

Automatic Compound Semantic Analysis using Wordnets

Zandré Botha, Roald Eiselen & Gerhard B. van Huyssteen

Centre for Text Technology
North-West University, Potchefstroom Campus
Potchefstroom, South Africa
{2218166; roald.eiselen; gerhard.vanhuissteen}@nwu.ac.za

Abstract— Compound semantic analysis is the task of finding the correct internal relation between the constituents of a compound [3, 10]. In this paper we use a measure of semantic similarity [14] based on the relations in the Afrikaans WordNet [2] to determine the similarity between two Afrikaans compounds. We infer that if the different constituents of two compounds are semantically similar, the compounds will have the same internal semantic relation between the constituents. This wordnet-based approach is compared to existing approaches and we show that the approach attains similar precision, but lower recall. To improve recall, we investigate a method for automatically extending the Afrikaans WordNet based on freely available wiki-resources.

Keywords—compound semantic analysis; semantic relations; Afrikaans WordNet; wordnet extension; Wiki-resources

I. INTRODUCTION

The automatic analysis of a compound is an important technology that is used in various fields of natural language processing, such as machine translation, information extraction, information retrieval, spelling checkers, and question answering systems [8, 15, 16]. Automatic compound analysis is usually concerned with two problems, namely finding compound boundaries and determining the semantic relations between the different constituents of a compound. This paper focuses on compound semantic analysis (CSA) of Afrikaans noun-noun (NN) compounds as part of the Automatic Compound Processing¹ (AuCoPro) project. CSA is the task of finding the semantic relation between the head noun in the compound and its modifier [3, 10]. This paper investigates a method for classifying the internal semantic relation of a NN compound in Afrikaans automatically by using the Afrikaans WordNet (AWN) [2] as knowledge base. Wordnets are useful for CSA because of the semantic relations (such as hyponymical and meronymical relations) a wordnet structure provides, which can be used to determine the internal semantic relation of a NN compound.

The result of using a wordnet for automatic CSA in Afrikaans is compared with existing research on two levels. Firstly, the wordnet-based approach is compared to a bag of

words approach to CSA proposed by [17] to determine if the wordnet-based approach can improve on their reported results. Secondly, the results are compared with English CSA using a wordnet to determine what impact wordnet size has on automatic CSA. When compared to other wordnets previously used for automatic CSA, such as the Princeton WordNet (PWN) [4] used by [10], the AWN is relatively small (10,045 synsets). Since the wordnet is an integral part of finding relational information, the size of the wordnet could have a significant impact on the CSA results.

As part of this investigation, we apply an approach to automatically extend the AWN to determine if results for automatic CSA in Afrikaans will improve with a larger wordnet. The extension of the AWN is done automatically by using existing, freely available lexical data [1] such as Wiktionary² and the Unicode Common Locale Data Repository³ (UCLDR) data.

The remainder of this paper is structured as follows. Section 2 gives a detailed description of how the AWN is used to automatically determine the internal semantic relations of unseen NN compounds in Afrikaans, and also includes an explanation of the evaluation procedure used to determine the success of this approach. Section 3 details the approach we followed for the automatic extension of the AWN, together with results of the extension procedure and results for CSA on an extended AWN. Section 4 provides a conclusion and outlines future work.

II. AUTOMATIC COMPOUND SEMANTIC ANALYSIS

A. Previous Work

A noun compound is a word that consists of two or more nouns [10]. As an example, the Afrikaans compound *koperpan* (“copper pan”) can be split into the two nouns *koper* and *pan*, and semantic analysis describes the relation between the two nouns as “a pan made out of copper”. The scheme for annotating these semantic relations, summarised in Table I, is described in [13] and was adopted by [17] and the AuCoPro

¹ tinyurl.com/aucopro

² <http://dumps.wikimedia.org/backup-index.html>

³ <http://cldr.unicode.org/>

project. In order to build a compound semantic analyser, an automatic method of finding the internal relation between two constituents of a compound needs to be developed.

This paper applies an approach similar to the one proposed by [10] of measuring the semantic similarity of an unseen NN compound to annotated NN compounds (i.e. compounds where the internal semantic relation between constituents has been assigned by a human) to determine the most likely internal semantic relation between the constituents of the unseen compound. This method is applied in order to determine how well it performs for automatic CSA in Afrikaans and is described in detail later in this section.

Most research concerning automatic CSA has been done for English [3, 10], although there has been research in other languages, such as Dutch [16, 17], German [8], and Afrikaans [16, 17] as well. A method where the top levels of a hierarchical knowledge base (such as a wordnet) are used to make distinctions between axioms (the starting point of the specific word in the wordnet hierarchy) and ontologies (sub-category of the wordnet) was applied in [3]. Using the relations within the PWN, [10] used a measure of semantic similarity to determine the semantic relatedness between two noun compounds. This method is described in greater detail later in this section. A deduction based approach was explored in [8] to infer the internal semantic relations of German compounds based on the meronymical hierarchy in GermaNet (the German WordNet).

Previous work related to automatic CSA for Afrikaans and Dutch was done by using a bag of words (BOW) approach [17]. For training data, each instance vector contained a category (the semantic specific category of the compound) and features such as the co-occurrence of each word. Using WEKA's sequential minimal optimisation (SMO) implementation of a support vector machine learning algorithm, [17] used 1,439 Afrikaans NN compounds and 1,447 Dutch NN compounds as the vectors for each language respectively. We use the results of [17] as a baseline for comparison with the results from our approach.

TABLE I. SEMANTIC RELATION ANNOTATION SCHEMA FOR ANNOTATED COMPOUNDS (N1 AND N2 ARE THE TWO CONSTITUENTS OF THE COMPOUND IN THAT ORDER)

Relation	Description	Example
BE	N2 which is (like) (a) N1.	<i>gidshond</i> ('guide dog')
HAVE	All compounds denoting some sort of possession.	<i>studieprobleme</i> ('study problems')
IN	Any compound denoting a location in place or time.	<i>varkhok</i> ('pigsty')
ACTOR	When there is an event denoted in the compound and one of the constituents is a salient entity.	<i>mielieboer</i> ('corn farmer')
INST	When there is an event and there is no salient entity present.	<i>voetspoor</i> ('foot imprint')
ABOUT	Describes 'an item is ABOUT something'	<i>geskiedenisboek</i> ('history book')
REL	Other Non-Lexicalised Relation.	<i>modenaam</i> ('fashion name')
LEX	Lexicalised Compound.	<i>lewensverskerking</i> ('life assurance')

UNKN	The meaning is unclear.	<i>naelhand</i> ('nail hand')
MISTAG	Incorrectly tagged as NN compound.	<i>snoeikunstenaar</i> ^a ('pruning artist')
NONC	Not a NN compound.	<i>walvis</i> ^b ('whale')

^a The compound *snoeikunstenaar* is classified as MISTAG because *snoei* is not a noun.

^b The word *walvis* ('whale') is not a compound, although it may be incorrectly classified as a compound because it can be split into two valid constituents, *wal* ('embankment' or 'mound') and *vis* ('fish').

B. Methodology

In this paper we follow a method similar to that of [10] for assigning internal semantic relations to previously unseen NN compounds in Afrikaans. As in [10], we focus exclusively on NN compounds consisting of two constituents, the head noun and the modifier. This is a pragmatic decision, as 78.3% of Afrikaans compounds consist of only two constituents [15].

In order to determine the internal semantic relation of a NN compound, we use a measure to calculate the semantic similarity between each of the constituents of an unseen compound and compounds annotated with the internal semantic relations. The annotated NN compound with the highest similarity score is identified and its semantic relation is assigned to the unseen NN compound.

For this paper we use the same annotated data as [17]. This data consists of 1,499 Afrikaans compounds, each of which has its internal semantic relation annotated. A compound can be classified as having one of eleven types of internal relations. Table I presents each relation, together with an example and a description. Table II gives the class distribution of each relation in the annotated data.

The first step of the automatic CSA process is to calculate the similarity score between annotated NN compounds and previously unseen compounds, based on the semantic relations in the AWN using the WordNet::Similarity module [14]. WordNet::Similarity is an open source Perl module that allows a user to calculate the semantic similarity between a pair of words based on the PWN. WordNet::Similarity provides six different measures of similarity, divided into path-based measures and information content based measures. We take a similar approach to [10], by combining four similarity measures: two path-based measures, LCH [11] and WUP [19], and two information content-based measures, JCN [9] and LIN

TABLE II. DISTRIBUTION OF SEMANTIC RELATION CLASSES IN THE ANNOTATED DATA

Relation	Number of instances
BE	365
HAVE	144
IN	302
ACTOR	130
INST	110
ABOUT	413
REL	25
LEX	0
UNKN	2
MISTAG	7
NONC	1

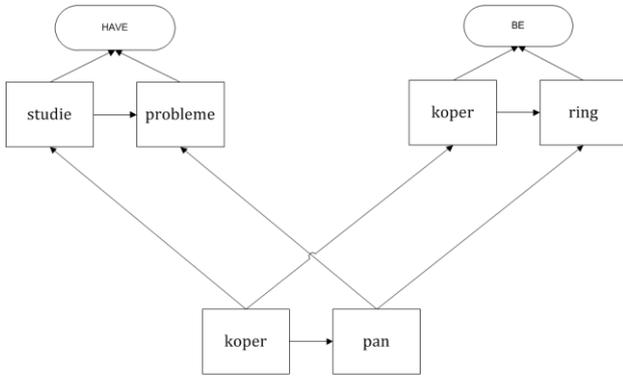


Fig. 1. Illustration of similarity between *kopperpan* and annotated NN compounds *studieprobleme* and *koperring* (adapted from [10])

[12]. Each score is normalized and summed to get a combined similarity score. The four measures are described as follows in [14]:

- **LCH:** Finds the shortest path between two concepts, and scales that value by the maximum path length found in the “*is-a*” hierarchy in which they occur.
- **WUP:** Finds the depth (distance to the root node) of the least common subsumer (LCS) of the concepts, and then scales that by the sum of the depths of the individual concepts.
- **JCN and LIN:** Augment the information content of the LCS with the sum of the information content of concepts A and B themselves. The LIN measure scales the information content of the LCS by this sum, while JCN takes the difference of this sum and the information content of the LCS.

We start by calculating the word similarity between the N1 constituents of the unseen and annotated compounds, followed by the word similarity between the N2 constituents of the unseen and annotated compounds. A simple example (see Figure 1) illustrates how the similarity between an unseen compound (*kopperpan*) and two annotated compounds, *studieprobleme* (study problems) and *koperring* (copper ring), is calculated. The word similarity score (see Table III) between N1 of the unseen compound and N1 of the annotated compound (*studieprobleme*) is 0.06, while the similarity score between N2 of the unseen and annotated compound is 0.24. The process of comparing annotated compounds with the unseen compound is repeated for every compound in the annotated data. Subsequently, there is a similarity score between N1 of the unseen compound and N1 of every compound in the annotated data, and a similarity score between N2 of the unseen compound and N2 of every compound in the annotated data.

TABLE III. WORD SIMILARITY SCORES OF CONSTITUENTS BETWEEN *KOPERPAN*, *STUDIEPROBLEME* AND *KOPERRING*

	Annotated NN compound	Unseen NN compound	Word similarity score	Compound similarity
N1	studie	koper	0.06	0.014
N2	probleme	pan	0.24	
N1	koper	koper	1	0.53
N2	ring	pan	0.53	

The second step in determining the internal relation of an unseen compound is to calculate the similarity of the annotated compounds to the unseen compound. This is done by multiplying the similarity scores of N1 and N2 of the annotated compound for every compound in the annotated data. The results in Table III show that the similarity score between *studieprobleme* and the unseen compound is lower than the similarity score between *koperring* and the unseen compound.

The compound with the highest similarity score in the annotated data is identified and its internal semantic relation is assigned to the unseen compound. The example results in Table III show that *kopperpan* is the most similar to *koperring* and we can infer that the internal semantic relation for *kopperpan* therefore is BE.

One of the shortcomings of using the AWN is that it does not necessarily contain inflections of the words that occur in compounds, such as the plural form *probleme* (“problems”) in *studieprobleme*. Of the 2,998 constituents in the data, only 1,875 are lemmas that are included in the AWN. To mitigate this problem, we used the Lemma-Identifier for Afrikaans (LIA) [6; 7] to lemmatise constituents not found in the AWN. LIA takes any Afrikaans word as input and gives the linguistic correct lemma as output. For example, LIA will lemmatise the inflected word *probleme* (“problems”) as *probleem* (“problem”). Of the remaining 1,123 constituents, an additional 410 were found in the AWN after lemmatisation. The impact on overall results of using LIA as part of the automatic CSA is discussed in the evaluation section.

C. Evaluation

The success of the wordnet-based approach to automatic CSA for Afrikaans is determined by measuring the precision, recall and *F*-score using 10-fold cross validation on the 1,499 annotated compounds. The previously described similarity method for automatic CSA is used to classify 10 randomly selected sets of 145 NN compounds from the annotated data, and calculating the similarity to the remaining 1,354 annotated compounds in each of the sets. The average precision, recall and *F*-score are then calculated as the 10-fold cross validation result. This process is also repeated for the same data set using LIA as an additional resource for identifying inflected constituents of which only the lemma is included in the AWN.

$$Precision = \frac{(\sum \text{unseen compounds correctly classified})}{(\sum \text{unseen compounds that were classified})} \quad (1)$$

$$Recall = \frac{(\sum \text{unseen compounds correctly classified})}{(\sum \text{unseen compounds in data set})} \quad (2)$$

$$F - score = 2 * \frac{(\text{precision}) * (\text{recall})}{(\text{precision}) + (\text{recall})} \quad (3)$$

TABLE IV. EVALUATION METRICS FOR AUTOMATIC CSA ON AFRIKAANS COMPOUNDS

	Precision	Recall	F-score
Baseline BOW	50.80%	51.60%	51.10%
AWN without LIA	48.02%	18.10%	26.22%
AWN with LIA	50.49%	29.27%	37.05%

The evaluation results in Table IV provide a comparison of the baseline CSA system developed by [17] to the wordnet-based approach explored in this paper and the approach including LIA as an additional resource. The results show that although the precision of the wordnet-based approach is comparable to the BOW approach, the recall is significantly lower for the wordnet-based approach. The inclusion of LIA in the process improves recall and the *F*-score by more than 10%, while also slightly improving the precision, but this still does not compare favourably with the baseline results.

The poor recall of the wordnet-based approach can largely be attributed to the small size of the AWN (10,045 synsets), since many of the constituents in the data do not have entries in the AWN. If one of the constituents of the unseen compound does not exist in the AWN, a similarity score for the unseen compound cannot be calculated. If a similarity score for an unseen compound cannot be calculated, the recall score, and subsequently the *F*-score, for this approach to automatic CSA drops. This problem can be mitigated by extending the AWN so that it has broader coverage. The following discussion details an automatic extension procedure of the AWN.

III. AUTOMATIC EXTENSION OF THE AFRIKAANS WORDNET

A. Previous Work

The advantage of using a wordnet as knowledge base for CSA is that the similarity measure compares the semantic similarity of two words, based on the different semantic relations of the wordnet. As explained in the previous section, the AWN (10,045 synsets) may be too small or incomplete in comparison to the PWN used in [10] (see Table IV) to achieve comparable results for automatic CSA. One way to mitigate this problem is to extend the AWN so that it has broader coverage.

Previous approaches to automatically extending a wordnet with freely available lexical data includes a method of looking at back-translations found in wiki resources [7], and a method of using data gathered from Wiktionary and the UCLDR to extend wordnets of various languages [1]. The method used by [7] is more applicable to languages with medium to large coverage wiki resources because it is dependent on synset-aligned wordnets and a large multilingual translation graph in as many languages as possible - none of which is currently available in Afrikaans. In this paper we apply a method similar to the one described in [1] to extend the AWN.

TABLE V. SIZE COMPARISON BETWEEN PWN, AWN AND AWN-E

	PWN	AWN	AWN-E
Nouns	79,689	6,673	7,657
Verbs	13,508	2,966	2,981
Adjectives	18,563	406	480
Adverbs	3,664	0	0
Total number of synsets	115,424	10,045	11,118

Wiktionary is an online dictionary that was designed as a lexical companion to Wikipedia and consists of lexical data, such as words, parts of speech, definitions, translations, synonyms and antonyms. Wiktionary data is useful for wordnet extension because it contains translations to the same Wiktionary entries in other language as well as different lexicographic information that can be integrated into a wordnet, and it is freely available.

The first step in the extension process [1] is to build a custom parser to extract words, parts of speech, definitions, synonyms and translations from the English Wiktionary pages.

By using multiple text similarity scores to compare a BOW for the gloss, set of lemmas, and possible example sentences for entries in Wiktionary and synsets in WordNet, the Wiktionary senses are linked to WordNet synsets. Once a Wiktionary sense is linked to a WordNet synset, the information from the Wiktionary entry is used to add or extend parts of the WordNet synset that may be incomplete or missing from the WordNet. In addition to the Wiktionary data, [1] also uses UCLDR data. UCLDR is a source of freely available lexical data that consist of information on date formats, numbers, currencies, times, and time zones, as well as help for choosing languages and countries by name.

B. Methodology

To extend the AWN, we first look at the data gathered from the Afrikaans Wiktionary. When compared to the Wiktionary of either English (36,650,000 entries) or Dutch (303,485 entries), the current Afrikaans Wiktionary (20,765 entries) is small and incomplete. Subsequently, the data from the Afrikaans Wiktionary that can be used to add new entries in the current AWN is relatively small. Even so, as the Afrikaans Wiktionary continues growing it can be used to further extend the AWN in the future.

To start the extension of the AWN with data from Wiktionary, we parse Afrikaans, English, and Dutch Wiktionaries for words, parts of speech, a short gloss, and translations (between English and Afrikaans, and Dutch and Afrikaans). We then create a list of entries in the Afrikaans Wiktionary with translations to the English Wiktionary.

Since each synset in the AWN has the same synset id as its counterpart in the PWN, the list of English translations is then aligned to the WordNet. Through a simple lookup we can determine which synset ids are not present in the AWN. New synsets are added to the AWN if they have a translation to an English Wiktionary entry. These Wiktionary entries are linked to their WordNet senses.

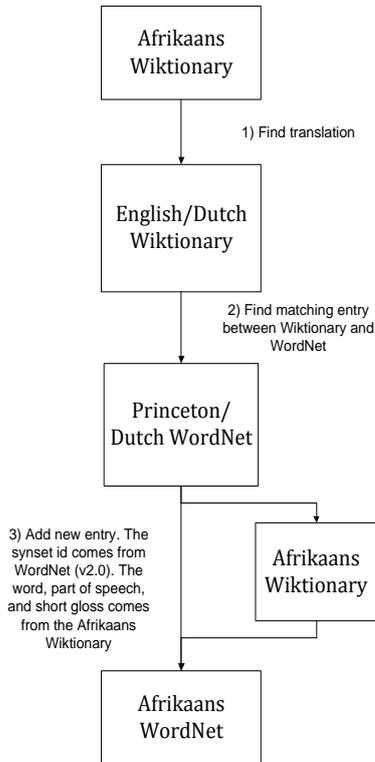


Fig. 2. Graphic illustration of the Afrikaans WordNet extension procedure

A Wiktionary entry that does not have the same part of speech as the translation in the Afrikaans Wiktionary, or that already exists in the AWN, is filtered out. New synsets are added to the AWN with the word, part of speech, and gloss from the Afrikaans Wiktionary entry, and the synset id from the WordNet. We follow the same procedure as described above to extend the AWN with data from the Dutch WordNet (Cornetto) and Dutch Wiktionary, because Cornetto also has a PWN synset id for each corresponding synset. Results of the extension are presented in Table VI, showing a total of 1,073 new synsets added to the AWN in this way.

The small number of synsets that were added to the original AWN is largely due to the small size of the Afrikaans Wiktionary. Many of the entries in the Afrikaans Wiktionary do not have translations to corresponding entries in the English or Dutch Wiktionary, or they already exist in the AWN, so that they cannot be added to the Afrikaans WordNet as a new synset. Also, because the Afrikaans WordNet has the same synset ids as the PWN only synsets that already exist in the PWN can be added as synsets to the Afrikaans WordNet.

TABLE VI. RESULTS OF THE AWN EXTENSION

	Wiktionary	UCLDR
Nouns	526	458
Verbs	15	0
Adjectives	74	0
Total	615	458

C. Evaluation

After the completion of the AWN extension, the same procedure for calculating metrics with 10-fold cross validation was run using the extended version of the Afrikaans WordNet (AWN-E) as knowledge base. Surprisingly this method achieved an F -score of 35.71% (with 47.96% precision and 28.46% recall), slightly worse than results from the results using the original AWN. The main reason for this drop in precision and recall is the fact that some synsets added to AWN-E are placed incorrectly in the wordnet hierarchy. The synsets are added incorrectly because we cannot link senses between the Afrikaans Wiktionary and PWN/Cornetto by using multiple text similarity scores as in [1]. Instead we use the translations found in the Afrikaans Wiktionary, and ambiguous entries were added to incorrect locations in the WordNet.

IV. CONCLUSION

This paper used a measure of word similarity based on the semantic relations in the Afrikaans WordNet to determine the semantic relations between compound constituents. We infer that if one compound is semantically very similar to another, then it will also have the same internal semantic relation.

The results of this approach to automatic CSA for Afrikaans (F -score of 37.05%) does not compare well to the BOW approach used by [17], which achieved an F -score of 51.1%. Although the precision of our approach is almost on par with [17] (50.8%), our recall score is much lower because it is dependent on a limited knowledge-base, which consequently has a negative impact on the overall F -score.

Our results compare more favourably to that of [10], which achieved an accuracy of 53% (calculated in the same way as our precision score), although our precision score is slightly lower at 50.49%. This is most likely due to the fact that the Afrikaans WordNet is still much smaller than the PWN used by [10], even after the automatic extension. As an attempt to validate this assumption, we applied an approach to automatically extending the AWN through openly available resources, adding 1,073 new synsets.

These initial results indicate that using a small wordnet limits the possible quality that can be attained for automatic CSA, but that the precision of the class assignment is similar to that of other approaches. Until larger wordnets or ontologies are available, it will be more successful to use machine learning techniques to learn these types of semantic relations. Further work on improving the automatic extension techniques may also be a beneficial resource development strategy, but requires larger and more comprehensive data for the strategy to be really beneficial.

REFERENCES

- [1] F. Bond, & R. Foster, "Linking and extending an open multilingual Wordnet," in *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, Sofia, Bulgaria, 2013.
- [2] CText. Afrikaans Wordnet. Potchefstroom: North-West University, 2011.
- [3] J. Fan, K. Barker & B. Porter, "The knowledge required to interpret noun compounds," in *Proceedings of the 18th International Joint*

- Conference on Artificial Intelligence*, Acapulco, Mexico, 2013, pp. 1483-1485.
- [4] C. Fellbaum, *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press, 1998.
- [5] H.J. Groenewald, "Automatic lemmatisation for Afrikaans," M.Sc. thesis, Dept. Elect. Eng, North-West Univ., Potchefstroom, 2006.
- [6] H.J. Groenewald, G.B. Van Huyssteen, A.S.J. Helberg, A. Van Den Bosch, *Lemma-Identifier for Afrikaans*, Potchefstroom: North-West University, 2011.
- [7] V. Hanoka, & B. Sagot, "Wordnet creation and extension made simple: a multilingual lexicon-based approach using wiki resources," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012, pp. 3473-3479.
- [8] E. Hinrichs, V. Henrich, & R. Barkey. (2013, May 13). *Using part-whole relations for automatic deduction of compound-internal relations in GermaNet* [Online]. Available: <http://link.springer.com/article/10.1007%2Fs10579-012-9207-y>
- [9] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan, 1997, pp. 19-33.
- [10] S.N. Kim & T. Baldwin, "Automatic interpretation of noun compounds using WordNet Similarity," in *Natural Language Processing-IJCNLP 2005*, Jeju Island, South Korea, 2005, pp. 945-956.
- [11] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," C. Fellbaum, Ed. Cambridge, MA: MIT press, 1998, pp. 265-283.
- [12] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the International Conference on Machine Learning vol. 98*, Madison, 1998, pp. 296-304.
- [13] D. Ó'Séaghdha, "Learning compound noun semantics," Univ. of Cambridge., Cambridge, MA, Tech. Rep. 735, 2008.
- [14] T. Pedersen, S. Patwardhan & T. Michelizzi, "WordNet::Similarity – measuring the relatedness of concepts," in *Demonstration Papers at HLT-NAACL 2004*, Stroudsburg, PA, 2004, pp. 38-41.
- [15] G.B. Van Huyssteen & M.M. Van Zaanen, "Learning compound boundaries for Afrikaans spelling checking," in *Pre-Proc Workshop on International Proofing Tools and Language Technologies*, Patras, Greece, 2004, pp. 101-108.
- [16] B. Verhoeven, "A computational semantic analysis of noun compounds in Dutch," M.Sc. thesis, Dept. Linguistics, Univ. of Antwerp, Antwerp, 2012.
- [17] B. Verhoeven, W. Daelemans & G.B. Van Huyssteen, "Classification of noun-noun compound semantics in Dutch and Afrikaans," in *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa*, Pretoria, South Africa, 2012, pp. 121-125.
- [18] P. Vossen, ed. *EuroWordNet: a multilingual database with lexical semantic networks*. Boston, MA: Kluwer Academic, 1998.
- [19] Z. Wu, and M. Palmer, "Verb semantics and lexical selection," in *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, 1994, pp. 133-138.