The Development of Dutch and Afrikaans Language Resources for Compound Boundary Analysis

van Zaanen, M.M.; van Huyssteen, Gerhard; Aussems, Suzanne; Emmery, Chris; Eiselen, Roald

Published in:
Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014); Reykjavik, Iceland

Publication date:
2014

Document Version
Peer reviewed version

The Development of Dutch and Afrikaans Language Resources for Compound Boundary Analysis

Menno van Zaanen[a], Gerhard van Huyssteen[b], Suzanne Aussems[a], Chris Emmery[a], Roald Eiselen[b]

Tilburg University[a],
Tilburg center for Cognition and Communication, P.O. Box 90153, 5000 LE Tilburg, the Netherlands,
North-West University[b]
Centre for Text Technology, Internal Box 395, Private Bag X6001, Potchefstroom 2520, South Africa
{mvzaanen, s.h.j.a.aussems, c.d.emmery}@tilburguniversity.edu, {gerhard.vanhuyssteen, roald.eiselen}@nwu.ac.za

Abstract

In most languages, new words can be created through the process of compounding, which combines two or more words into a new lexical unit. Whereas in languages such as English the components that make up a compound are separated by a space, in languages such as Finnish, German, Afrikaans and Dutch these components are concatenated into one word. Compounding is very productive and leads to practical problems in developing machine translators and spelling checkers, as newly formed compounds cannot be found in existing lexicons. The Automatic Compound Processing (AuCoPro) project deals with the analysis of compounds in two closely-related languages, Afrikaans and Dutch. In this paper, we present the development and evaluation of two datasets, one for each language, that contain compound words with annotated compound boundaries. Such datasets can be used to train classifiers to identify the compound components in novel compounds. We describe the process of annotation and provide an overview of the annotation guidelines as well as global properties of the datasets. The inter-annotator agreement between the annotators was considered highly reliable. Furthermore, we show the usability of these datasets by building an initial automatic compound boundary detection system, which assigns compound boundaries with approximately 90% accuracy.

Keywords: compound boundary annotation, language resource development, Dutch, Afrikaans

1. Introduction

Compounding, the process of combining two or more stems or words into a complex lexical unit, is considered a very productive word formation process in a large variety of languages (Aussems et al., 2013a; Aussems et al., 2013b). In languages such as English, compounds are created by combining components but keeping them separated by a space, such as *trapeze artist*. In other languages, such as Finnish, German, Afrikaans, or Dutch, the components of a compound are concatenated into one word, such as Finnish *trapetsitaiteilija*, German *Trapezkünstler*, Afrikaans *sweefstokarties*, or Dutch *trapezeartiest*. The concatenation of two or more words into one word is a very productive process, which allows for the construction of new compounds on the fly. Due to this frequently used process to create new words, such idiosyncratic compound words often cannot be found in a dictionary. As a result, the productivity of compounding leads to problems in tools with predefined lexicons, such as spelling checkers, or automatic translators (van Huyssteen and van Zaanen, 2004; Koehn and Knight, 2003).

In order to allow us to build tools that identify boundaries between the components of compounds, annotated training data is required. Although several morphologically annotated datasets exist, most of these datasets incorporate additional morphological annotations next to the compound boundaries. As such, these datasets are not perfectly suited to develop compound boundary detection systems.

Additionally, the lack of datasets specifically targeting compound boundary information makes research on the process of compounding difficult to achieve. As compounding is productive and used in a variety of languages, it is interesting from a linguistic point of view to investigate, for instance, compounding as a language independent process.

The research described here forms the basis for research that addresses the cross-language comparison of the process of concatenative compounding between the closely-related West Germanic languages Afrikaans and Dutch. Afrikaans originally stems from Dutch dialects from the 17th century (Raidt, 1991), but due to the geographical distance between the two languages, as well as the complex language contact situation in South Africa, Afrikaans evolved over time into its own independent form that we know today. However, despite various lexical, phonological, morphological, syntactic and semantic changes, the two languages are still considered by and large mutually intelligible (Gooskens and Bezooijen, 2006).

With regard to lexical similarity, approximately 90 to 95% of Afrikaans vocabulary originated in Dutch dialects (Mesthrie, 2002; Kamwangamalu, 2004), while other words in Afrikaans are originally borrowed from other languages such as Bantu, Khoisan, Malay and Portuguese (Sebba, 1997; Niesler et al., 2005). Many of the words in Afrikaans from Dutch origin are not graphologically identical to their original cognates anymore, which means that it is impossible to blindly use Dutch technologies to process Afrikaans text (Pilon et al., 2010). One therefore needs to distinguish between identical cognates (i.e., etymologically-related words from two different languages), non-identical cognates, false friends, and non-cognates when dealing with these two languages. Compare

for instance identical cognates like *donker* (dark) and *periode* (period), non-identical cognates like Afrikaans *beskryf* (describe) with Dutch *beschrijven, beschrijf, beschrijft, beschreven, beschreef* (all inflected forms of the verb), false friends like Afrikaans *aalmoesenier* (almoner) and Dutch *aalmoezenier* (chaplain), and non-cognates like Afrikaans *gottabeentje* and Dutch *telefoonbotje* (ulnar nerve).

Morphosyntactic differences between Afrikaans and Dutch are found, for instance, in Afrikaans verb inflection of main verbs, which follows a much simpler paradigm than that of Dutch. Afrikaans has also lost the distinction between strong and weak verbs, which is noticeable in the conjugation of verbs (De Villiers, 1978). Other systematic differences occur, amongst others, in the gender system, the genitive system and the pronominal system (van Huyssteen and Pilon, 2009).

With specific reference to compounding, a few similarities and differences between Afrikaans and Dutch could be noted. When compound components are identical or non-identical cognates, it most often results in (near) identical compounds, e.g., *alarmknop* (alarm button), or Afrikaans *huurkontrak* and Dutch *huurcontract* (rental agreement). In both languages noun-noun compounds are by far the most productive form of compounding, while verb-noun compounds also occur frequently, e.g., Afrikaans *kookboek* (recipe book), or Dutch *knooppunt* (junction). (See (Verhoeven and van Huyssteen, 2013) for a discussion on the interpretation of verb-noun compounds in Afrikaans and Dutch.). Interestingly, adjective-noun compounds, like Afrikaans *geelwortel* (carrot), occur much more frequently in Afrikaans than in Dutch. Both languages have constructions where a preposition combines with a verb to form a so-called particle verb (also called separable complex verb (Booij, 2010)), and both languages allow for recursive compounding.

In both Afrikaans and Dutch linking morphemes (also called interfixes) play an important role in compounding. Linking morphemes often increase the valency of two components to concatenate in a compound, e.g., in Afrikaans *hondekos* where the *-e-* has a prosodic function. In some cases, linking morphemes occur systematically after certain left-hand components, such as after words ending in *-(i)teit* (e.g., Dutch *faculteitsraad* (faculty board)), or after *wild* in Afrikaans (e.g. *wildskamp* (game enclosure)). In both Afrikaans and Dutch the *-s-* and *-e-* linking morphemes occur frequently, while the *-en-* linking morpheme occurs most in Dutch. For the purposes of this project, we consider the hyphen also as a linking morpheme (linking grapheme), since it occurs in compounds as a means to increase the valency of components ending in vowels to combine with components beginning with vowels, such as Dutch *zee-eend* (scoter).

Up to now, no datasets consisting of compounds annotated using the same annotation guidelines were available for these two languages, which made a cross-lingual analysis of compounding impossible.

Here, we describe the development of uniform annotation guidelines, which are used to annotate compound boundaries in both Afrikaans and Dutch compound words. Using these annotation guidelines, datasets for Afrikaans and Dutch have been developed and inter-annotator agreements have been calculated to evaluate the reliability between annotators. Next, we show the practical usability of the datasets by evaluating an initial automatic compound boundary detection system based on data from the datasets. Even though these datasets are developed with the aim to facilitate a cross-lingual comparison of compounding, the developed datasets may also serve as language resources for other types of research, such as the development or evaluation of language adaptation of computational tools, or cognitively-oriented research on (differences between) the use of compounds in closely-related languages.

First, we will describe the project in which the datasets have been developed. This provides the context of why the guidelines have been developed and how the data has been annotated. Next, we describe the process of the development of the datasets. The datasets are then evaluated in a qualitative way (describing the problems identified during the annotation process) and a quantitative way (indicating the size of the datasets and their inter-annotator agreement). Based on the datasets, initial compound boundary detectors have been developed and their results are discussed briefly as well.

1.1. The AuCoPro project

The Automatic Compound Processing (AuCoPro) project, which is collaborative research between Tilburg University (The Netherlands), University of Antwerp (Belgium), and North-West University (South Africa), deals with the analysis of compounds in both Dutch and Afrikaans. The AuCoPro project has several aims. Most importantly, the project is a first step in the analysis and comparison of compounding in closely-related languages.

Even though research on compound analysis in Dutch and Afrikaans exists (van Huyssteen and van Zaanen, 2004; Pilon et al., 2008), this research is performed on either ad-hoc datasets or datasets that contain additional, more fine-grained morphological information, which introduces noise when focusing on compound boundary information only. To be able to research the use of compound boundaries, we present the development of resources that are specifically designed for compound boundary analysis.

The AuCoPro project consists of two closely-related sub-projects. The sub-project described here deals with the identification of compound boundaries. The second sub-project focuses on the semantic relations that exist between the components found in the compounds (Verhoeven et al., 2012; Verhoeven and van Huyssteen, 2013). To allow for the comparison of compounding in both languages, compound data is collected and manually annotated on two levels: compound boundaries and semantic relations between the components.

Here, we describe results of the sub-project that deals with the manual annotation of compound boundaries. In this sub-project three phases are recognized. Firstly, compounds need to be identified. Initial work on the automatic identification of compounds has already been published (Aussems et al., 2013a; Aussems et al., 2013b). Secondly, the compounds need to be manually annotated or corrected. Finally, the annotation procedures are evaluated

on both the intrinsic properties of the datasets as well as on their usability for the development of compound boundary annotation tools.

2. Approach

The first step in creating datasets containing compounds annotated with their boundaries is to compile a list of compounds to be annotated (manually or automatically). Large corpora are available (for instance, for Dutch the SoNaR corpus (Oostdijk et al., 2008; Oostdijk et al., 2012) consists of 500 million tokens), but it is non-trivial to identify compounds within these texts.

Aussems et al. (2013a) describe an approach developed within the AuCoPro project that can be used to identify Dutch (and potentially Afrikaans) compounds given a set containing both simplex and compound words. This unsupervised system searches for compounds by identifying potential compound boundaries. If a word contains potential compound boundaries according to the system, it is considered a compound.

Even though the unsupervised system works well when used to identify compounds, the potential compound boundaries it identifies do not produce highly accurate annotations of compound boundaries. It seems that manually annotated data are essential in order to build highly accurate compound boundary detection systems.

Instead of relying on only the unsupervised approach, we start by identifying potential compounds from existing Dutch datasets that contain complete morphological information. The underlying idea is that removing undesired morphological information from a dataset containing compounds and their boundaries is easier than identifying the compounds and learn their boundaries in an unsupervised manner.

The compound dataset for Dutch stems from morphologically annotated datasets, which are then modified. However, for Afrikaans no such datasets exist. To allow for the identification of compounds in Afrikaans data, an unsupervised approach is used to identify potential compounds. This approach is based on a longest string matching (LSM) method that identifies potential compounds and inserts provisional compound component boundaries (van Huyssteen and van Zaanen, 2004).

For both languages, all compounds in the datasets are checked manually. This approach enabled faster and more accurate compound boundary annotation compared to annotation from scratch, since most compounds only required boundary verification.

2.1. Initial Dutch dataset

For Dutch, a list of potential compounds is extracted from two initial datasets: the e-Lex[1] dataset and a dataset created by Lieve Macken (personal communication). e-Lex contains approximately 1.1 million morphologically annotated words (including many morphologically complex non-compounds). Based on the morphological structure, 68,855 words contain compound boundaries and these words are selected. This list is extended with the dataset by Macken, which contains 51,249 annotated compounds. Combining the two datasets and removing duplicate words results in a dataset of 71,274 potential compounds. This dataset is already annotated with morphological information, which in many cases corresponds to compound boundaries. The structure of the words found in the e-Lex dataset have been stripped of their morphological information except the potential compound boundary information.

[1] http://tst-centrale.org/producten/lexica/e-lex/7-25

2.2. Initial Afrikaans dataset

The Afrikaans compound dataset originates from two sources, namely the Afrikaans PUK-Protea corpus and the CTexT Afrikaans spelling checker lexicon, originally developed as part of the CKarma project (CText, 2005; Pilon et al., 2008). This dataset is extended by adding unique compounds from the TK corpora (Taalkommissie van die Suid-Afrikaanse Akademie vir Wetenskap en Kuns, 2011). The initial datasets are plain text corpora and a such the words from the corpora do not contain any relevant morphological information.

To identify likely compound boundaries, the LSM algorithm (van Huyssteen and van Zaanen, 2004), which identifies words consisting of two or more correctly spelled components, is used. The output of the LSM algorithm also inserts potential boundary markers for the identified components. This information is retained, allowing for boundary verification. The resulting set contains 77,651 potential compounds.

2.3. Manual annotation

Both initial Dutch and Afrikaans datasets contain morphologically annotated words. These annotations may still be incorrect (when they were automatically annotated) or may denote non-compounds. To identify and correct these potentially incorrect boundaries, the datasets are manually verified to correct erroneous boundary markers and insert missing markers according to annotation guidelines.

2.3.1. Annotation guidelines

The annotation guidelines (Verhoeven et al., In Press) are developed with the underlying aim to provide a consistent annotation in both the Afrikaans and Dutch datasets. The guidelines used in this project are based on the guidelines used for the CKarma project (CText, 2005; Pilon et al., 2008). However, the guidelines are extended for Dutch, additional examples are added and several rules in the guidelines are made more explicit.

The task of annotating the compound boundaries consists of inserting boundary markers between each of the components of the compound. Such boundaries are annotated using the **+** symbol, e.g., Dutch **fiets + schuur** (bike shed). The components of the compounds have to be lexical items that can occur by themselves. In practice, a range of exceptions can be identified. These will be discussed in more detail in Section 3.1..

As mentioned above, both Dutch and Afrikaans make use of linking morphemes (for instance, *-s-*, or *-e-*), which are sometimes required in the construction of compounds and always occur between components. Linking morphemes

are annotated using the _ symbol preceding the linking morpheme, e.g., Dutch *paardenbloemwijn* (dandelion wine, *lit.* "horse flower wine") consists of three stems and a single linking morpheme, *-en-*, which is annotated as **paard _ en + bloem + wijn**. This annotation is shallow without any further hierarchical ordering.

2.3.2. Data annotation

For Afrikaans, seven native Afrikaans annotators participated in the annotation process. In total, $25,266$ potential Afrikaans compounds have been analyzed. For the Dutch dataset, two native Dutch speakers have annotated a total of $26,000$ potential compounds.

Before annotation, the datasets were split into parts of $1,000$ potential compounds each. The annotation of a list of $1,000$ items took approximately one hour. Splitting the entire datasets into parts allowed for easy intermediate saving of progress and also made bookkeeping of annotated items easier.

From the total number of items, a subset was selected which is used to measure annotation quality. For Dutch and Afrikaans, the selection of items for inter-annotator agreement was performed on the level of $1,000$ item parts. For Afrikaans, a total of $12,818$ items were annotated by at least two annotators. Annotations of each of the seven annotators were compared to at least two other annotators. For Dutch, the first part and each following fifth part were annotated by both annotators. Overall, this approach resulted in six overlapping sets (consisting of $6,000$ items in total) that were used to calculate initial inter-annotator reliability for Dutch.

After the completion of a part annotated by multiple annotators, the annotators and supervisor evaluated the between-annotator inconsistencies, identified annotation problems, and adapted the annotation manual if required. Based on the results of the discussions, the annotators went back to all data annotated so far to correct any inconsistencies. These inconsistencies included differences that existed between annotations of the different annotators, but also the items that had to be corrected due to changes (both modifications and extensions) of the guidelines. This process was repeated until the parts used for the calculation of inter-annotator reliability were identical.

After annotation, the resulting Afrikaans dataset consists of $18,497$ true compounds (out of the $25,266$ that have been analyzed) and for Dutch $21,997$ compounds remain from the initial $26,000$ potential compounds. All of these items have at least one compound boundary annotated.

3. Evaluation

To evaluate the process of annotation as well as the resulting annotated datasets, we have evaluated three aspects: the use and modification of the annotation guidelines, the consistency of the annotations by the annotators, and an initial attempt at building a classifier that identifies compound boundaries. Each aspect will be discussed below.

3.1. Annotation guidelines

The annotation guidelines were based on the CKarma annotation guidelines (CText, 2005; Pilon et al., 2008). Given that these guidelines have already been used in the CKarma project, it led us to believe that they would form a good basis for this project as well. The guidelines were specifically designed for Afrikaans, so they had to be extended to handle Dutch compounds as well.

During the annotation process, several problematic cases were identified and the annotation manual was adjusted or extended to handle these problems. In particular, compounds containing prepositions, allomorphs, synthetic and derived compounds were identified during the annotation process.

The annotation of compounds that include prepositions, such as Dutch *aanval* (attack), is problematic, as the potential components *aan* and *val* (on + fall) do not describe the meaning of the word (i.e., the meaning is non-compositional). It was decided that in these cases, prepositions are not annotated as separate components. However, in the situation where two prepositions are combined as a part of a compound, they do serve as proper (semantic) components in the compound. For instance, Dutch *achteruitkijken* (looking backwards) is structured as **achteruit + kijken**. These compounds are annotated in the datasets.

Even though the general rule used during annotation is that the components should be proper lexical items, this is not always the case. For instance, Dutch *botenschuur* (boat shed) is analyzed as **bot _ en + schuur**, where the component *bot-* is an allomorph of the word *boot* (boat). Such allomorphs were therefore allowed in the datasets.

The problem of synthetic compounds, such as Afrikaans *besluitneming* (decision making) initially seem to consist of meaningful components, respectively *besluit* (decision) and **neming* (taking). However, in the case of synthetic compounds, the combined components of the compound are morphologically modified by the compounding process. Since **besluitneem* is not a verb in Afrikaans, *besluitneming* cannot be analyzed as a derived compound (i.e., **besluitneem* + *ing*), and since **neming* is not a valid word, it can also not be analyzed as **besluit* + *neming*. Its morphological structure is rather that of a verb phrase (*besluit neem* that combines with a suffix (*-ing*). This complexity has led to the decision that synthetic compounds are not annotated in these datasets.

Similarly, derived compounds such as Dutch *persoonlijk* (personal), may initially seem to be composed of the components *persoon* (person) and *lijk* (corpse). However, it may be clear that the meaning of the components of derived compounds do not correspond to the meaning of the compound as a whole and as such, it has been decided that they are not annotated as compounds.

In addition to providing rules on how to deal with the problematic cases of compounds containing prepositions, allomorphs, synthetic and derived compounds, the guidelines describe a range of potentially difficult situations. These include rules on how to deal with compounds with multiple analyses, affixes that also correspond to regular words, highly lexicalized words, or compounds containing names. Additionally, special situations, such as Internet-related words (such as Twitter hash-tags or URLs), compounds containing dashes, numbers (either in digits or de-

	Afrikaans	Dutch
Initial number of items	25,266	26,000
Number of remaining compounds	18,497	21,997
Average number of compound boundaries	1.13	1.07
Average number of linking morphemes	0.33	0.31
Items used for evaluation	12,818	6,000
Number of annotators	7	2
Average Cohen's kappa	98.6 (0.8)	97.6 (0.7)
Average word-level agreement	96.8% (2.1)	95.3% (1.8)
Classification accuracy	88.28%	91.48%

Table 1: Quantitative properties of the Afrikaans and Dutch datasets. Standard deviations are given between brackets.

scribed in words), brackets or other non-letter characters are described. Finally, nonsense words, typos and foreign or archaic words used as components on compounds are also discussed.

During the annotation and verification process, the annotators found several differences between the initial structures (that were extracted from the original dataset) and the annotation according to the guidelines. These differences fall in roughly three categories, namely particle verb errors, incorrect semantic boundary detection, and identifying non-words as compound components.

Particle verbs, such as Dutch *omkopen* (to bribe) are not annotated. However, from a morphological perspective, there is a boundary between *om* and *kopen* as they are often split when used in sentential context. In the initial dataset, particle verbs when found in larger compounds, such as Dutch *omkoopgeld* (bribe money) were annotated incorrectly with a compound boundary between *om* and *koop* (with the additional component *geld*).

Incorrect boundary detection also leads to dividing non-compounds into semantically improper components, e.g., Afrikaans *stereotipe* (stereotype) was sometimes split into *stereo* (stereo) and *tipe* (type). In this case there is no semantic basis for analyzing the word, although there are two correct components in the word.

Identifying non-words as compound components produces errors where words are split into components that are not stems, words, or linking morphemes, e.g., Dutch **tentoon + stelling** (exhibition), where *tentoon* is not a recognized Dutch word. Even though these problems are addressed in the guidelines, annotators were still making these mistakes.

3.2. Annotators

To evaluate the consistency between the annotators, inter-annotator reliability has been measured. For evaluation purposes, each position between letters in a compound is considered an annotation. Therefore, a string such as *abc* has two annotations, namely between the *a* and the *b* and between the *b* and the *c*. The inter-annotator reliabilities for the datasets presented here are computed directly after the completion of a part of the dataset that was annotated by two annotators, before any corrections were made.

The raters' overall agreement is computed using Cohen's kappa (Cohen, 1960) and is averaged over the six different parts for Dutch and thirteen parts for Afrikaans, each consisting of $1,000$ compounds. The average Cohen's kappas and their standard deviations are given in Table 1. Both kappas are considered being highly reliable ($k > 80$).

Additionally, we computed word-level agreement. Per annotator pair, we computed the percentage of identically annotated words. These results are also presented in Table 1.

3.3. Classification

The original reason for developing the compound datasets for both Afrikaans and Dutch was to allow cross-lingual comparison of compounding. However, we already mentioned that such datasets could also be useful for other purposes.

To show practical usability of the datasets, we report here on an initial attempt to build automatic compound boundary detection systems. This experiment is not meant to show the state-of-the-art of automatic compound boundary detection systems, but to illustrate the usefulness of the datasets for such a task. We have decided on this particular problem because it can be performed completely automatically and does not require a deep analysis of the results (which is the case in, for instance, a cross-linguistic analysis of compounding).

The process of compound boundary detection is quite similar to that of syllabification (which identifies syllable boundaries in words) or hyphenation (which identifies potential breaks in words allowing for their hyphenation). This lead us to use the well-known and practically successful hyphenation system of Liang (1983). This system is also used in the LaTeX typesetting system.

Given the annotated data, we run `patgen` in several iterations to attain good results on the training data. `Patgen` is the pattern generation system that comes with LaTeX. The result of this step is a list of patterns that indicate positions between letters that typically do or do not allow for compound boundaries. Using the Tex-Hyphen-1.01 Perl module[2], we can now apply the patterns generated by `patgen` words in order to identify compound boundaries.

The datasets are both evaluated using leave-one-out. This means that each compound in the dataset is used as testing data once, while the remainder of the dataset is used as training data. A split between test and training data is essential (otherwise a simple lookup system would lead to perfect results). However, we want to keep as much training data as possible. Applying leave-one-out means that for

[2] `http://www.adelton.com/perl/TeX-Hyphen/`

Dutch 21,997 experiments are run and 18,496 experiments for Afrikaans. (One compound in the dataset for Afrikaans, **algemene _ - + Onderwys _ - + en _ - + Opleiding + sertifikaat** is too long to be handled by `patgen`. We left this out of the evaluation.) The classification accuracies can be found in Table 1.

4. Discussion & Conclusion

To enable researchers to investigate cross-language comparisons of linguistic processes, having access to comparable data in different languages is essential. Here, we have discussed the development of datasets containing compounds and their component boundaries using the same annotation guidelines, applied to the two closely-related languages Afrikaans and Dutch.

In order to ensure a high inter-annotator reliability, annotation guidelines, that were originally developed for Afrikaans, were used as a starting point. These guidelines were modified and extended to support the annotation of Dutch compounds as well. The development of a cross-language annotation manual already provided some insights in the differences between Afrikaans and Dutch.

The evaluation of the data was performed on three levels. Firstly, during the annotation process, regular discussions with the annotators took place, which indicated difficult situation that required more extensive explanation in the guidelines as well as problematic cases that were not (yet) handled by the guidelines. Secondly, given the level of inter-annotator reliability as well as the word-level agreement for both languages, the cross-language transfer of knowledge in the guidelines was very successful. Finally, the datasets have been successfully used in an example system that automatically identifies compound boundaries.

The availability of the datasets enables a wide range of future research directions. The quality of the datasets indicate that both monolingual as well as cross-linguistic analyses of Afrikaans and Dutch from different perspectives are now possible. This research could focus on linguistic similarities and differences between the languages.

The datasets can also be used for a variety of applications. For instance, they could serve as basis for the development of (language independent) compound analysis techniques. These compound analyzers can be used in different natural language processing technologies to improve their overall performance. Additionally, these datasets allow for the development and evaluation of domain or language-adaptation approaches, in which a compound analysis tool in one domain or language benefits from data in another.

To conclude, we have described the development of datasets for Afrikaans and Dutch containing compounds and their shallow morphological structure. The evaluation shows that the annotation efforts resulted in useful language resources, which provide a good basis for compound analysis related tasks.

5. Acknowledgments

We would like to thank the anonymous reviewers for their useful comments. This research was funded by a joint research grant of the Nederlandse Taalunie (Dutch Language Union) and the Department of Arts and Culture (DAC) of the South African Government for a project on automatic compound processing (AuCoPro[3]). The project was also supported through a grant from the South African National Research Foundation (grant number 81794). Views expressed in this publication cannot be assigned to any of the funders, but remain that of the research groups of the North-West University (South Africa), the University of Antwerp (Belgium) and Tilburg University (The Netherlands).

6. References

S. Aussems, S. Bruys, B. Goris, V. Lichtenberg, N. van Noord, R. Smetsers, and M. van Zaanen. 2013a. Automatically identifying compounds. In *Book of abstracts of the 23rd meeting of Computational Linguistics in the Netherlands*, page 10, Enschede, University of Twente.

S. Aussems, B. Goris, V. Lichtenberg, N. van Noord, R. Smetsers, and M. van Zaanen. 2013b. Unsupervised identification of compounds. In *Proceedings of BENELEARN, Nijmegen*, pages 18–25.

G. Booij. 2010. *Construction morphology*. Oxford University Press, Oxford.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

CText. 2005. Ckarma: C5 kompositumanaliseerder vir robuuste morfologiese analise. Technical report, Centre for Text Technology, North-West University, Potchefstroom, South Africa.

M. de Villiers. 1978. *Nederlands en Afrikaans (Dutch and Afrikaans)*. Nasou, Cape Town.

C. Gooskens and R.V. Bezooijen. 2006. Mutual comprehensibility of written Afrikaans and Dutch: symmetrical or asymmetrical? *Literary and Linguistic Computing*, 21:543–557.

N. Kamwangamalu. 2004. The language policy/language economics interface and mother-tongue education in post-apartheid South Africa. *Language Problems and Language Planning*, 28:131–146.

P. Koehn and K. Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistic*, volume 1, pages 187–193.

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

R. Mesthrie. 2002. *Language and social history: Studies in South African sociolinguistics.* David Philip, Cape Town.

T.R. Niesler, P.H. Louw, and J.C. Roux. 2005. Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases. *Southern African Linguistics and Applied Language Studies*, 23(4):459–474.

N. Oostdijk, M. Reynaert, P. Monachesi, G. van Noord, R. Ordelman, I. Schuurman, and V. Vandeghinste. 2008. From D-Coi to SoNaR: A reference corpus for Dutch. In *Proceedings of the sixth international conference on language resources and evaluation (LREC)*, pages 1437–1444, Marrakech, Marokko. ELRA.

[3] `http://tinyurl.com/aucopro`

N. Oostdijk, M. Reynaert, V. Hoste, and I. Schuurman. 2012. The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns and J. Odijk, editors, *Essential speech and language technology for Dutch: Results by the STEVIN-programme*, chapter 13, pages 201–226. Springer-Verlag.

S. Pilon, M.J. Puttkammer, and G.B. van Huyssteen. 2008. Die ontwikkeling van n woordafbreker en kompositumanaliseerder vir Afrikaans (the development of a hyphenator and compound analyser for Afrikaans). *Literator*, 29:21–41.

S. Pilon, G.B. van Huyssteen, and L. Augustinus. 2010. Converting Afrikaans to Dutch for technology recycling. In *Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa*, pages 219–224.

E.H. Raidt. 1991. *Afrikaans en sy Europese verlede (Afrikaans and its European past)*. Nasou, Cape Town.

M. Sebba. 1997. *Contact languages: pidgins and creoles*. Palgrave Macmillan.

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

G.B. van Huyssteen and S. Pilon. 2009. Rule-based conversion of closely-related languages: a Dutch-to-Afrikaans convertor. In *Proceedings of the Twentieth Annual Symposium of the Pattern Recognition Association of South Africa*, pages 23–28.

G.B. van Huyssteen and M.M. van Zaanen. 2004. Learning compound boundaries for afrikaans spelling checking. In *Pre-Proceedings of the Workshop on International Proofing Tools and Language Technologies*, pages 101–108.

B. Verhoeven and G.B. van Huyssteen. 2013. More than only noun-noun compounds: Towards an annotation scheme for the semantic modelling of other noun compound types. In *Proceedings of the Ninth Joint ISO - ACL Workshop on Interoperable Semantic Annotation*, pages 59–66.

B. Verhoeven, W. Daelemans, and G.B. van Huyssteen. 2012. Classification of noun-noun compound semantics in dutch and afrikaans. In *Proceedings of the Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2012)*, pages 121–125.

B. Verhoeven, G.B. van Huyssteen, M. van Zaanen, and W. Daelemans. In Press. Annotation guidelines for compound analysis. *CLiPS Technical Report Series (CTRS)*, 5.