Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity

Lonneke van der Plas & Jörg Tiedemann
Alfa-Informatica
University of Groningen
P.O. Box 716
9700 AS Groningen
The Netherlands
{vdplas,tiedeman}@let.rug.nl

Abstract

There have been many proposals to extract semantically related words using measures of distributional similarity, but these typically are not able to distinguish between synonyms and other types of semantically related words such as antonyms, (co)hyponyms and hypernyms. We present a method based on automatic word alignment of parallel corpora consisting of documents translated into multiple languages and compare our method with a monolingual syntax-based method. The approach that uses aligned multilingual data to extract synonyms shows much higher precision and recall scores for the task of synonym extraction than the monolingual syntax-based approach.

1 Introduction

People use multiple ways to express the same idea. These alternative ways of conveying the same information in different ways are referred to by the term *paraphrase* and in the case of single words sharing the same meaning we speak of *synonyms*. Identification of synonyms is critical for many NLP tasks. In information retrieval the information that people ask for with a set of words may be found in in a text snippet that comprises a completely different set of words. In this paper we report on our findings trying to automatically acquire synonyms for Dutch using two different resources, a large monolingual corpus and a multilingual parallel corpus including 11 languages.

A common approach to the automatic extraction of semantically related words is to use distributional similarity. The basic idea behind this is that similar words share similar contexts. Systems based on distributional similarity provide ranked lists of semantically related words according to the similarity of their contexts. Synonyms are expected to be among the highest ranks followed by (co)hyponyms and hypernyms, since the highest degree of semantic relatedness next to identity is synonymy.

However, this is not always the case. Several researchers (Curran and Moens (2002), Lin (1998), van der Plas and Bouma (2005)) have used large monolingual corpora to extract distributionally similar words. They use grammatical relations[1] to determine the context of a target word. We will refer to such systems as *monolingual syntax-based* systems. These systems have proven to be quite successful at finding semantically related words. However, they do not make a clear distinction between synonyms on the one hand and related words such as antonyms, (co)hyponyms, hypernyms etc. on the other hand.

In this paper we have defined context in a multilingual setting. In particular, translations of a word into other languages found in parallel corpora are seen as the (translational) context of that word. We assume that words that share translational contexts are semantically related. Hence, relatedness of words is measured using distributional similarity in the same way as in the monolingual case but with a different type of context. Finding translations in parallel data can be approx-

---

[1]One can define the context of a word in a non-syntactic monolingual way, that is as the document in which it occurs or the n words surrounding it. From experiments we have done and also building on the observations made by other researchers (Kilgarriff and Yallop, 2000) we can state that this approach generates a type of semantic similarity that is of a *looser* kind, an associative kind,for example *doctor* and *disease*. These words are typically not good candidates for synonymy.

imated by automatic word alignment. We will refer to this approach as *multilingual alignment-based* approaches. We expect that these translations will give us synonyms and less semantically related words, because translations typically do not expand to hypernyms, nor (co)hyponyms, nor antonyms. The word *apple* is typically not translated with a word for *fruit* nor *pear*, and neither is *good* translated with a word for *bad*.

In this paper we use both monolingual syntax-based approaches and multilingual alignment-based approaches and compare their performance when using the same similarity measures and evaluation set.

2 Related Work

Monolingual syntax-based distributional similarity is used in many proposals to find semantically related words (Curran and Moens (2002), Lin (1998), van der Plas and Bouma (2005)).

Several authors have used a monolingual parallel corpus to find paraphrases (Ibrahim et al. (2003), Barzilay and McKeown (2001)). However, bilingual parallel corpora have mostly been used for tasks related to word sense disambiguation such as target word selection (Dagan et al., 1991) and separation of senses (Dyvik, 1998). The latter work derives relations such as synonymy and hyponymy from the separated senses by applying the method of semantic mirrors.

Turney (2001) reports on an PMI and IR driven approach that acquires data by querying a Web search engine. He evaluates on the TOEFL test in which the system has to select the synonym among 4 candidates.

Lin et al. (2003) try to tackle the problem of identifying synonyms among distributionally related words in two ways: Firstly, by looking at the overlap in translations of semantically similar words in multiple bilingual dictionaries. Secondly, by looking at patterns specifically designed to filter out antonyms. They evaluate on a set of 80 synonyms and 80 antonyms from a thesaurus.

Wu and Zhou's (2003) paper is most closely related to our study. They report an experiment on synonym extraction using bilingual resources (an English-Chinese dictionary and corpus) as well as monolingual resources (an English dictionary and corpus). Their monolingual corpus-based approach is very similar to our monolingual corpus-based approach. The bilingual approach is different from ours in several aspects. Firstly, they do not take the corpus as the starting point to retrieve word alignments, they use the bilingual dictionary to retrieve multiple translations for each target word. The corpus is only employed to assign probabilities to the translations found in the dictionary. Secondly, the authors use a parallel corpus that is bilingual whereas we use a multilingual corpus containing 11 languages in total. The authors show that the bilingual method outperforms the monolingual methods. However a combination of different methods leads to the best performance.

3 Methodology

3.1 Measuring Distributional Similarity

An increasingly popular method for acquiring semantically similar words is to extract distributionally similar words from large corpora. The underlying assumption of this approach is that semantically similar words are used in similar contexts. The contexts a given word is found in, be it a syntactic context or an alignment context, are used as the features in the vector for the given word, the so-called *context vector*. The vector contains frequency counts for each feature, i.e., the multiple contexts the word is found in.

Context vectors are compared with each other in order to calculate the distributional similarity between words. Several measures have been proposed. Curran and Moens (2002) report on a large-scale evaluation experiment, where they evaluated the performance of various commonly used methods. Van der Plas and Bouma (2005) present a similar experiment for Dutch, in which they tested most of the best performing measures according to Curran and Moens (2002). Pointwise Mutual Information (I) and *Dice*† performed best in their experiments. *Dice* is a well-known combinatorial measure that computes the ratio between the size of the intersection of two feature sets and the sum of the sizes of the individual feature sets. *Dice*† is a measure that incorporates weighted frequency counts.

$$Dice\dagger = \frac{2\sum_f min(I(W_1, f), I(W_2, f))}{\sum_f I(W_1, f) + I(W_2, f)}$$

,where f is the feature

W_1 and W_2 are the two words that are being compared,

and I is a weight assigned to the frequency counts.

3.2 Weighting

We will now explain why we use weighted frequencies and which formula we use for weighting.

The information value of a cell in a word vector (which lists how often a word occurred in a specific context) is not equal for all cells. We will explain this using an example from monolingual syntax-based distributional similarity. A large number of nouns can occur as the subject of the verb *have*, for instance, whereas only a few nouns may occur as the object of *squeeze*. Intuitively, the fact that two nouns both occur as subject of *have* tells us less about their semantic similarity than the fact that two nouns both occur as object of *squeeze*. To account for this intuition, the frequency of occurrence in a vector can be replaced by a weighted score. The weighted score is an indication of the amount of information carried by that particular combination of a noun and its feature.

We believe that this type of weighting is beneficial for calculating similarity between word alignment vectors as well. Word alignments that are shared by many different words are most probably mismatches.

For this experiment we used Pointwise Mutual Information (I) (Church and Hanks, 1989).

$$I(W, f) = log \frac{P(W,f)}{P(W)P(f)}$$

,where W is the target word

P(W) is the probability of seeing the word

P(f) is the probability of seeing the feature

P(W,f) is the probability of seeing the word and the feature together.

3.3 Word Alignment

The multilingual approach we are proposing relies on automatic word alignment of parallel corpora from Dutch to one or more target languages. This alignment is the basic input for the extraction of the *alignment context* as described in section 5.2.2. The alignment context is then used for measuring distributional similarity as introduced above.

For the word alignment, we apply standard techniques derived from statistical machine translation using the well-known IBM alignment models (Brown et al., 1993) implemented in the open-source tool GIZA++ (Och, 2003). These models can be used to find links between words in a source language and a target language given sentence aligned parallel corpora. We applied stan-dard settings of the GIZA++ system without any optimisation for our particular input. We also used plain text only, i.e. we did not apply further pre-processing except tokenisation and sentence splitting. Additional linguistic processing such as lemmatisation and multi-word unit detection might help to improve the alignment but this is not part of the present study.

The alignment models produced are asymmetric and several heuristics exist to combine directional word alignments to improve alignment accuracy. We believe, that precision is more crucial than recall in our approach and, therefore, we apply a very strict heuristics namely we compute the intersection of word-to-word links retrieved by GIZA++. As a result we obtain partially word-aligned parallel corpora from which translational context vectors are built (see section 5.2.2). Note, that the intersection heuristics allows one-to-one word links only. This is reasonable for the Dutch part as we are only interested in single words and their synonyms. However, the distributional context of these words defined by their alignments is strongly influenced by this heuristics. Problems caused by this procedure will be discussed in detail in section 7 of our experiments.

4 Evaluation Framework

In the following, we describe the data used and measures applied.

The evaluation method that is most suitable for testing with multiple settings is one that uses an available resource for synonyms as a gold standard. In our experiments we apply automatic evaluation using an existing hand-crafted synonym database, Dutch EuroWordnet (EWN, Vossen (1998)).

In EWN, one synset consists of several synonyms which represent a single sense. Polysemous words occur in several synsets. We have combined for each target word the EWN synsets in which it occurs. Hence, our gold standard consists of a list of all nouns found in EWN and their corresponding synonyms extracted by taking the union of all synsets for each word. Precision is then calculated as the percentage of candidate synonyms that are truly synonyms according to our gold standard. Recall is the percentage of the synonyms according to EWN that are indeed found by the system. We have extracted randomly from all synsets in EWN 1000 words with a frequency

above 4 for which the systems under comparison produce output.

The drawback of using such a resource is that coverage is often a problem. Not all words that our system proposes as synonyms can be found in Dutch EWN. Words that are not found in EWN are discarded.[2]. Moreover, EWN's synsets are not exhaustive. After looking at the output of our best performing system we were under the impression that many correct synonyms selected by our system were classified as incorrect by EWN. For this reason we decided to run a human evaluation over a sample of 100 candidate synonyms classified as incorrect by EWN.

5 Experimental Setup

In this section we will describe results from the two synonym extraction approaches based on distributional similarity: one using syntactic context and one using translational context based on word alignment and the combination of both. For both approaches, we used a cutoff n for each row in our word-by-context matrix. A word is discarded if the row marginal is less than n. This means that each word should be found in any context at least n times else it will be discarded. We refer to this by the term *minimum row frequency*. The cutoff is used to make the feature space manageable and to reduce noise in the data. [3]

5.1 Distributional Similarity Based on Syntactic Relations

This section contains the description of the synonym extraction approach based on distributional similarity and syntactic relations. Feature vectors for this approach are constructed from syntactically parsed monolingual corpora. Below we describe the data and resources used, the nature of the context applied and the results of the synonym extraction task.

5.1.1 Data and Resources

As our data we used the Dutch CLEF QA corpus, which consists of 78 million words of Dutch

subject-verb	*cat_eat*
verb-object	*feed_cat*
adjective-noun	*black_cat*
coordination	*cat_dog*
apposition	*cat_Garfield*
prep. complement	*go+to_work*

Table 1: Types of dependency relations extracted

grammatical relation	# pairs
subject	507K
object	240K
adjective	289K
coordination	400 K
apposition	109K
prep. complement	84K
total	1629K

Table 2: Number of word-syntactic-relation pairs (types) per dependency relation with frequency > 1.

newspaper text (Algemeen Dagblad and NRC Handelsblad 1994/1995). The corpus was parsed automatically using the Alpino parser (van der Beek et al., 2002; Malouf and van Noord, 2004). The result of parsing a sentence is a dependency graph according to the guidelines of the Corpus of Spoken Dutch (Moortgat et al., 2000).

5.1.2 Syntactic Context

We have used several grammatical relations: subect, object, adjective, coordination, apposition and prepositional complement. Examples are given in table 1. Details on the extraction can be found in van der Plas and Bouma (2005). The number of pairs (types) consisting of a word and a syntactic relation found are given in table 2. We have discarded pairs that occur less than 2 times.

5.2 Distributional Similarity Based on Word Alignment

The alignment approach to synonym extraction is based on automatic word alignment. Context vectors are built from the alignments found in a parallel corpus. Each aligned word type is a feature in the vector of the target word under consideration. The alignment frequencies are used for weighting the features and for applying the frequency cutoff. In the following section we describe the data and resources used in our experiments and finally the results of this approach.

---

[2]Note that we use the part of EWN that contains only nouns

[3]We have determined the optimum in F-score for the alignment-based method, the syntax-based method and the combination independently by using a development set of 1000 words that has no overlap with the test set used in evaluation. The minimum row frequency was set to 2 for all alignment-based methods. It was set to 46 for the syntax-based method and the combination of the two methods.

5.2.1 Data and Resources

Measures of distributional similarity usually require large amounts of data. For the alignment method we need a parallel corpus of reasonable size with Dutch either as source or as target language. Furthermore, we would like to experiment with various languages aligned to Dutch. The freely available Europarl corpus (Koehn, 2003) includes 11 languages in parallel, it is sentence aligned, and it is of reasonable size. Thus, for acquiring Dutch synonyms we have 10 language pairs with Dutch as the source language. The Dutch part includes about 29 million tokens in about 1.2 million sentences. The entire corpus is sentence aligned (Tiedemann and Nygaard, 2004) which is a requirement for the automatic word alignment described below.

5.2.2 Alignment Context

Context vectors are populated with the links to words in other languages extracted from automatic word alignment. We applied GIZA++ and the intersection heuristics as explained in section . From the word aligned corpora we extracted *word type links*, pairs of source and target words with their alignment frequency attached. Each aligned target word type is a feature in the (translational) context of the source word under consideration.

Note that we rely entirely on automatic processing of our data. Thus, results from the automatic word alignments include errors and their precision and recall is very different for the various language pairs. However, we did not assess the quality of the alignment itself which would be beyond the scope of this paper.

As mentioned earlier, we did not include any linguistic pre-processing prior to the word alignment. However, we post-processed the alignment results in various ways. We applied a simple lemmatizer to the list of bilingual word type links in order to 1) reduce data sparseness, and 2) to facilitate our evaluation based on comparing our results to existing synonym databases. For this we used two resources: CELEX – a linguistically annotated dictionary of English, Dutch and German (Baayen et al., 1993), and the Dutch snowball stemmer implementing a suffix stripping algorithm based on the Porter stemmer. Note that lemmatization is only done for Dutch. Furthermore, we removed word type links that include non-alphabetic characters to focus our investigations on 'real words'. In order to reduce alignment

noise, we also applied a frequency threshold to remove alignments that occur only once. Finally, we restricted our study to Dutch nouns. Hence, we extracted word type links for all words tagged as noun in CELEX. We also included words which are not found at all in CELEX assuming that most of them will be productive noun constructions.

From the remaining word type links we populated the context vectors as described earlier. Table 3 shows the number of context elements extracted in this manner for each language pair considered from the Europarl corpus[4]

	#word-transl. pairs		#word-transl. pairs
DA	104K	FR	90K
DE	133K	IT	96K
EL	60K	PT	86K
EN	119K	SV	97K
ES	119K	ALL	994K
FI	89K		

Table 3: Number of word-translation pairs for different languages with alignment frequency > 1

6 Results and Discussion

Table 4 shows the precision recall en F-score for the different methods. The first 10 rows refer to the results for all language pairs individually. The 11th row corresponds to the setting in which all alignments for all languages are combined. The penultimate row shows results for the syntax-based method and the last row the combination of the syntax-based and alignment-based method.

Judging from the precision, recall and F-score in table 4 Swedish is the best performing language for Dutch synonym extraction from parallel corpora. It seems that languages that are similar to the target language, for example in word order, are good candidates for finding synonyms at high precision rates. Also the fact that Dutch and Swedish both have one-word compounds avoids mistakes that are often found with the other languages. However, judging from recall (and F-score) French is not a bad candidate either. It is possible that languages that are lexically different from the target language provide more synonyms. The fact that Finnish and Greek do not gain high scores might be due to the fact that there are only a limited amount of translational contexts (with a frequency > 1) available for these language (as is shown in table 3). The reasons are twofold.

---

[4]abbreviations taken from the ISO-639 2-letter codes

	# candidate synonyms								
	1			2			3		
	Prec	Rec	F-sc	Prec	Rec	F-sc	Prec	Rec	F-sc
DA	19.8	5.1	8.1	15.5	7.6	10.2	13.3	9.4	11.0
DE	21.2	5.4	8.6	16.1	7.9	10.6	13.1	9.3	10.9
EL	18.2	4.5	7.2	14.0	6.5	8.9	11.8	7.9	9.4
EN	19.5	5.3	8.3	14.7	7.8	10.2	12.4	9.7	10.9
ES	18.4	5.0	7.9	14.7	7.8	10.2	12.1	9.4	10.6
FI	18.0	3.9	6.5	14.3	5.6	8.1	12.1	6.5	8.5
FR	20.3	5.5	8.7	15.8	8.3	10.9	13.0	10.1	11.4
IT	18.7	4.9	7.8	14.7	7.5	9.9	12.3	9.2	10.5
PT	17.7	4.8	7.6	14.0	7.4	9.7	11.6	8.9	10.1
SV	22.3	5.6	9.0	16.4	7.9	10.7	13.3	9.3	10.9
ALL	**22.5**	**6.4**	**10.0**	**16.6**	**9.4**	**12.0**	**13.7**	**11.5**	**12.5**
SYN	8.8	2.5	3.9	6.9	4.0	5.09	5.9	5.1	5.5
COMBI	19.9	5.8	8.9	14.5	8.4	10.6	11.7	10.1	10.9

Table 4: Precision, recall and F-score (%) at increasing number of candidate synonyms

Firstly, for Greek and Finnish the Europarl corpus contains less data. Secondly, the fact that Finnish is a language that has a lot of cases for nouns, might lead to data sparseness and worse accuracy in word alignment.

The results in table 4 also show the difference in performance between the multilingual alignment-method and the syntax-based method. The monolingual alignment-based method outperforms the syntax-based method by far. The syntax-based method that does not rely on scarce multilingual resources is more portable and also in this experiment it makes use of more data. However, the low precision scores of this method are not convincing. Combining both methods does not result in better performance for finding synonyms. This is in contrast with the results reported by Wu and Zhou (2003). This might well be due to the more sophisticated method they use for combining different methods, which is a weighted combination.

The precision scores are in line with the scores reported by Wu and Zhou (2003) in a similar experiment discussed under related work. The recall we attain however is more than three times higher. These differences can be due to differences between their approach such as starting from a bilingual dictionary for acquiring the translational context versus using automatic word alignments from a large multilingual corpus directly. Furthermore, the different evaluation methods used make comparison between the two approaches difficult. They use a combination of the English Word-Net (Fellbaum, 1998) and Roget thesaurus (Roget, 1911) as a gold standard in their evaluations. It is obvious that a combination of these resources leads to larger sets of synonyms. This could explain the relatively low recall scores. It does however not explain the similar precision scores.

We conducted a human evaluation on a sample of 100 candidate synonyms proposed by our best performing system that were classified as incorrect by EWN. Ten evaluators (authors excluded) were asked to classify the pairs of words as synonyms or non-synonyms using a web form of the format *yes/no/don't know*. For 10 out of the 100 pairs all ten evaluators agreed that these were synonyms. For 37 of the 100 pairs more than half of the evaluators agreed that these were synonyms. We can conclude from this that the scores provided in our evaluations based on EWN (table 4) are too pessimistic. We believe that the actual precision scores lie 10 to 37 % higher than the 22.5 % reported in table 4. Over and above, this indicates that we are able to extract automatically synonyms that are not yet covered by available resources.

7 Error Analysis

In table 5 some example output is given for the method combining word alignments of all 10 foreign languages as opposed to the monolingual syntax-based method. These examples illustrate the general patterns that we discovered by looking into the results for the different methods.

The first two examples show that the syntax-

	ALIGN(ALL)	SYNTAX
consensus *consensus*	eensgezindheid *consensus*	evenwicht *equilibrium*
herfst *autumn*	najaar *autumn*	winter *winter*
eind *end*	einde *end*	begin *beginning*
armoede *poverty*	armoedebestrijding *poverty reduction*	werkloosheid *unemployment*
alcohol *alcohol*	alcoholgebruik *alcohol consumption*	drank *liquor*
bes *berry*	charme *charm*	perzik *peach*
definitie *definition*	definie *define+incor.stemm.*	criterium *criterion*
verlamming *paralysis*	lam *paralysed*	verstoring *disturbance*

Table 5: Example candidate synonyms at 1st rank and their translations in italics

based method often finds semantically related words whereas the alignment-based method finds synonyms. The reasons for this are quite obvious. Synonyms are likely to receive identical translations, words that are only semantically related are not. A translator would not often translate *auto* (*car*) with *vrachtwagen* (*truck*). However, the two words are likely to show up in identical syntactic relations, such as being the object of *drive* or appearing in coordination with *motorcycle*.

Another observation that we made is that the syntax-based method often finds antonyms such as *begin* (*beginning*) for the word *einde* (*end*). Explanations for this are in line with what we said about the semantically related words: Synonyms are likely to receive identical translations, antonyms are not but they do appear in similar syntactic contexts.

Compounds pose a problem for the alignment-method. We have chosen intersection as alignment method. It is well-known that this method cannot cope very well with the alignment of compounds because it only allows one-to-one word links. Dutch uses many one-word compounds that should be linked to multi-word counterparts in other languages. However, using intersection we obtain only partially correct alignments and this causes many mistakes in the distributional similarity algorithm. We have given some examples in rows 4 and 5 of table 5.

We have used the distributional similarity score only for ranking the candidate synonyms. In some cases it seems that we should have used it to set a threshold such as in the case of *berry* and *charm*.

These two words share one translational context : the article *el* in Spanish. The distributional similarity score in such cases is often very low. We could have filtered some of these mistakes by setting a threshold.

One last observation is that the alignment-based method suffers from incorrect stemming and the lack of sufficient part-of-speech information. We have removed all context vectors that were built for a word that was registered in CELEX with a PoS-tag different from 'noun'. But some words are not found in CELEX and although they are not of the word type 'noun' their context vectors remain in our data. They are stemmed using the snowball stemmer. The candidate synonym *definie* is a corrupted verbform that is not found in CELEX. *Lam* is ambiguous between the noun reading that can be translated in English with *lamb* and the adjective *lam* which can be translated with *paralysed*. This adjective is related to the word *verlamming* (*paralysis*), but would have been removed if the word was correctly PoS-tagged.

8   Conclusions

Parallel corpora are mostly used for tasks related to WSD. This paper shows that multilingual word alignments can be applied to acquire synonyms automatically without the need for resources such as bilingual dictionaries. A comparison with a monolingual syntax-based method shows that the alignment-based method is able to extract synonyms with much greater precision and recall. A human evaluation shows that the synonyms the alignment-based method finds are often missing in EWN. This leads us to believe that the precision scores attained by using EWN as a gold standard are too pessimistic. Furthermore it is good news that we seem to be able to find synonyms that are not yet covered by existing resources.

The precision scores are still not satisfactory and we see plenty of future directions. We would like to use linguistic processing such as PoS-tagging for word alignment to increase the accuracy of the alignment itself, to deal with compounds more effectively and to be able to filter out proposed synonyms that are of a different word class than the target word. Furthermore we would like to make use of the distributional similarity score to set a threshold that will remove a lot of errors. The last thing that remains for future work is to find a more adequate way to combine the

syntax-based and the alignment-based methods.

References

R.H. Baayen, R. Piepenbrock, and H. van Rijn. 1993. The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania,Philadelphia.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Meeting of the Association for Computational Linguistics*, pages 50–57.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–296.

K.W. Church and P. Hanks. 1989. Word association norms, mutual information and lexicography. *Proceedings of the 27th annual conference of the Association of Computational Linguistics*, pages 76–82.

J.R. Curran and M. Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, pages 59–67.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Meeting of the Association for Computational Linguistics*, pages 130–137.

Helge Dyvik. 1998. Translations as semantic mirrors. In *Proceedings of Workshop Multilinguality in the Lexicon II, ECAI 98, Brighton, UK*, pages 24–44.

C. Fellbaum. 1998. Wordnet, an electronic lexical database. MIT Press.

A. Ibrahim, B. Katz, and J. Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora.

A. Kilgarriff and C. Yallop. 2000. What's in a thesaurus? In *Proceedings of the Second Conference on Language Resource an Evaluation*, pages 1371–1379.

Philipp Koehn. 2003. Europarl: A multilingual corpus for evaluation of machine translation. unpublished draft, available from `http://people.csail.mit.edu/koehn/publications/europarl/`.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *IJCAI*, pages 1492–1493.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774.

Robert Malouf and Gertjan van Noord. 2004. Wide coverage parsing with stochastic attribute value grammars. In *IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and stati stical modeling for deep analyses*, Hainan.

Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. 2000. CGN syntactische annotatie. Internal Project Report Corpus Gesproken Nederlands, see http://lands. let.kun.nl/cgn.

Franz Josef Och. 2003. GIZA++: Training of statistical translation models. Available from `http://www.isi.edu/˜och/GIZA++.html`.

P. Roget. 1911. Thesaurus of English words and phrases.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Peter D. Turney. 2001. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, 2167:491–502.

Leonoor van der Beek, Gosse Bouma, and Gertjan van Noord. 2002. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7(4):353–374.

Lonneke van der Plas and Gosse Bouma. 2005. Syntactic contexts for finding semantically similar words. Proceedings of the Meeting of Computational Linguistics in the Netherlands (CLIN).

P. Vossen. 1998. Eurowordnet a multilingual database with lexical semantic networks.

Hua Wu and Ming Zhou. 2003. Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, Sapporo, Japan.