Automatic Acquisition of Lexico-semantic Knowledge for QA

Lonneke van der Plas
Information Science
University of Groningen
Postbus 716, 9700 AS Groningen
vdplas@let.rug.nl

Gosse Bouma
Information Science
University of Groningen
Postbus 716, 9700 AS Groningen
gosse@let.rug.nl

Abstract

We present an experiment for finding semantically similar words on the basis of a parsed corpus of Dutch text and show that the acquired information correlates with relations found in Dutch EuroWordNet. Next, we demonstrate how the acquired knowledge can be used to boost the performance of an open-domain question answering system for Dutch. Automatically acquired lexico-semantic information is used to improve the recall of a method for extracting function relations (such as *Wim Kok is the prime minister of the Netherlands*) from corpora, and to improve the precision of our QA system on general WH-questions and definition questions.

1 Introduction

Lexico-semantic knowledge is increasingly important in NLP, especially for applications such as Word Sense Disambiguation, Information Extraction and Question Answering (QA). Although the coverage of handmade resources such as Wordnet (Fellbaum, 1998) in general is impressive, coverage problems remain for applications involving specific domains or involving languages other than English.

We are interested in using lexico-semantic knowledge in an open-domain question answering system for Dutch. Obtaining such knowledge from existing resources is possible, but only to a certain extent. The most important resource for our research is the Dutch part of EuroWordNet (Vossen, 1998), but its size is only half of that of the English WordNet. Many of the lexical items used in the CLEF QA[1] corpora for Dutch, for instance, cannot be found in EuroWordNet. In addition, information about the classes to which named entities belong (i.e. *Narvik* ISA *harbour*) has been shown to be useful for QA. However, such information is typically absent from handbuilt resources. For these reasons, we are interested in methods for acquiring lexico-semantic knowledge automatically from text corpora.

The remainder of the paper is organized as follows. In the next section we briefly describe the question types for which we want to use lexico-semantic knowledge and in section 3, we describe related work. In section 4 we describe our approach to finding distributionally similar words. Sections 5 and 6 decribe how the acquired knowledge is used for improving the performance of our QA system on specific question types, i.e. questions asking for the name of persons which have a specific function in an organization (i.e. *Who is the secretary general of the UN?*), general WH-questions, and definition questions. We report the results of an evaluation on CLEF 2005 data in section 7. Section 8 contains our conclusions and suggestions for future research.

2 Lexico-semantic Knowledge for QA

We will now briefly describe the three question types whose performance we hope to improve using automatically acquired lexical knowledge.

[1] http://clef-qa.itc.it/

Often questions are asked about the function of a particular person:

Who is the chair of Unilever?

Off-line methods (Fleischman et al., 2003) can be used to improve the performance of the system on such questions. In off-line QA plausible answers to highly likely questions are extracted before the actual question has been asked.

Bouma et al. (2005a) describe how syntactic patterns are used to extract answers for frequently occurring question types. The following syntactic pattern could serve to extract ⟨*Person,Role,Organization*⟩-tuples from the corpus:

$$name(PER) \xleftarrow{\text{app}} noun \xrightarrow{\text{mod}} name(ORG)$$

Here, the *name(PER)* constituent provides the *Person* argument of the relation, the *noun* provides the role, and the *name(ORG)*-constituent provides the name of the *Organization*. An important source of noise in applying this pattern to the parsed corpus are cases where the *noun* is not indicating a role or a function:

colleague Henk ten Cate of Go Ahead

Here, the noun *colleague* does not represent a role within the organization *Go Ahead*.

To remedy this problem, we collected a list of nouns denoting functions or roles from Dutch EWN, and restricted the search pattern to nouns occurring in this list:

$$name(PER) \xleftarrow{\text{app}} function \xrightarrow{\text{mod}} name(ORG)$$

While this helps to improve precision, it also hurts recall, as many valid function words present in the corpus are not present in EWN. In section 5, we will report on an experiment where we expanded the list of function words extracted from EWN semi-automatically with distributionally similar words found in the corpus.

A second question type where the use of lexical knowledge is potentially useful are general WH-questions such as:

Which vulcano errupted in june 1991?

A QA system may find various named entities (such as *Fillipines* and *Pinatubo*) as potential answers to the question. Knowing that *Pinatubo* ISA *vulcano* can help to identify the correct answer. Information about named entities is typically absent in hand-made lexical resources. In section 6, we describe a method for acquiring such information automatically from a parsed corpus.

A final question type where lexical knowledge is useful are definition questions:

Who is Javier Solana?

For CLEF 2005, definition questions were restricted to persons and organizations and answers should provide "some fundamental information" to users who know nothing about the named entity. Determining which information should be used to provide an answer to such questions in general is hard. We tried an approach were we used automatically acquired ISA-labels for named entities to find an appropriate category which needs to be included in the answer. In section 6, we describe how this information can be used to find answers to definition questions.

3 Related Work

Syntactic relations have been shown to provide information which can be used to acquire clusters of semantically similar words automatically (Lin, 1998a). As we have a fully parsed version of the Dutch CLEF QA corpus (78 million words, 4.1 million sentences) at our disposal, we were interested in applying this method to Dutch. In particular, we followed the strategy of Curran and Moens (2002) which evaluates various similarity measures and weight functions against various thesauri (MacQuarie (Bernard, 1990), Moby (Ward, 1996) and Roget (Roget, 1911)). We implemented most of the best performing similarity measures and weights according to the evaluation of Curran and Moens (2002) and evaluated their performance against Dutch EuroWordNet. Some results are given in section 4.

Automatically acquired clusters of semantically similar words can be used to extend or enrich existing ontological resources. Alfonseca and Manandhar (2002), for instance,

describe a method for expanding WordNet automatically. New concepts are placed in the WordNet hierarchy according to their distributional similarity to words that are already in the hierarchy. Their algorithm performs a top-down search and stops at the synset that is most similar to the new concept. In section 5, we are using a similar technique to expand the class of function words obtained from EuroWordNet.

Pasca (2004) and Pantel and Ravichandran (2004) present methods for acquiring class labels for instances (categorised named entities) from unstructured text. Pasca (2004) applies lexico-syntactic extraction patterns based on Part-of-Speech tags. Patterns were hand-built initially, and extended automatically by scanning the corpus for the pairs of named entities and classes found with the initial patterns. Patterns which occur frequently in matching sentences can be added as additional extraction patterns. Pasca (2004) applies this information to websearch for example for processing list-type queries. For example, *SAS, SPSS, Minitab* and *BMDP* are returned in addition to the top documents for the query *statistical packages*. Pantel and Ravichandran (2004) propose an algorithm that takes a list of semantic classes in the form of clusters of words as input. Labels for these clusters are found by looking at four lexico-syntactic relationships *apposition* (*ayatollah Khomeini*), *nominal subject* (*Khomeini is an ayatollah*), *such as* (*Ayatollahs such as Khomeini*), and *like* (*Ayatollahs like Khomeini*). Apart from judging the quality of their results manually, they conducted two QA experiments: answering definition questions and performing QA information retrieval (IR). They show that both tasks benefit from the use of automatically acquired class labels.

4 Extracting semantically similar words

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 context of a word W may be defined as the document in which W occurs or the n words surrounding W (n-grams, bag of words). Alternatively, the context may be defined syntactically. In that case, the words with which the target word is in a specific syntactic relation form the context of that word. Approaches which do not use syntax tend to find more associative relations between words (i.e. between *patient* and *hospital*), whereas approaches using syntactic context tend to find concepts belonging to the same class (i.e. *doctor* and *surgeon*). As we are ultimately interested in extending the coverage of a resource such as Dutch EuroWordNet, we focussed on the second approach.

Most research has been done using a limited number of syntactic relations ((Lee, 1999), (Weeds, 2003)). However, (Lin, 1998a) shows that a system which uses a range of grammatical relations outperforms Hindle's (1990) results that were based on using information from just the subject and object relation. Apart from the subject and object relation we have used several other grammatical relations: adjective, coordination, apposition and prepositional complement. Examples are given in table 1.

4.1 Data collection

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

From these dependency graphs, we extracted tuples consisting of the (non-pronominal) head of an NP (either a common noun or a proper name), the dependency relation, and either (1) the head of the dependency relation (for the object, subject, and apposition relation), (2) the head plus a preposition (for NPs occurring inside PPs which are prepositional complements), (3) the head of the dependent (for the adjective and apposition relation) or (4) the head of the other elements of a coordination (for the coordination relation). Examples are given in table 1. The number of tuples and the number of non-identical ⟨`Noun,Relation,OtherWord`⟩ triples (types) found are given in table 2.Note that

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	tuples	types
subject	5.639.140	2.122.107
adjective	3.262.403	1.040.785
object	2642.356	993.913
coordination	965.296	2.465.098
prep. complement	770.631	389.139
apposition	526.337	602.970

Table 2. Number of tuples and non-identical dependency triples (types) extracted per dependency relation.

a single coordination can give rise to various dependency triples, as from a single coordination like*bier, wijn, en noten* (*beer, wine, and nuts*) we extract the triples ⟨*bier, coord, wijn*⟩, ⟨*bier, coord, noten*⟩, ⟨*wijn, coord, bier*⟩, ⟨*wijn, coord, noten*⟩, ⟨*noten, coord, bier*⟩, and ⟨*noten, coord, wijn*⟩. Similarly, from the apposition *premier Kok* we extract both ⟨*premier, hd_app, Kok*⟩ and ⟨*Kok, app, premier*⟩.

For each noun that was seen at least 10 times in any dependency relation, we built a vector. After applying this cut-off, vectors are present for 83.479 nouns.

4.2 Similarity measures and weights

Various vector-based methods can be used to compute the distributional similarity between words. 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 (MI) and *Dice*† performed best in the experiments. We will now explain this weight and similarity measure in further detail.

The information value of a cell in a word vector (which lists how often a word occurred in a specific grammatical relation to a specific word) is not equal for all cells. A large number of nouns can occur as the subject of the verb *hebben* (*have*), for instance, whereas only a few nouns may occur as the object of *uitpersen* (*squeeze*). Intuitively, the fact that two nouns both occur as subject of *hebben* tells us less about their semantic similarity than the fact that two nouns both occur as object of *uitpersen*. 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 (the grammatical relation, and the word heading the grammatical relation). For this experiment we used Pointwise Mutual Information (MI) (Church and Hanks, 1989).

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

To compute the similarity of two word vectors, we used a variant of the Dice-measure, which Curran and Moens (2002) refer to as *Dice*†:

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

4.3 Performance

The Dutch version of the multilingual resource EuroWordNet (EWN) (Vossen, 1998) was used for evaluation. We randomly selected 1000 target words from Dutch EWN with a frequency of more than 10, according to the frequency information present in Dutch EWN. For each word we collected its 100 most similar words (nearest neighbours) according to the system under evaluation, and for each pair of words (target word + one of the most similar words) we calculated the semantic similarity according to Dutch EWN. A system scores well if the nearest neighbours found by the system also have a high semantic similarity according to EWN.

Lin (1998b) evaluates a number of measures for computing WordNet similarity. From the measures which are defined in terms of IS-A relations only, the Wu and Palmer (1994) measure correlated best with human judgements. The

	EWN Similarity at					
hline k=	1	5	10	20	50	100
system	.60	.54	.52	.49	.46	.44

Table 3. Average EWN similarity at k candidates when combining dependency relations based on Dice†+ MI.

Wu/Palmer measure for computing the semantic similarity between two words W1 and W2 in a word net, whose most-specific common ancestor is W3, is defined as follows:

$$Sim = \frac{2(D3)}{D1 + D2 + 2(D3)}$$

where, D1 (D2) is the distance from W1 (W2) to the lowest common ancestor of W1 and W2, W3. D3 is the distance of that ancestor to the root node.

Table 3 reports average EWN similarity for the 1, 5, 10, 20, 50, and 100 most similar words for the 1000 words in ours test set. If a word is ambiguous according to EWN (i.e. is a member of several synsets), the highest similarity score is used. The EWN similarity of a set of word pairs is defined as the average of the similarity between the pairs. The baseline for this task is 0.26, which is the score obtained by picking 100 random words as nearest neighbours of a given target word. van der Plas and Bouma (2005) show that the system using data obtained from all syntactic relations outperforms systems using only a subset of the syntactic relations. Furthermore, they show that Dice†+ MI outperforms various other combinations of weight functions and similarity measures.

5 Using automatically acquired role and function words

In section 2, we explained that for QA we are interested in extracting, off-line, all instances of the following pattern in our corpus:

$$\textit{name(PER)} \xleftarrow{\text{app}} \textit{function} \xrightarrow{\text{mod}} \textit{name(ORG)}$$

To obtain a list of words describing a role or function, we extracted from Dutch EWN all words under the node *leider* (*leader*) (255 in total). The majority of hyperonyms of this node seemed to indicate function words we were interested in (i.e. it contained (the Dutch equivalents of) *king, queen, president, director, chair*, etc.), while other potential candidates (such as *beroep* (*profession*) seemed less suitable. However, the coverage of this list, when tested on a newspaper corpus, is far from complete. On the one hand, the list contains a fair amount of archaic items, while on the other hand, many functions that occur frequently in newspaper text are missing (i.e. Dutch equivalents of *banker, boss, national team coach, captain, secretary-general*, etc.).

To improve recall, we extended the list of function words obtained from EWN semi-automatically with distributionally similar words. In particular, for each of the 255 words in the EWN list, we retrieved its 100 most distributionally similar words. We gave each retrieved word a score that corresponds to its reverse rank (1st word: 100, 2nd: 99, 3rd: 98 etc.). The overall score for a word was the sum of the scores it obtained for the individual key words. Thus, words that are semantically similar to several words in the original list will obtain a higher score than words that were returned only once or twice. Words that were present already in the EWN-list were filtered.

An informal evaluation of the result learned that many false positives in the expanded list were either named entities or nouns referring to groups of people (*board, committee, ,...*). The distinction between groups and functions of individuals is hard to make on the basis of distributional data. For instance, both a *board* and a *director* can take decisions, report results, be criticized, etc. We tried to filter both proper names and groups automatically, by discarding noun stems that start with a capital, and noun stems which are listed under the node *groep* (*group*) in EWN.

Finally, we selected the top-1000 of the filtered list, and validated it manually. The list contained 644 valid role or function nouns, which are absent in EWN. A substantial number of the errors are nouns which refer to a group but which are not listed as such in EWN.

The 644 valid nouns were merged with the original EWN list, to form a list of 899 function or role nouns. Next, the off-line extraction process was executed using both the original EWN

EWN		EWN+	
tuples	unique	tuples	unique
34191	16530	77028	46589

Table 4. Coverage of function table with (EWN+) and without (EWN) expansion.

list and the expanded list. The effect on recall is illustrated in table 4. The number of extracted tuples increases with 125%, while the number of unique tuples increases with 181%. The effect of this increase on the performance of our QA system is described in section 7.

6 Using automatically acquired instances

Both Pasca (2004) and Pantel and Ravichandran (2004) describe methods for acquiring labels for named entities from large text corpora and evaluate the results in the context of web search and question answering. Pantel and Ravichandran (2004) use the apposition relation to find potential labels for named entities. As we already had extracted all appositions from the CLEF corpus as part of the vector-based method for finding semantically similar words, we decided to use this information for two other QA tasks as well.

As can be seen in table 2, we extracted 602K apposition relations (301K regardless of direction), from a total of over 526K appositions tuples found in the corpus. This database contains, for instance, 112 appositions with names of *ferry boats* (*Estonia, Anna Maria Lauro, Sally Star* etc.) and no less than 2951 appositions with names of national team coaches (*Bobby Robson, Jack Charlton, Menotti, Berti Vogts* etc.). The class labels extracted for each named entity may contain a certain amount of noise. However, by focussing on the most frequent label for a named entity, most of the noise can be discarded. For instance, *Guus Hiddink* occurs 197 times in the extracted apposition tuples, 170 times as *bondscoach* (*national team chef*), and not more than 5 times with various other labels (*coach, colleague, guest, newcomer, ...*). Regarding the ambiguity of the classified named entities we can say that on average a named entity has 1.7 labels. The distribution is skewed: 80 % has only 1 label and for example the most ambiguous named entity, *the Netherlands*, has 515 labels in total.

We used the extracted class labels to improve the performance of our QA system on general WH-questions such as:

> Which ferry sank southeast of the island Utö?

Question analysis and classification tells us that this is a question of type `which(ferry)`. Candidate answers that are selected by our system are: *Tallinn, Estonia, Raimo Tiilikainen* etc. The QA system uses various strategies to rank potential answers, i.e. the score assigned to the passage by Information Retrieval(IR), the presence of named entities from the question in the sentence in which the answer is found, the syntactic similarity between question and answer sentence, the frequency of the answer in the set of potential answers etc. Still, selecting the correct named entity for answers to general WH-questions poses considerable problems for our system.

To improve the performance of the system on these questions, we incorporated an additional strategy for selecting the correct answer. Potential answers which have been assigned the class corresponding to the question stem (i.e. `ferry` in this case) are ranked higher than potential answers for which this class label cannot be found in the database of ISA-relations. Since *Estonia* is the only potential answer which ISA ferry, according to our database, this answer is selected. Note that in answering WH-questions we do not select only the most frequent label assigned to a named entity, but simply check whether the named entity occurs at least once with the appropriate class label.

A second question type where the acquired class labels are relevant are definition question. The CLEF 2005 QA test set contains no less than 60 questions of the form:

> What is Sabena?

The named entity *Sabena* occurs frequently in the corpus, but often with class labels assigned to it, which are not suitable for inclusion in a definition (*possibility, partner, company,....* By focussing on the most frequent class label assigned to a named entity (*airline company* in this case),

a more appropriate label for a definition can be found. Frequency is important but often the class label by itself is not sufficient for an adequate definition. Therefore we expand the class label with modifiers which typically need to be included in a definition.

More in particular, our strategy for answering definition questions consisted of two phases:

- Phase 1: The most frequent class found for a named entity is taken.
- Phase 2: The sentences which mention the named entity and the class are selected, and searched for additional information which might be relevant. Snippets of information that are in a adjectival relation or a prepositional complement to the class label are selected.

For the example above, our system produces *Belgian airline company* as answer.

However, deciding beforehand what information is relevant is not trivial. As explained we decided to only expand the label with adjectival and PP modifiers that are adjacent to the class label in the corresponding sentence. This is the reason for a number of answers being inexact. Given the constituent *the museum Hermitage in St Petersburg*, this strategy fails to include *in St Petersburg*, for instance. We did not include relative clause modifiers, as these tend to contain information which is not appropriate for a definition. However, for the question, *Who is is Iqbal Masih*, this leads the system to answer *twelve year old boy*, extracted from the constituent *twelve year old boy, who fought against child labour and was shot sunday in hist home town Muritke*. Here, at least the first conjunct of the relative clause should have been included. Similarly, we did not include purpose clauses, which leads the system to respond *large scale American attempt* to the question *what was the Manhattan project*, instead of *large scale American attempt to develop the first (that is, before the Germans) atomic bomb*.

7 Evaluation

We compared the performance of two versions of our QA system on the Dutch questions from CLEF 2005. As no official results for CLEF 2005 were known to us at the time of the experiment, [2] answers were judged for correctness by ourselves and two additional project members. Answers were judged correct if at least three of the four judges considered them correct. Note that in CLEF, systems must return only a single, exact, answer.

In table 5 the performance of the baseline and improved system is shown. In the first column the question type is given (question types not relevant for this paper are left out). In the second and fourth column the number of questions classified as being of the corresponding question type is shown. In colums 3 and 5 the corresponding CLEF score is given.

The baseline of our QA system, was the Joost QA system, without a special question type for function questions, and without access to ISA-relations. The baseline treats function questions as person questions, i.e. as questions which require a named entity of type *person* as an answer. General WH-questions and definition questions are answered by selecting the most highly ranked answer from the list of relevant paragraphs returned by the IR component. Answers to definition questions are basically selected by means of the same strategy as described for the improved system above, except that answers must now be selected from the documents returned by IR, rather than from sentences known to contain a relevant class label.

The improved system makes use of the question type *function* and the related table in which information about functions is stored. Furthermore it uses ISA-relations in answering general WH questions and definition questions.

The overall effect of these additions is an improvement in (estimated) CLEF score of 8% and an error reduction of 16%.

Adding a question class for functions, and a related table with (off-line extracted) answers to such questions has the effect that 19 person questions and one general WH-question in the baseline system are now classified as function questions. The effect on accuracy of this change seems small (as person questions are already answered relatively well), but is nevertheless positive. Of the 20 questions that are classified as function

[2] see Bouma et al. (2005b) for official results

question_type	baseline		improved	
	# q	score	# q	score
WH-questions	36	0.31	35	0.46
definition	60	0.53	60	0.68
person	26	0.69	7	0.71
function	0	0.00	20	0.75
...	...	...	...	...
total	**200**	**0.49**	**200**	**0.57**

Table 5. Overall performance of the baseline and improved QA system on the CLEF 2005 Dutch QA test set.

questions in the improved system, 4 involve the question stems *weduwe* (*widdow*), *adviseur* (*advisor*), *secretaris-generaal* (*secretary-general*) and *vriendin* (*girl friend*), which were present in our extended list of function nouns only.

Adding ISA-relations as an additional knowledge source for answering WH-questions improves the CLEF score of 36 WH-questions with 15 % and gives an error reduction of 22%. Using the same information to provide answers to definition questions improves the CLEF score on 60 definition with almost 15%, which is an error reduction of 22%.

8 Conclusions and future work

We have demonstrated that lexico-semantic knowledge can be acquired from syntactically parsed corpora, and that the inclusion of such knowledge in a QA system has a positive effect on the overall performance of the QA system. Firstly, the use of off-line techniques in general has a positive effect on the accuracy of QA. Here, we have demonstrated that the resources required to do off-line extraction accurately can be acquired semi-automatically by expanding a given list of relevant function words. Secondly, the performance of the system on general WH-questions and definition questions was shown to improve considerably if it has access to automatically acquired class labels.

The research reported here can be extended in several ways. For instance, while we used a considerable number of grammatical relations for finding semantically similar words, we did not use predicative complements. Sentences containing such a complement (i.e. *Garfield is a cat*) do seem to provide useful information for learning semantic similarity. In addition, this relation may be used to expand the number categorised named entities.

Alternative ways of exploiting the class labels in QA can be explored as well. Pantel and Ravichandran (2004), for instance, use class labels to index the document collection. I.e. every paragraph which mentions a named entity known to be a *ferry*, is labeled with this class as well. This strategy allows the IR component to make use of class information. Pantel and Ravichandran (2004) show that this improves the precision of IR considerably. In future work, we would like to explore this possibility as well.

Acknowledgements

This research was carried out as part of the research program for *Interactive Multimedia Information Extraction*, IMIX, financed by NWO, the Dutch Organisation for Scientific Research.

References

E. Alfonseca and S. Manandhar. 2002. Improving an ontology refinement method with hyponymy patterns. In *Proceedings of the international conference on Language Resources and Evaluation (LREC-2002)*.

J.R.L. Bernard. 1990. The Macquairie encyclopedic thesaurus. The Macquairie Library, Sydney, Australia.

Gosse Bouma, Jori Mur, and Gertjan van Noord. 2005a. Reasoning over dependency relations for QA. In *Proceedings of the IJCAI workshop on Knowledge and Reasoning for Answering Questions (KRAQ)*, pages 15–21, Edinburgh.

Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann. 2005b. Question answering for Dutch using dependency relations. In *Proceedings of CLEF 2005*. To appear.

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.

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

Michael Fleischman, Eduard Hovy, and Abdessamad Echihabi. 2003. Offline strategies for online question answering: Answering questions before they are asked. In *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, pages 1–7, Sapporo, Japan.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of ACL-90*, pages 268–275.

L. Lee. 1999. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

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

Dekang Lin. 1998b. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

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.

Patrick Pantel and Deepak Ravichandran. 2004. Automatically labeling semantic classes. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 321–328, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

M. Pasca. 2004. Acquisition of categorized named entities for web search. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management*, pages 137 – 145.

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

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. In *Proceedings of CLIN 2004*, Leiden University. To Appear.

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

G. Ward. 1996. Moby thesaurus. Moby Project.

J. Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, University of Sussex.

Z. Wu and M. Palmer. 1994. Verb semantics and lexical selection. In *The 23rd Annual Meeting of the Association for Computational Linguistics*, pages 133–138.