, Valia. 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.

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

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.

Talmy, L. 2000. The semantics of causation. *In: Toward a Cognitive Semantics, Volume 1: Concept Structuring Systems*. Cambridge, MA, USA: MIT Press.

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

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

Van Zaanen, Menno, & Van Huyssteen, Gerhard. 2002. Improving a Spelling Checker for Afrikaans. *Page 143–156 of: Computational Linguistics in the Netherlands 2002—-Selected Papers from the Thirteenth CLIN Meeting.*

van Zaanen, Menno, van Huyssteen, Gerhard, Aussems, Suzanne, Emmery, Chris, & Eiselen, Roald. 2014. The Development of Dutch and Afrikaans Language Resources for Compound Boundary Analysis. *In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014).*

Verhoeven, Ben. 2012. *A computational semantic analysis of noun compounds in Dutch.* M.Phil. thesis, University of Antwerp, Antwerp, Belgium.

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

Verhoeven, Ben, & Van Huyssteen, Gerhard. 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.*

Verhoeven, Ben, Daelemans, Walter, & Van Huyssteen, Gerhard B. 2012. Classification of Noun-Noun Compound Semantics in Dutch and Afrikaans. *Pages 121–125 of: Proceedings of the Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2012).*

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

Appendix A

Annotation Guidelines for Compound Segmentation

1. Introduction

This document describes annotation guidelines that are used within the Automatic Compound Processing (AuCoPro) project for the annotation of boundaries in compounds. This project is a collaboration between Tilburg University (Tilburg, the Netherlands), University of Antwerp (Antwerp, Belgium) and North-West University (Potchefstroom, South Africa). The project is funded jointly by the Nederlandse Taalunie (Dutch Language Union) as the Belgian/Dutch sponsor and the Department of Arts and Culture as well as the South African National Research Foundation (grant number 81794) as the South African sponsors.

Within the AuCoPro project, research is performed in the context of compound processing, both on Dutch and Afrikaans natural language data. Firstly, annotated data is produced that indicates boundaries between elements in compounds. This document describes the annotation guidelines for this part of the project. Secondly, annotation of the semantic relations between the elements of compounds are annotated, which is described in another document, and which can be found on either of the project's websites (http://tinyurl.com/aucopro and/or https://sourceforge.net/projects/aucopro/). The technical report of the project provides more context and details, including a terminology list of terms, as well as guidelines with regard to glosses used in this document.

For the analysis of Afrikaans compound boundaries, an annotation guidelines document exists, which has been used for the annotation of compound boundaries within the CKarma project (CText, 2005; Pilon and Puttkammer, 2008). This document has been used as the basis for these current annotation guidelines for both Afrikaans and Dutch. Note that since the original CKarma guidelines have been modified in the AuCoPro project, the AuCoPro Afrikaans data differs from the CKarma data; the annotation of the Afrikaans and Dutch AuCoPro data follow each other as closely as possible.

2. Theoretical background

In Dutch and Afrikaans, there are several processes that allow for the creation of words based on base words. Generally, we identify two major types of word-formation processes, viz. derivation and compounding (Booij, 2007). Derivation is the process of extending a simplex (for instance Afr. *man* 'man') with an affix. This may be, amongst other things, a prefix (Afr. *be_man* **VR_man** 'to man') or a suffix (Afr. *man_lik* **man_ADJR** 'manly'). This process may be applied recursively.

The process of compounding is what concerns us in this document. Compounding is the creation of a new word based on two (or more) words (which may be simplexes or complexes). For instance, Du. *deur+bel* **door+bell** 'door bell' is created from the two simplexes *deur* 'door' and *bel* 'bell'. In other words, a compound is a word consisting of parts that can function as words by themselves. The process of compounding may also be repeated a number of times (in theory an infinite number of times, but in practice this is a process of limited recursiveness).

Haeseryn et al. (1997) state that it is unclear whether the process of synthetic compounding should be identified as a third category of word-formation, or whether it can be subsumed under either the process of derivation or that of compounding, or as a simple combination of both. Examples of synthetic compounds are Du. *vijf+jaar_s* **five+year_ADVR** 'five-yearly' (where *vijf jaar* is an NP, written disjunctively outside the context of a compound), or Afr. *besluit+nem_ing* **decision+take_NR** 'decision making' (where *besluit neem* is a VP, also written disjunctively in other contexts). In this project, synthetic compounds are not analyzed.

In this project, we analyze the elements of a compound in a shallow manner. No deep hierarchical ordering of parts is performed. This means that when a compound consists of more than two elements, these all occur on the same level. For instance, Du. *paard_en+bloem+wijn* **horse_LINK+flower+wine** 'dandelion wine' consists of four components, viz. *paard, -en-, bloem,* and *wijn*, without any indication of their syntagmatic relations. The parts *paard, bloem* and *wijn* are simplexes, which we will call constituents (i.e. meaningful components in 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. Du. *krant_en+boer* **newspaper_LINK+farmer** 'newspaper seller'). In this example, the *-en-* element is called a linking morpheme, a component that is required to "glue" the constituents together.

In some cases a word may undergo morphophonological changes in the context of a compound. For instance, in Du. *bot_en+schuur* **boat_LINK+shed** 'boat shed', the dependent constituent *bot-* is an allomorph of the independent word *boot* 'boat'.

3. Scope

Compounds annotated in this project are restricted by their parts-of-speech. Only compounds of the part-of-speech types: noun, verb, adjectives and adverbs are annotated. The constituents of the annotated compounds do not have any restrictions of their part-of-speech. This means that function words, such as Du. *omdat* 'because' (conjunctive), or Afr. *jouself* 'yourself' (pronoun), are not annotated.

However, not all nouns and verbs are annotated. Particle verbs like Du. *om+kopen* **around+buy** 'bribe', and synthetic or derived nouns, such as Afr. *besluit+nem_ing* **decision+take_NR** 'decision making' are not annotated. However, affixoids are identified as constituents, so, for instance, Du. *kei+mooi* **stone+beautiful** 'very beautiful' is annotated as kei + mooi, and Afr. *onder+voorsitter* **under+chairperson** 'vice chair' is annotated as onder + voorsitter.

In principle, lexicalised compounds are not annotated. However, when the relation between the constituents in a lexicalised compound (found in a dictionary) is opaque, it may be added to the dataset and be analysed.

4. Methodology

Two datasets have been developed during this project. Each dataset is a list of compounds. (Non-compounds are not included.) One dataset consists of Dutch compounds and the other of Afrikaans compounds.

The compounds will be extracted from existing corpora. For Afrikaans, the compounds are based on the list from the CKarma project. Additionally, the NCHLT corpus (http://rma.nwu.ac.za/) have been used to extract Afrikaans compounds. For Dutch, compounds from eLex (http://tst-centrale.org/nl/producten/lexica/e-lex/7-25), combined with a compound list created by Lieve Macken (LT3, Ghent University), compounds extracted from the Lassy corpus (http://tst-centrale.org/nl/producten/corpora/lassy-groot-corpus/6-67).

5. Dataset

The dataset consists of a list of compound words. The dataset is in plain text (utf-8) format with each compound on a new line. Each compound is analyzed and annotated. The boundary between constituents in the compound or between a constituent followed by a linking morpheme and a following constituent is done using a plus sign: “+”. The boundary between a word and a linking morpheme is indicated using an underscore: “_”. Note that a linking morpheme is morphologically more closely associated with the left-hand constituent. This is indicated using an underscore on the left-hand side of a linking morpheme. Both “+” and “_” are surrounded by spaces (indicated by “␣”).

In extended Bachus-Naur form, the dataset can be described as follows (where eol represents an end-of-line):

⟨*dataset*⟩	::=	(⟨*compound*⟩ eol)*
⟨*compound*⟩	::=	⟨*constituent*⟩ ␣+␣⟨*restcompound*⟩
	|	⟨*constituent*⟩ ␣_␣⟨*linking*⟩ ␣+␣⟨*restcompound*⟩
⟨*restcompound*⟩	::=	⟨*constituent*⟩
	|	⟨*compound*⟩

In this grammar, the terms ⟨*constituent*⟩ and ⟨*linking*⟩ correspond to a constituent and linking morpheme, respectively, as defined in section 2.

6. Issues

There are several situations in which it is unclear whether a word should be annotated as a compound or not. In this section we will provide an overview of situations that can occur together with the decision whether or not the compounds should be included in the dataset.

6.1. Multiple analyses

Some compounds may have multiple valid analyses, for example, Du. *massagebed* can be analyzed as either *massa+gebed* **mass+prayer** 'mass prayer', or *massage+bed* **massage+bed** 'massage bed'. Similarly, Du. *minister* can be analyzed as either *mini+ster* **mini+star** 'small star' or left unanalyzed as *minister* 'minister', and Afr. *brandertjie* can be analyzed as *brand+ertjie* **burn+pea** 'burning pea' or *brander_tjie* **wave_DIM** 'small wave'. In the case where all alternatives are compounds (like *massagebed*), the different analyses will be given in the dataset. If an alternative reading is not a compound (as is the case with *minister* and *brandertjie*, only the compound reading is added to the dataset.

6.2. Lexicalized words

Certain compounds may be seen as highly lexicalized and as such are not annotated. For instance, Du. *wed+strijd* **bet+contest** 'match' is not annotated as wed + strijd, and Afr. *hand+skoen* **hand+shoe** 'glove' is not annotated either.

6.3. Words with prepositions

Compounds containing prepositions are not annotated. So, for instance, Du. *aan+val* **on+fall** 'attack', *aan+was* **on+grow** 'growth', and *af+stand* **off+stance** 'distance' are not annotated. Note that one should distinguish between prepositions as particles (which are also not analyzed) and prepositions as affixoids (which are analyzed): while Afr. *onder+neem* **under+take** 'undertake' is not analyzed, Afr. *onder+offisier* **under+officer** 'non-commissioned officer' is analyzed, since *onder* is considered an affixoid in this context.

In some cases, combinations of prepositions functioning as adjectives/adverbs are found. For instance, Du. *achter+uit+kijken* **back+out+look** 'look backwards' is annotated as achteruit + kijken and Du. *onder+door+gang* **under+through+way** 'underpass' is annotated as onderdoor + gang.

6.4. Words with names

There are several words that have a proper name as a component, for instance, company names like Du. *EMC_-+project* **EMC_-+project** 'EMC project' and *HP_-+journalist* **HP_-+journalist** 'journalist of the HP magazine'; compounds with names of artists, e.g. Du. *Zevon_-+plaat* **Zevon_-+record** 'record by Zevon'; and street names, e.g. Du. *Rozen+straat* **roses+street** 'Rose street'. Similar situations occur in compounds like Du. *Shamal+wind* **Shamal+wind** 'Shamal wind', where *Shamal* is the name of a type of wind, or Du. *Kiekeboek+prijs* **Kiekeboek+prize** 'Kiekeboek award'. Compounds with names as components (like all the above examples) should be annotated as compounds.

However, if the word under consideration is a proper name, it should not be analyzed. For instance, Afr. *Johannesburg* should not be analyzed as a compound (as it is a proper name), but Afr. *Johannesburg+goud* **Johannesburg+gold** 'gold from Johannesburg' is considered a compound and is analyzed as Johannesburg + goud.

6.5. Affixes as words

Some affixes also have homographs that exist as independent words (and those can be used as a component in compounds). For instance, the constituent Du. *ster* can either be a word meaning 'star', or a feminine suffix (e.g. Du. *ren_ster* **run_F** 'female runner'). Similarly, Du. *lijk* can be used as an independent word meaning 'corpse', or a suffix (e.g. Du. *burger_lijk* **civilian_ADJR** 'civil', which could therefore also be analyzed as *burger+lijk* **civilian+corpse** 'civilian corpse'). Similarly, Du. *mee* or *mede* 'with' in words like Du. *mee+gaan* **with+go** 'go along with'; Du. *toe* 'close' in words like Du. *toe+zicht* **close+view** 'supervision' are affixoids and hence analyzed as mee + gaan and toe + zicht.

In situations where the component that looks like an affix functions as a constituent in the compound, such as Du. *Kerst+ster* **Christmas+star** 'Christmas star'; Afr. *tiener+ster* **teenager+star** 'teen star', we do annotate these words as compounds: kerst + ster and tiener + ster respectively. With other affixes, this happens less often. Compare for instance Du. *ver+gezicht* **far+view** 'vista' is one of them (and is analyzed as ver + gezicht), but in contrast Du. *ver_zenden* **VR_send** 'send' is not annotated. In cases like these, in which an affix actually functions as an affix, the word is not analyzed as a compound. For instance, Du. *gevaarlijk* 'dangerous' is not analyzed as gevaar + lijk 'danger corpse', since this interpretation is highly improbable. Similarly, words with affixes like *wel, mis, tal, tuig* are not annotated, as in Du. *misdaad* 'crime', Du. *welvaart* 'prosperity', Du. *toegang* 'entrance', Du. *jaartal* 'date', Du. *vliegtuig* 'airplane', respectively. By default, the affix analysis is assumed (which means that the word is not analyzed as a compound) unless that reading is not available (in which case the word is analyzed as a compound), for instance in Du. *neutron_en+ster* **neutron_LINK+star** 'neutron star'.

A somewhat more complex situation occurs when analyzing Du. *minister* 'minister' or 'mini star'. In this case, *ster* does not function as an affix in either readings, so we follow the principle in section 6.1. This means that Du. *minister* is annotated as mini + ster.

6.6. Synthetic and derived compounds

In general, as mentioned earlier, synthetic compounds are not analyzed. For instance, Du. *tegen+draad_s_heid* **against+thread_LINK_NR** 'contrariness' cannot be split into *tegen* and *draadsheid*, as the second component is not a valid word. Similarly, Du. *on_vader+land_s+lieven_d* **un_father+country_LINK+love_PRTC** 'unpatriotic' is not annotated. However, Du. *terein+afbaken_ing* **terrain+delimitate_NR** 'terrain delimitation' is annotated as terein + afbakening, since both components are valid.

The same applies to derivations of existing compounds, such as Afr. *boek+rak_agtig* **book+case_ADJR** 'like a book case', which is not analysed.

Some synthetic or derived compounds seem to have valid components. However, if these components are not valid in this context, the words should not be annotated. For instance, words ending on Du. *kundig* 'knowledgeable', such as Du. *natuur+kundig* **nature+knowledgeable** 'physical', or Du. *ziekte+kundig* **sickness+knowledgeable** 'pathological' are not annotated. Similarly, words ending on *-ig* as in *aardig, jarig, vormig* in words like Du. *boos+aard_ig* **malice+nature_ADJR** 'malicious', Du. *negentig+jar_ig* **ninety+year_ADJR** 'ninety yearly', Du. *komma+vorm_ig* **comma+shape_ADJR** 'shaped like a comma' are not annotated either.

6.7. Names of days of the week

Names of days of the week are not annotated. Even though Du. *maan+dag* **moon+day** 'Monday' might be considered a compound by some, it is not annotated to keep consistency with respect to other days of the week, such as Du. *woensdag* 'Wednesday'. If the name of a day occurs as a component of a compound, the compound itself is still structured. For instance, Du. *woensdag+middag* **Wednesday+afternoon** 'Wednesday afternoon' is analysed as woensdag + middag.

6.8. Internet-related tokens

The data may contain Internet related tokens (for instance, the SoNaR corpus contains a collection of tweets). This not only includes hashtags and @mentions (found, for example, in Twitter data), but also URLs. All of these tokens are special in the sense that these types of tokens are modifications of regular tokens (or sequences or characters that are not part of the language). The modifications are, for instance, adding a hash symbol: # (such as *#vroegevogels*, or *#vrouwendag*) or an at symbol: @ in front of the token (for instance, *@loopmaatjes* or *@spatiegebruik*).

The situation with URLs is a little bit more complex. A URL may contain a valid token (which can also be a compound). However, at the least, the URL has a top-level domain, which is the rightmost part of a domain name. For instance in the domain name (which can be used as a URL): *voedingscentrum.nl*, the part *.nl* is the top-level domain. The other part, Du. *voeding_s+centrum* **feeding_LINK+centre** 'feeding center' is a regular token and can be analysed as a compound: voeding _ s + centrum.

URLs and also other Internet-derived tokens, such as hashtags and @mentions should not be annotated as compounds. The reason for this is that users that come up with the URLs, hashtags and @mentions may generate new tokens on the fly, while at the same time, spaces are not allowed within these tokens, which means that these types of tokens are more likely to look like a compound, even though the token would not be a compound in a non-Internet related setting.

6.9. Words with non-letter characters

Some words contain characters that are not letters. In particular, we describe the situations in which words contain hyphens, numbers, and brackets.

6.9.1. Words with hyphens

Words may in some cases contain hyphens. Firstly, some words require hyphens to combine components into a compound. Examples of this type of word are Du. *foto_-+enscenering* **photo_LINK+staging** 'photo-staging', Afr. *see_-+eend* **sea_LINK+duck** 'seaduck', and Du. *italiaans_-+somalische* **italian_LINK+somali** 'italian-somali'. Also included are phrasal compounds (Du. *wereld_-+vedetten_-+van_-+vroeger* **world_LINK+celebrities_LINK+of_LINK+past** 'world celebrities of the past'), and compounds with abbreviations or acronyms as modifiers (Afr. *ATKV_-+lid* **ATKV_LINK+member** 'member of the ATKV').

All these types of compounds should be annotated as compounds. Here we consider such hyphens as (orthographical) linking morphemes, since they also link two constituents to form a complex word. As such, we use an underscore before the hyphen in our annotations; e.g. see _ - + eend.

Note to distinguish between constituents in a compound, and affixes. In Dutch, the component *ex-* is a prefix, despite the fact that it could be used as an independent word (meaning 'ex-partner'), as in Du. *ex-society+journalist* **ex_LINK+society+journalist** 'former society journalist'; it is therefore not analysed as a compound. The same applies to Afr. *nie-* 'non-', as in *nie_-+frustrerende* **non_LINK+frustrating** 'non-frustrating', which is not analyzed as a compound, because *nie-* is a prefix.

6.9.2. Words with numbers

A subset of the set of compounds with hyphens contains compounds where one component (usually the modifier) is a number; such compounds are analysed. For instance, compounds such as Du. *12_-+punt_en+programma* **12_LINK+point_LINK +programme** '12-point programme' are annotated as 12 _ - + punt _ en + programma, or Afr. *.22_-+geweer* **.22_LINK+rifle** '.22 rifle' is annotated as .22 _ - + geweer. Note that a word like *12_-+voud_ig* **12_-+fold_ADJR** '12-fold' should not be analyzed as a compound (because *voudig* is not an independent word).

6.9.3. Words with numbers as words

Words with numbers written as words are only annotated if the number is above twenty. Hence, Afr. *sewe_-+en_-+twintig* **seven_LINK+and_LINK+twenty** 'twenty seven' is analysed as sewe _ - + en _ - + twintig.

6.9.4. Words with brackets

Words that contain other words separated by brackets are also considered compounds and therefore analysed. For instance, Du. *(omgevings)factor* **environment_LINK +factor** '(environmental) factor' should be annotated without the brackets: omgeving _ s + factor. Essentially the brackets in these cases indicate that the part between brackets is optional: *(omgevings)factor* can be read as either *omgevingsfactor* or *factor*.

6.10. Nonsense words and typos

When dealing with naturally occurring linguistic data, several types of tokens can be found that are not considered to be part of the "standard" language. For instance, nonsense words, such as Afr. *vagelgalm* or Du. *koninkrijkkon*, should not be annotated as compounds.

Words containing typos should not be annotated either. This refers to components that are non-real word errors (i.e. components that are not part of the language), such as Afr. *hadtekening* (instead of Afr. *hand+tekening* **hand+drawing** 'signature'). However, if the typo leads to a real word (a correct word in the language), this compound should be annotated. For instance, if Afr. *kanselier* 'chancellor' is spelled incorrectly with a double letter "l", it could be interpreted as Afr. *kansel+lier* **pulpit+lyre** 'lyre used on the pulpit'.

6.11. Words in other languages and archaic words

Loanwords are commonplace in all the languages of the world. In Afrikaans and Dutch, such loanwords can be used as constituents in a compound. However, loanwords are not annotated themselves. On the one hand, a word can consist of only loanword constituents, such as Du. *kriegwissenschaft, bachforelle, hatesphere*, or *weekend*; these words are not annotated. On the other hand, a word can consist of a combination of loan words and non-loanwords, such as Afr. *speaker+posisie* **speaker+position** 'position of the speaker (in parliament)', or Du. *weekend+retour* **weekend+retour** 'weekend return trip'; these cases should be analysed. Note that some compounds might look like loanwords, for instance, Du. *blues+rally* **blues+rally** 'blues rally'; however, both *blues* and *rally* are considered Dutch words, so this is analyzed as blues + rally.

In addition to clear loanwords as compounds or as constituents, compounds in dialects can also be found, such as Du. (dialect) *kervel+soepke* **chervil+soup** 'chervil soup'. Similar to compounds consisting only of loanword constituents (which are not analysed), compounds consisting only of dialect words are not annotated. However, compounds containing at least one non-dialect word are annotated.

Archaic words, such as Du. *desgevraagd* 'as requested', Afr. *desalnietemin* 'nevertheless', are not annotated.

Appendix B

Annotation Guidelines for the Semantic Analysis of Noun-Noun Compounds

These guidelines were taken and adapted from Ó Séaghdha's PhD thesis 'On Compound Semantics' (2008). They are developed to be able to describe the semantic relation between the constituents of two-noun compounds. We have only annotated those compounds that are not in the dictionary, but of which the constituent nouns can in fact be found in the dictionary. If a compound already has a gloss, we do not have to analyse it to find its meaning, but we do need to know the meaning of each constituent to be able to find the compound meaning. This means that a lot of common, lexicalised and exocentric compounds are excluded from the annotation. These compounds will be removed from the annotation data by crosschecking the data with a dictionary before the annotation commences. Should we still encounter such compounds in our data, rule 1.4 explains what to do with them.

More details on the adaptation of these guidelines can be found in chapter 4 of Verhoeven's master dissertation 'A Computational Semantic Analysis of Noun Compounds in Dutch' (2012).

A classification scheme with paraphrasing prepositions and predicates and a decision tree are made available to aid the annotators in making their acquaintance with the annotation process and guidelines. These are to be considered tools that can aid in the apprehension of the annotation process or when struggling with the classification of a certain compound. These tools can be found on the project website[1].

1. General Guidelines

The task is to annotate each compound noun N1 N2 with regard to the semantic relation that holds between the constituent nouns N1 and N2. It is assumed that compounds are either copulative or semantically right-headed.

Rule 1.1 The general annotation format is <RELATION-DIRECTION-RULE>.

[1] http://tinyurl.com/aucopro

RELATION is one of the 11 relation labels defined in section 2 of these guidelines. DIRECTION specifies the order of the constituent nouns in the chosen relation's argument structure – in particular, direction will have the value 1 if the first noun in the compound (N1) fits in the first noun slot mentioned in the rule licensing the chosen relation, and will have value 2 if the second noun in the compound (N2) fits in the rule's first noun slot. RULE is the number of the rule licensing the relation.
For example:

water fern
IN-2-2.1.3.1
This aquatic water fern is a rosette plant which has dense, fibrous roots.

enemy provocation
ACTOR-1-2.1.4.1
The army said at the weekend that troops had reacted to enemy provocations and intervened to protect local citizens.

In the case of water fern the IN relation is licensed by Rule 2.1.3.1 (N1/N2 is an object spatially located in or near N2/N1). Mapping the compound's constituent nouns onto the rule definition, we see that the first slot (N1/N2 is. . .) is filled by N2 *fern* and hence the direction is 2. For the categories BE, REL, LEX, UNKNOWN, MISTAG and NONCOMPOUND there is no salient sense of directionality, in these cases the direction will be annotated as 1.

cedar tree
BE-1-2.1.1.1
On rising ground at the western end of the churchyard of St Mary's at Morpeth in Northumberland stands, sheltered by cedar trees, a funerary monument.

In practice, we will assign every compound a direction to have uniformity in the encoding. Every compound from a category that has no sense of directionality (see above) will be encoded with direction 1. In the examples of section 2 you will find the direction of the example in brackets behind the compound.

Rule 1.2 Each compound is presented with its sentential context and should be interpreted within that context. Knowledge of other instances of the compound type are irrelevant.

A given compound type can have different meanings in different contexts. A school book is frequently a book read IN school, but it could also be a book ABOUT school. A wood table might be a table that IS wood (BE), but it might also be a table for chopping wood on (IN). The intended meaning of a compound is often clarified by the sentence it appears in.

Rule 1.3 Where a compound is ambiguous and is not clarified by the sentential context, the most typical meaning of the compound is favoured.

Compound interpretation must sometimes rely on world knowledge. In these cases, the annotator will have to rely on his or her intuition. Querying Google for the most typical meaning would be a viable option, but would take too much time in the annotation process.

The compound school book is not clarified by a sentence such as This is a school book'. In this case, book read IN school is the most typical interpretation. If the compound's ambiguity arises from the polysemy of a constituent, the same consideration applies. University can refer to an institution or its physical location, but in the case of university degree the institutional meaning must be correct as locations cannot award degrees, and the compound is labelled ACTOR. If the meaning of the compound is unclear, the appropriate tag is UNKNOWN.

Rule 1.4 There are number of special cases that would normally not appear in our training data. If they should be present, they are to be treated differently than other compounds, they will all be annotated REL or LEX.

- When a compound is used metaphorically, it will not be considered a regular compound and it should be labelled LEX.

For example: the compound bird brain is often used to refer to someone stupid, not to an actual bird's brain. Luckily, a lot of metaphorical compounds have such a typical meaning that they can be found in a dictionary and will therefore not be present in the annotation data.

- Where a compound consisting of two common nouns is used as a proper noun, it will be discarded from our annotation. Also compounds that exist of one or more proper nouns, abbreviations or acronyms will be left out. All these special cases receive the REL tag.

Many names, while constructed from two common nouns, do not seem to encode the same kind of semantics as non-name compounds, e.g. Penguin Books, Sky Television, Dolphin Close, Coronation Street. These names encode only a sense of non-specific association between the constituents. All compounds that are used as a proper noun will therefore be classified as REL, even those that could be classified otherwise. For example: the Telecommunications Act, The Old Tea Shop, Castle Hill. The task of identifying these proper noun compounds should be passed on to a named entity recognition (NER) module.

Rule 1.5 Where there is a characteristic situation or event that characterizes the semantic relation between the constituents, it is necessary to identify which constituents of the compound are participants and which roles they play. Whether such a situation exists for a given compound, and the roles played by its constituents in the situation, will determine which relation labels are available.

Participants take on roles that can be described as Agent, Instrument, Object or Result:

- **Agent** The instigator of the event, the primary source of energy
- **Instrument** An intermediate entity that is used/acted on by the Agent and in

turn exerts force on or changes the Object; more generally, an item which is used to facilitate the event but which is not the Object

- **Object** The entity on which a force is applied or which is changed by the event and which does not exert force on any participant other than the Result. Recipients (e.g. of money or gifts, but not outcomes) also count as Objects.
- **Result** An entity which was not present before and comes into being through the event

For example, the meaning of cheese knife seems to involve an event of cutting, in which cheese and knife take object and instrument roles respectively. Similarly, taxi driver evokes an event of driving and gevangenisbewaker (prison guard) evokes an event of guarding. The INST and ACTOR relations apply only where such a situation or event is present and where the compound identifies its participant(s). The application of HAVE assumes that the most salient aspect of the underlying situation is possession. It is not strictly necessary to identify the precise nature of the situation or event, only to identify the general roles played by the participants.

Some role-tagged examples: cheese$_O$ knife$_I$, taxi$_O$ driver$_A$, sneezing$_R$ powder$_I$, gevangenis$_O$bewaker$_A$. It follows from the role descriptions that locations and topics do not count as participants - compounds encoding such roles receive IN and ABOUT labels instead of the ACTOR and INST labels reserved for participants. The participant role types are listed in order of descending agentivity. We thus have an agentivity hierarchy Agent > Instrument > Object > Result[2]. This ordering plays an important role in distinguishing ACTOR compounds from INST compounds (see Sections 2.1.4 and 2.1.5). It is not necessary to annotate this information, and it is not always necessary to identify the exact participant role of a constituent, so long as the hierarchical order of the constituents can be identified. Identifying participants is only needed to distinguish between relations (ACTOR vs. INST) and directionalities (see the discussion under Rule 2.1.5.2).

2. Semantic Relations

2.1 Main Relations

2.1.1 BE

Rule 2.1.1.1 X is N1 and X is N2.

Eng. *woman driver, elm tree, distillation process, human being*
Afr.: *digter-skrywer, briefbom, kommapunt, gasarbeider, vroueatleet, seunsvriend, wuifgroet*
Du.: *geluidhinder, rundsvlees, bombrief, puntkomma, gastarbeider, getuige-deskundige*

This rule does not admit sequences such as Eng. *deputy chairman, fellow man, chief executive* or Du. *hoofdverantwoordelijke*, where it is not correct to state that an [N1

[2] This agentivity hierarchy was informed by the semantic roles hierarchy in Talmy (2000).

N2] is an N1 (a chief executive is not a chief). Such sequences are not to be considered compounds, and their modifiers are to be considered (mistagged) adjectives – see Rule 2.2.1.1.

Rule 2.1.1.2 N2 is a form/shape taken by the substance N1.

Eng. *stone obelisk, chalk circle, plastic box, steel knife*
Afr. *plastiekrekkie, glasbak, houtstoel, ysterhek, silverring*
Du. *gummiband, betonsteen, staalkabel*

This rule is not very productive in Dutch since substances are most often written as adjectives, e.g. *plastieken doos, stalen mes*.

Rule 2.1.1.3 N2 is ascribed significant properties of N1 without the ascription of identity. The compound roughly denotes "an N2 like N1".

Eng. *father figure, angler fish, chain reaction, pie chart*
Afr. *mensaap, vaderfiguur, kettingreaksie*
Du. *hagelpatroon, rondegang, manwijf, mensaap*

2.1.2 HAVE

Rule 2.1.2.1 N1/N2 owns N2/N1 or has exclusive rights or the exclusive ability to access or to use N2/N1 or has a one-to-one possessive association with N2/N1

Eng. *army base(1), customer account(1), government power(1)*
Afr. *skoolgrond(1), kantooradres(1), menseregte(1)*
Du. *straatnaam(1), koningsdochter(1)*

The term one-to-one possessive association is intended to cover cases where it seems strange to speak of ownership, for example in the case of inanimate objects (street name, planet atmosphere).

Rule 2.1.2.2 N1/N2 is a physical condition, a mental state or a mentally salient entity experienced by N2/N1

Eng. *polio sufferer(1), cat instinct(2), student problem(2), union concern(2)*
Afr. *kankerpasiënt(1)*
Du. *lepralijder(1), studentenprobleem(2)*

Rule 2.1.2.3 N1/N2 has the property denoted by N2/N

Eng. *water volume(1), human kindness(1)*
Afr. *stooftemperatuur(1)*
Du. *productietijd(1)*

A "property" is something that is not an entity or a substance but which an entity/-substance can be described as having. Redness, temperature, dignity, legibility are all examples of properties.

Rule 2.1.2.4 N1/N2 has N2/N1 as a part or constituent

Eng. *car door(1), motor boat(2), cat fur(1), chicken curry(2), pie ingredient(1), tree sap(1)*
Afr. *deurknop(1), kasdeur(1), stortkop(1), geweerloop(1), sjokoladekoek(2), melktert(2)*
Du. *houtweefsel(1), bladzijde(1), moutjenever(2), hamersteel(1), grafzerk(1), tafelblad(1)*

The test for the presence of a part-whole relation is whether it seems natural and accurate in the context to say "The N1/N2 has/have N2/N1" and "The N1/N2 is/are part of N2/N1". Furthermore, substances which play a functional role in a biological organism are classed as parts: *human blood, tree sap, whale blubber*. This is the case even when the substance has been extracted, as in *olive oil*.

A part is often located in its whole, but in these cases the part-whole relation is to be considered as prior to the co-location, and HAVE is preferred to IN. Complications arise with cases such as *sea chemical*, where both HAVE and IN seem acceptable. One principle that can be used tests whether the candidate part is readily separated (perceptually or physically) from the candidate whole. Chemicals in sea water (HAVE) are not typically separable in this way and can be viewed as parts of a whole. On the other hand, a *sea stone* or a *sea (oil) slick* are perceptually distinct and physically separable from the sea and are therefore IN.

Rule 2.1.2.5 N1/N2 is a group/society/set/collection of entities N2/N1

Eng. *stamp collection(2), character set(2), lecture series(2), series lecture(1), committee member(1), infantry soldier(1)*
Afr. *seëlversameling(2), keramiekversameling(2), studentegroep(2)*
Du. *postzegelverzameling(2), schoenenhoop(2), groepslid(1)*

2.1.3 IN

In the following rules, an opposition is drawn between events/activities and objects. The class of events includes temporal entities such as times and durations. Objects are perceived as non-temporal and may be participants in an event (the term participant is used as defined under Rule 1.5). To assign the correct rule, the annotator must decide whether the located thing is an event or an object, and whether the location is temporal or spatial. Events may also sometimes be participants – in the sense of Rule 1.5 and in these cases the rules dealing with objects and participants will apply – a nursing college is a college where nursing is taught as a subject, but not necessarily one where the activity of nursing takes place, so Rule 2.1.3.1 applies. In contrast a nursing home, being a home where the event of nursing takes place, would come under Rule 2.1.3.2, analogous to dining room. Some nouns are polysemous and can refer to both objects (play as a written work, harvest as harvested crops) and events (play as performance, harvest as activity). The annotator must decide whether the temporal or physical aspect is primary in a given context.

Rule 2.1.3.1 N1/N2 is an object spatially located in or near N2/N1

Eng. *forest hut(2), shoe box(1), side street(2), top player(2), crossword page(1), hospital doctor(2), sweet shop(1)*
Afr. *vleismark(1), hospitaalbed(2), begrafenisrys(2)*
Du. *waterplant(2), rivierleem(2), ziekenhuisbed(2), havenkantoor(2), kerkdief(2)*

Where the location is due to part-whole constituency or possession, HAVE is preferred (as in *car door, sea salt*). Source-denoting compounds such as country boy and spring water are classed as IN as the underlying relation is one of location at a (past) point in time.

Rule 2.1.3.2 N1/N2 is an event or activity spatially located in N2/N1

Eng. *dining room(1), hospital visit(2), sea farming(2), football stadium(1)*
Afr. *plaasbesoek(2), wildtuintoer(2), harsingontsteking(2)*
Du. *biljartzaal(1), distributiecentrum(1), tuinfeest(2), zeeslag(2)*

Rule 2.1.3.3 N1/N2 is an object temporally located in or near N2/N1, or is a participant in an event/activity located there

Eng. *night watchman(2), coffee morning(1)*
Du. *nachtuil(2), sterrennacht(1), lenteweertje(2), weekblad(2)*
Afr. *dagblad(2), nagapie(2), maanskynaand(1)*

Rule 2.1.3.4 N1/N2 is an event/activity temporally located in or near N2/N1

Eng. *future event(2), midnight mass(2)*
Afr. *rugbyseizoen(1), somerdiens(2)*
Du. *avondfeest(2), nachtvoorstelling(2), jaarvergadering(2)*

2.1.4 ACTOR

The distinction between ACTOR and INST is based on sentience. Only certain classes of entities may be actors:

1. Sentient animate lifeforms: membership of the animal kingdom (regnum animalia) is a sufficient condition. Bacteria and viruses are not sentient enough (flu virus is annotated INST).
2. Organisations or groups of people: for example finance committee, consultancy firm, manufacturing company, council employee. Some words referring to institutions are polysemous in that they can denote its physical aspect or its social/organisational aspect – university often denotes a physical location, but in the compounds university degree and university decision it is functioning as an organisation and count as agents (granting a degree and making a decision are actions only humans or organisations can carry out). On the other hand, in research university it is not clear whether we have a university that does research (agentive)

> or a university in which research is done (non-agentive). In such cases, the physical denotation should be considered the primary meaning of the word, and the organisational denotation is derived through metonymy – the non-agentive interpretation of these compounds is favoured unless the underlying event requires the institution to act as an agent. Such events often involve the institution acting as a legal entity. Hence university degree (degree awarded by a university), school decision (decision made by a school), shop employee (employee employed by a shop) are ACTOR; research university, community school, school homework and sweet shop are IN.

A compound can be labelled ACTOR only if the underlying semantic relation involves a characteristic situation or event. In the following definitions, the term participant is used in the sense of Rule 1.5.

Rule 2.1.4.1 N1/N2 is a sentient participant in the event N2/N1

Eng. *student demonstration(1), government interference(1), infantry assault(1)*
Afr. *werkerstaking(1), vrouekonferensie(1)*
Du. *burgeroorlog(1), arbeidsvrouw(2), aanslagpleger(2)*

That N2/N1 denote an event is not sufficient for this rule – it must be the characteristic event associated with the compound. Hence this rule would not apply to a singing teacher, as the characteristic event is teaching, not singing. Instead, Rule 2.1.4.2 would apply. As only one participant is mentioned in the current rule 2.1.4.1, there is no need to establish its degree of agentivity.

Rule 2.1.4.2 N1/N2 is a sentient participant in an event in which N2/N1 is also a participant, and N1/N2 is more agentive than N2/N1

Eng. *honey bee(2), bee honey(1), company president(2), history professor(2), taxi driver(2), student nominee(1)*
Afr. *spankaptein(2), voortrekkerleier(2)*
Du. *aasdier(2), hartendief(2)*

Relative agentivity is determined by the hierarchy given under Rule 1.5. The underlying event cannot be one of possession (car owner = HAVE) or location (city inhabitant = IN). Profession-denoting compounds often have a modifier which is a location – street cleaner, school principal, restaurant waitress, school teacher. A distinction can be drawn between those where the profession involves managing or changing the state of the location, i.e. the location is an object (school principal, street cleaner = ACTOR), and those where the profession simply involves work located there (school teacher, restaurant waitress = IN by Rule 2.1.3.1). Note that modifiers in -ist such as expressionist, modernist, socialist, atheist are treated as nouns, so that an expressionist poem is analysed as a poem such as an expressionist would characteristically write.

2.1.5 INST

The name INST(rument) is used to distinguish this category from ACTOR, though the scope of the category is far broader than traditional definitions of instrumentality. Again, the term participant is used in the sense of Rule 1.5.

Rule 2.1.5.1 N1/N2 is a participant in an activity or event N2/N1, and N1/N2 is not an ACTOR.

Eng. *skimming stone(2), gun attack(1), gas explosion(1), combustion engine(2), drug trafficking(1), rugby tactics(2), machine translation(1)*
Afr. *beesveiling(1), bomdril(1), ontstekingsknoppie(2)*
Du. *smaakbederf(1), zaadhandel(1), leengoed(2)*

Compounds identifying the location of an event (such as street demonstration) should be labelled IN by Rule 2.1.3.2 or 2.1.3.4, and compounds identifying the focus of or general motivation for a human activity or mental process (such as crime investigation), but not its direct cause, should be labelled ABOUT by Rule 2.1.6.3. As only one participant is mentioned, there is no need to establish its degree of agentivity.

Rule 2.1.5.2 The compound is associated with a characteristic event in which N1/N2 and N2/N1 are participants, N1/N2 is more agentive than N2/N1, and N1/N2 is not an ACTOR.

Eng. *rice cooker(2), tear gas(2), blaze victim(1)*
Afr. *traangas(2), voedselverwerker(2)*
Du. *cadeaubon(2), worstmachine(2)*

The directionality of the relation is determined by the more agentive participant in the hierarchy given in Rule 1.5: cheeseO knifeI (INST2), wineO vinegarR (INST1), windA damageR (INST1), humanO virusA (INST1). Sometimes it may be difficult to distinguish Agents from Instruments (gun wound) or Objects from Results (blaze victim) – this is not important so long as it is possible to identify which participant is more agentive.

In some cases, it may not be clear what the exact underlying event is, but the more agentive participant may still be identified – a transport system is a system that in some way provides or manages transport, but it is nonetheless clear that the appropriate label is INST2. In other cases, where both participants affect each other, it may be less clear which is more agentive – motor oil can be construed as oil that lubricates/enables the function of the engine or as oil the engine uses. Likewise petrol motor, computer software, electron microscope. At least where the relation is between a system or machine and some entity it uses to perform its function, the former should be chosen as more agentive. Hence motor oil is INST1, petrol motor is INST2, and so on.

As in Rule 2.1.5.1, where one of the constituents is the location of the associated event, then IN is the appropriate label by Rule 2.1.3.1 or 2.1.3.3. If the more agentive participant meets the criteria for ACTOR status (2.1.4), then that label should be applied instead. If the interaction between the constituents is due to one being a part of the other (as in car engine), HAVE is the appropriate label by Rule 2.1.2.4. A border with ABOUT must be drawn in the case of psychological states and human activities whose cause or focus is N1. As described further under Rules 2.1.6.3, the criterion adopted is based on whether there is a direct causal link between N1 and N2 in the underlying event – a bomb can by itself cause bomb terror (INST1), but a spider phobia is not a reaction to any particular spider and is classed as ABOUT.

2.1.6 ABOUT

Rule 2.1.6.1 N1/N2's descriptive, significative or propositional content relates to N2/N1

Eng. *fairy tale(2), flower picture(2), tax law(2), exclamation mark(2), film character(2), life principles(2), sitcom family(1)*
Afr. *mensekennis(2), balletmusiek(2)*
Du. *vakjargon(2), contactstoornis(2), praktijktheorie(2), vakdeskundigheid(2)*

In English, a lot of speech acts belong to this category. Direction 2 is a lot more prominent with this rule. Properties and attributes that seem to have a descriptive or subjective nature are still to be labelled HAVE by Rule 2.1.2.3 – street name and music loudness are HAVE1.

Rule 2.1.6.2 N1/N2 is a collection of items whose descriptive, significative or propositional content relates to N2/N1 or an event that describes or conveys information about N2/N1

Eng. *history exhibition(2), war archive(2), science lesson(2)*
Afr. *kunsuitstalling(2), musiekversameling(2)*
Du. *tijdreeks(2), muziekbibliotheek(2)*

Rule 2.1.6.3 N1/N2 is a mental process or mental activity focused on N2/N1, or an activity resulting from such

Eng. *crime investigation(2), science research(2), research topic(1), exercise obsession(2), election campaign(2), football violence(2), holiday plan(2)*
Afr. *taalnavorsing(2), selfondersoek(2)*
Du. *darmonderzoek(2), plantenobsessie(2)*

In the case of activities, N1/N2 cannot belong to any of the participant categories given under Rule 1.5; rather it is the topic of or motivation for N2/N1. The sense of causation in, for example, oil dispute is not direct enough to admit an INST classification – the state of the oil supply will not lead to an oil dispute without the involved parties taking salient enabling action. In the case of emotions, there is also a risk of overlapping with INST; bomb terror is INST and bomb dislike is classed as ABOUT, but examples such as bomb fear are less clearcut. A line can be drawn whereby immediate emotional reactions to a stimulus are annotated INST, but more permanent dispositions are ABOUT. In the case of bomb fear, the relation must be identified from context. Problems (debt problem) and crises (oil crisis) also belong to this category, as they are created by mental processes.

Rule 2.1.6.4 N1/N2 is an amount of money or some other commodity given in exchange for N2/N1 or to satisfy a debt arising from N2/N1

Eng. *share price(2), printing charge(2), income tax(2)*
Afr. *goudprys(2), boedelbelasting(2), petrolprys(2), bankkoste(2)*
Du. *olieprijs(2), loonarbeid(1), gokbedrag(2)*

N2/N1 is not the giver or recipient of N1/N2 – an *agency fee* would be INST under the interpretation feeI paid to an agencyO – but the thing exchanged or the reason for the transaction.

2.1.7 REL

Rule 2.1.7.1 The relation between N1 and N2 is not described by any of the above relations but seems to be produced by a productive pattern

Eng. *Baker Street, sodium chloride*
Afr. *Akkerlaan, waterstofkarbonaat*
Du. *Vaarttheater, Plataanlei, waterstofcarbonaat, adjudant-onderofficier*

A compound can be associated with a productive pattern if it displays substitutability. If both of the constituents can be replaced by an open or large set of other words to produce a compound encoding the same semantic relation, then a REL annotation is admissible. For example, the compound reading skill (in the sense of degree of skill at reading) is not covered by any of the foregoing categories, but the semantic relation of the compound (something like ABILITY) is the same as that in football skill, reading ability and learning capacity. This contrasts with an idiosyncratic lexicalised compound such as home secretary (= LEX), where the only opportunities for substitution come from a restricted class and most substitutions with similar words will not yield the same semantic relation. Another class of compounds that should be labelled REL are names of chemical compounds such as carbon dioxide and sodium carbonate, as they are formed according to productive patterns. There are also several special cases that receive the REL tag. Take a look at Rule 1.4 for the descriptions.

2.1.8 LEX

Rule 2.1.8.1 The meaning of the compound is not described by any of the above relations and it does not seem to be produced by a productive pattern

Eng. *turf accountant, monkey business*
Afr. *spierpaleis*
Du. *loftrompet, prins-gemaal, spierbundel*

These are noncompositional in the sense that their meanings must be learned on a case-by-case basis and cannot be identified through knowledge of other compounds. This is because they do not have the property of substitutability - the hypothetical compounds horse business or monkey activity are unlikely to have a similar meaning to monkey business. LEX also applies where a single constituent has been idiosyncratically lexicalised as a modifier or head such as X secretary meaning 'minister responsible for X'.

2.1.9 UNKNOWN

Rule 2.1.9.1 The meaning of the compound is too unclear to classify.

Some compounds are simply uninterpretable, even in context. This label should be avoided as much as possible but is sometimes unavoidable.

2.2 Noncompounds

2.2.1 MISTAG

Rule 2.2.1.1 One or both of N1 and N2 have been mistagged and should not be counted as (a) common noun(s)

Eng. *fruity bouquet* (N1 is an adjective), *London town* (N1 is a proper noun)
Afr. *geelwortel* (N1 is an adjective), *hoofkok* (N1 is adjective-like), *afhaal* (N is a verb), *Paryspad* (N1 is a proper noun)
Du. *Juratijdperk* (N1 is a proper noun), *voortuin* (N1 is a preposition), *hoofdbewaker* (N1 is adjective-like)

In the case of blazing fire, N1 is a verb, so this is also a case of mistagging; in superficially similar cases such as dancing teacher or swimming pool, however, the -ing form can and should be treated as a noun. The annotator must decide which analysis is correct in each case – a dancing teacher might be a teacher who is dancing (MISTAG) in one context, but a teacher who teaches dancing (ACTOR) in another context. Certain modifiers might be argued to be nouns but for the purposes of annotation are stipulated to be adjectives. Where one of assistant, key, favourite, deputy, head, chief or fellow appears as the modifier of a compound in the data, it is to be considered mistagged. This only applies when these modifiers are used in adjective-like senses – key chain or head louse are clearly valid compounds and should be annotated as such.

2.2.2 NONCOMPOUND

Rule 2.2.2.1 The extracted sequence, while correctly tagged, is not a 2-noun compound

There are various reasons why two adjacent nouns may not constitute a compound:

1. An adjacent word should have been tagged as a noun, but was not.
2. The modifier is itself modified by an adjacent word, corresponding to a bracketing [[X N1] N2]. For example: [[real tennis] club], [[Liberal Democrat] candidate], [[five dollar] bill]. However compounds with conjoined modifiers such as land and sea warfare and fruit and vegetable seller can be treated as valid compounds so long as the conjunction is elliptical (land and sea warfare has the same meaning

as land warfare and sea warfare). Not all conjoined modifiers satisfy this condition – a salt and pepper beard does not mean a beard which is a salt beard and a pepper beard, and the sequence pepper beard is a NONCOMPOUND.

3. The two words are adjacent for other reasons. For example: ‘the question politicians need to answer’, structureless lists of words.

4. The modifier is not found as a noun on its own, because it would not appear in the dictionary. For example: multiparty election, smalltown atmosphere.

Appendix C

Annotation Guidelines for the Semantic Analysis of Other Nominal Compounds

This document contains the annotation guidelines for the semantic analysis of XN compounds (nominal compounds with a non-noun as a first constituent) for Dutch and Afrikaans. These guidelines were first described in Verhoeven & van Huyssteen (2013). The non-noun can be a verb, preposition, adjective, adverb or quantifier. This protocol thus only serves for two-constituent compounds, although we will sometimes mention multi-word compounds in the classification scheme.

1. General Guidelines

In the introduction that follows, we describe some relevant remarks for annotation:

a. Endo- vs. exocentric

A compound is considered either semantically or syntactically exocentric when it does not fulfil the same semantic (does not refer to same type of entity) or syntactic function (e.g. POS category) as one of its parts. Exocentric compounds are always lexicalised, while endocentric compounds are mostly non-lexicalised (but might of course also be lexicalised).

b. Nominalized verbs

See section 2 of Verhoeven & van Huyssteen (2013).

c. Ambiguity

Where a compound is ambiguous and its sentential context does not clarify its meaning, the most typical meaning of the compound is favoured.

d. Structure of this document

This protocol describes different classes. Most of them contain paraphrases (between brackets) that can help the annotator with his task. However, these paraphrases should not be considered as an undeniable truth. It is possible that the compound fits the class, but not the paraphrase.

e. Structure of the annotation

Every considered compound should receive a tag between 1.1.1. and 5.3., this means only the most specific (and thus lowest) subcategories in the enumeration below can be used as classes during annotation. When objections to the classification of this compound with this class can be raised, these should be noted in the 'comment' column.

2. Semantic Relations

2.1. Verb-Noun Compounds

2.1.1. Event

The verb describes an action in which the noun is some sort of participant.

2.1.1.1. SUBJECT

('N that Vs; the goal of N is to V')
Afr. *snydokter* **cut+doctor** 'doctor that cuts; surgeon', *beskermheer, kookmeester, werksesel, ploegos*
Du. *gloeilamp* **glow+lamp** 'lamp that glows; lightbulb', *druppelkraan, sleepboot, schuifdeur*[1]

2.1.1.2. OBJECT

('N that is (being) V-ed; VN is the result of V- INF; the goal of N is to be V-ed')
Afr. *snyblomme* **cut+flowers** 'the goal of the flowers is to be cut', *fakspapier, suiglemoen, drinkwyn, stoofappel, eetdruiwe, suiglekker, suigstokkie, maalvleis, drukstuk*
Du. *werpbal* **throw+ball** 'ball that is thrown', *snuffelpaal, hakhout*

2.1.1.3. INSTRUMENT

('N is used to V-INF')
Afr. *kapbyl* **chop+axe** 'axe used to chop down trees', *klimraam, wiegstoel, kapbyl, faksmasjien, opdienblad, waslap, snydiamant*
Du. *leesbril* **read+glasses** 'glasses that are used to read; reading glasses', *bewaarkluis, zoeklicht, bouwplan, eetlepel, leerboek*[1]*, verkooptechniek*[1]

2.1.2. Location

The noun describes a spatial or temporal location of the action described by the verb.

2.1.2.1. SPACE

('V in (neighbourhood of) N; N where one Vs')
Afr. *herstelsentrum* **recover+centre** 'centre where people recover from injuries or operations', *kapsalon, eetsaal, kuierplek, drinkgat, wasbak, snyhoek* 'hoek waar twee lyne mekaar sny'

[1] (Don, 2009: 374)

Du. *slaapkamer* **sleep+room** 'room where one sleeps; bed room', *schrijftafel, leeszaal, stemdistrict, opvangcentrum, speelveld*[1]*, rookpaal, praatpaal, reisbestemming*

2.1.2.2. TIME

('N during which one Vs')
Afr. *bakleifase* **quarrel+fase** 'fase during which one quarrels', *ruspouse, werkvakansie*
Du. *regeerperiode* **rule+period** 'period during which someone rules', *kruipstadium*

2.1.3. Composed of

The noun is some sort of collection of the action described by the verb.
('N consists of V')
Afr. *skokterapie* **shock+therapy** 'therapy that consists of shocking the patient', *vegsport*
Du. *niesbui* **sneeze+shower** 'rapid succession of sneezes'

2.1.4. Lexicalised

2.1.4.1. ENDOCENTRIC

Afr. *snyhou* **cut+stroke** 'kind of tennis stroke', *snykoekie* 'plat koek in olie gebraai'
Du. *draaibal* **turn+ball** 'ball that is kicked with a turning effect'

2.1.4.2. EXOCENTRIC

Afr. *speeltuin* **play+garden** 'playground'
Du. *verzamelwoede* **collect+anger** 'urge or mania to collect things'

2.2. Adjective-Noun Compounds

The adjective-noun compounds are (probably) all lexicalised in Dutch and Afrikaans, since the normal pattern in Germanic languages is to consider A + N as a syntactic phrase.

2.2.1. Lexicalised

2.2.1.1. ENDOCENTRIC

2.2.1.1.1. Duration

('kind of N that is A')
Afr. *langverlof* **long+leave** 'kind of leave that is longer than what is normally taken', *kortasem, kortpad*
Du. no examples found so far

2.2.1.1.2. Colour

('kind of N that is A')
Afr. *geelrys* **yellow+rice** 'kind of rice that is yellow', *groenslaai, bruinbrood, witwyn*
Du. *rodekool* **red+cabbage** 'kind of cabbage that is red'

2.2.1.1.3. Other quality/property

('kind of N that has the quality expressed by A')

Afr. *sterkstroom* **strong+current** 'high volt- age; the power current is strong'
Du. *hogeschool* **high+school** 'school for higher education'

2.2.1.2. EXOCENTRIC

2.2.1.2.1. Attributive

Afr. *luigat* **lazy+bottom** 'person that is lazy'
Du. *kaalkop* **bald+head** 'person that has a bald head', *roodhuid, wijsneus, bleekgezicht, kaalkop, groenboek, blauwdruk*

2.2.1.2.2. Other

Afr. *groenskrif* **green+script** 'first draft of legislation; green paper', *blyspel* Du. *blijspel* **happy+game** 'theatre play that is supposed to amuse people'

2.3. Quantifier-Noun Compounds

2.3.1. Quantity-Object

To our knowledge, the only productive form of QN compounding is when the quantifier specifies the quantity of N within a larger phrasal compound (i.e. [[Q+N]$_{NP}$ N]$_{N}$), e.g. Afr. *sewejaardroogte* **seven+year+drought** 'seven-year drought'.

These kinds of compounds will not appear in our data. Should they appear, they should be annotated as MISTAG or NONCOMPOUND.

2.3.2. Lexicalised

2.3.2.1. ENDOCENTRIC

No examples so far.

2.3.2.2. EXOCENTRIC - ATTRIBUTIVE

(compound is 'entity that has Q number of N')
Afr. *vierkleur* **four+colour** 'flag of the old Transvaal Republic', *agthoek, driepoot*
Du. *duizendpoot* **thousand+leg** 'centipede', *drietand, eenoog, driekleur*

2.4. Preposition-Noun Compounds

2.4.1. Location

The concept described by N is at position P of an undefined other concept. The paraphrases of the categories described here use the reference point 'G' (i.e. grounding point) to refer to these undefined concepts.

2.4.1.1. SPACE

('N is spatially at position P relative to G')
Afr. *onderrok* **under+skirt** 'skirt worn under other skirt"
Du. *achterlicht* **behind+light** 'light at behind of car or bike; rear light', *voordeur, bovenkamer*

2.4.1.2. TIME

('N is temporally at position P relative to G')

Afr. *voormiddag* **before+noon** 'forenoon'
Du. *nagesprek* **after+talk** 'conversation after previous event'

2.4.1.3. ABSTRACT/METAPHORICAL

('N is at abstract position P relative to G')
Afr. *byverdienste* **by+income** 'additional income to normal income', *oormaat, byvoordeel, byvoegsel, bysaak*
Du. *overgewicht* **over+weight** 'the weight that is over the normal', *bijbaantje, onderofficier, bijzaak*

2.4.2. Process-based

This kind PN compound is related to some kind of process. The noun goes in the (possibly abstract) direction described by the preposition.
('N goes in direction P')
Afr. *opmars* **up+march** 'march', *uitvaart*
Du. *overstap* **over+step** 'transfer on public transport', *uittocht*

2.4.3. Lexicalised

2.4.3.1. ENDOCENTRIC

Afr. *optog* **up+trip** 'procession'
Du. *uitgroeisel* **out+growth** 'excrescence', *optocht*

2.4.3.2. EXOCENTRIC

Afr. *insig* **in+sight** 'insight' Du. *nageboorte* **after+birth** 'afterbirth', *opmaat, opdracht, najaar, voorjaar, vooravond*

2.5. Unclassifiable

2.5.1. Mistag

At least one of the two constituents has been tagged with the wrong part-of-speech.

2.5.2. Noncompound

Although the parts-of-speech are tagged correctly, this is not a correctly formed two-part compound.

2.5.3. Unknown

The meaning of the compound is too unclear to classify.