Extracting Dependency Frames from Existing Lexical Resources

Gosse Bouma
Alfa-informatica
Rijksuniversiteit Groningen
Postbus 716
9700 AS Groningen
The Netherlands
gosse@let.rug.nl

Abstract

One method for evaluating a wide-coverage parser involves measuring how accurately it identifies dependency relations. The construction of a grammar which outputs dependency relations requires a lexicon with detailed information on subcategorization and dependency. We discuss how such a lexicon can be constructed automatically for a wide-coverage lexicalist grammar for Dutch by extracting the relevant information from existing lexical resources. We compare the coverage of the two sources relative to each other, and the coverage of the resulting lexicon with respect to a dependency treebank.

1 Introduction

It has been observed that accurate, wide-coverage, parsing of unrestricted text requires a lexical component with detailed subcategorization frames. A lexicon that is incomplete in this respect can seriously degrade parser performance. Carroll and Briscoe (1996) observe, for instance, that for their initial system *the largest source of error on unseen input is the omission of appropriate subcategorization values for lexical items (mostly verbs).*

Carroll et al. (1998) propose an evaluation method for parsers and grammars based on *dependency relations*. Such an evaluation scheme has advantages over tree-based methods, especially for languages with a relatively free word order. However, for some dependency relations it implies that lexical subcategorization frames must be enriched with the relevant information explicitly (i.e. in order to distinguish between direct and indirect objects, etc.). We refer to such enriched subcategorization frames as *dependency frames*.

Lexical databases providing subcategorization information are rare and therefore researchers have focussed on the question of how to obtain such information automatically, from raw or annoted text. For Dutch, the tools or corpora to do automatic acquisition are not available. On the other hand, at least two lexical resources provide dependency frames. In this paper, we address the question to what extent using these lexical resources can lead to an adequate lexical component for a wide-coverage computational grammar.

Below, we introduce dependency relations as a means for syntactic annotation of corpora and evaluation and we discuss dependency and subcategorization in the Alpino grammar for Dutch. Next, we explain to what extent detailed dependency frames can be extracted from two lexical resources (CGN/Celex and Parole), how the extracted information is incorporated in the Alpino lexicon, and we provide an indication of the coverage of the resulting lexicon.

2 Dependency Relations

Dependency relations are generally seen as triples consisting of a head word, a relation label, and the head word of the dependent. For instance, for the example in (1) we might define the dependency relations in (2).

(1) chevrolet announced a new model for 1975

(2) ⟨announced, su, chevrolet⟩
⟨announced, obj1, model⟩
⟨announced, pc, for⟩
⟨model, mod, new⟩
⟨model, det, a⟩
⟨for, obj1, 1975⟩

Dependency relations of this type can be extracted from a syntactic analysis tree by identifying for each constituent what its (lexical) head is, and what the relations are between (the constituent containing) the lexical head and its sisters. In figure 1, a phrase structure tree for (1) is given, where head daughters are marked, and dependency relations are added to the non-head nodes.

A number of researchers have stressed the importance of dependency relations for evaluation and training of wide-coverage grammars. Carroll et al. (1998) argue that evaluation in terms of dependency relations avoids some of the drawbacks of tree-based evaluation methods (such as counting the number of inconsistent (crossing) brackets in parser output and annotation). Furthermore, dependency relations are relatively easy to obtain from the output of a wide range of grammars.

Accuracy of statistical parsers can be improved by lexicalization, i.e. by making parse-decisions sensitive to the lexical head of a phrase (Magerman, 1994; Collins, 1999). In addition, Collins (1999) shows that accuracy can be improved by taking into account depen-

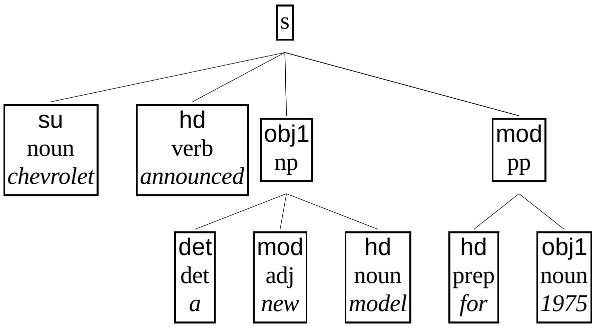

Figure 1: Phrase structure tree enriched with dependency relations.

dency relations between lexical heads. To obtain dependency relations from the Penn Treebank, which provides labelled constituents only, Collins used an automatic method where relation names are triples consisting of the category of the dependent, head word, and dominating node, respectively. Collecting statistics on the co-occurence of lexical heads in certain syntactic configurations obviously requires large amounts of parsed or annotated data. One might try to reduce this sparse data problem by abstracting from the actual noun and verb stems to (semantic) classes of nouns and verbs.

Recently, treebanks (such as the German Negra corpus (Skut et al., 1997) and the Prague Dependency Treebank (Hajicova et al., 1998)) have been constructed in which dependency relations are marked explicitly, thus allowing the same dependency relation to appear in different syntactic configurations, as well as allowing the same syntactic configuration to be labelled with different dependency relations. Skut et al. (1997) argue that for languages with a relatively free word order, dependency trees are a more natural and valuable form of annotation than phrase structure (alone).

For Dutch, there are currently no corpora available providing phrase structure or dependency relations. However, within the project *Corpus Gesproken Nederlands* (*Corpus Spoken Dutch*) (Oostdijk, 2000), guidelines have been developed for syntactic annotation, using dependency trees similar to those used for the German Negra corpus. Figure 1 illustrates the proposed annotation format.

3 Alpino

The Alpino Grammar is a wide-coverage, lexicalist, grammar for Dutch.[1] The grammar formalism is based on a fragment developed previously for use in a spoken dialogue system (van Noord et al., 1999) and supports the implementation of feature-based and constraint-based grammars. The formalism is carefully designed to allow linguistically sophisticated analyses as well as efficient and robust processing.

$$\left[\begin{array}{ll}
\multicolumn{2}{l}{\textit{verb}}\\
\text{PHON} & \boxed{1}\\
\text{DEPS} & \left\langle \begin{bmatrix} \textit{np} & \\ \text{CASE} & \textit{nom} \\ \text{DT} & \boxed{2}\end{bmatrix}, \begin{bmatrix} \textit{np} & \\ \text{CASE} & \textit{acc} \\ \text{DT} & \boxed{3}\end{bmatrix}\right\rangle\\
\text{DT} & \begin{bmatrix} \text{HD} & \begin{bmatrix}\text{POS} & \textit{verb}\\ \text{WORD} & \boxed{1}\end{bmatrix}\\ \text{SU} & \boxed{2}\\ \text{OBJ1} & \boxed{3}\end{bmatrix}
\end{array}\right]$$

$$\left[\begin{array}{ll}
\multicolumn{2}{l}{\textit{verb}}\\
\text{PHON} & \boxed{1}\\
\text{DEPS} & \left\langle \begin{bmatrix} \textit{np} & \\ \text{CASE} & \textit{nom} \\ \text{DT} & \boxed{2}\end{bmatrix}, \begin{bmatrix} \textit{np} & \\ \text{CASE} & \textit{dat} \\ \text{DT} & \boxed{3}\end{bmatrix}\right\rangle\\
\text{DT} & \begin{bmatrix} \text{HD} & \begin{bmatrix}\text{POS} & \textit{verb}\\ \text{WORD} & \boxed{1}\end{bmatrix}\\ \text{SU} & \boxed{2}\\ \text{OBJ2} & \boxed{3}\end{bmatrix}
\end{array}\right]$$

Figure 2: Schematic lexical entry for transitive verbs taling a direct object (OBJ1), and for transitive verbs taking an indirect object (OBJ2).

The grammar design is inspired by Head-driven Phrase Structure Grammar (Pollard and Sag, 1994). The grammar currently contains over 100 rules, defined in terms of a few general rule schemata, and covering the basic constructions of Dutch (including main and subordinate clauses, (indirect) questions, imperatives, relative clauses, a wide range of verbal and nominal complementation and modification patterns, and coordination). The lexicon contains definitions for various nominal types (nouns with various complementation patterns, proper names, pronouns, temporal nouns, deverbalized nouns), complementizer types, determiner types, adverb types, adjectives, and 36 verbal subcategorization types.

The formalism supports the use of recursive constraints over feature-structures (using delayed evaluation, van Noord and Bouma (1994)). This allowed us to incorporate an analysis of cross-serial dependencies based on argument-inheritance (Bouma and van Noord, 1997) and a trace-less account of extraction similar to that in Bouma et al. (2001).

4 Subcategorization and Dependency

HPSG does not represent dependency relations explicitly. As we want to use dependency trees for evaluation and annotation of corpora, a new level of representation has been added to the grammar. The attribute DT

[1] Alpino is being developed as part of the project *Algorithms for Linguistic Processing,* `www.let.rug.nl/~vannoord/alp`

dominates a dependency tree, with attributes for the lexical head and the various dependency relations. The values of these relations are dependency trees or leaf nodes consisting of a POS-tag and word only. The construction of dependency trees is driven by the lexicon. For each subcategorization type recognized in the lexical hierarchy, a mapping between elements on the list-valued feature which specifies basic subcategorization properties (DEPS) and attributes of DT is defined. Examples are given in figure 2. In rule schemata where a head combines with an element it subcategorizes for, the DT attribute can simply be shared between head daughter and mother. As there is an strong correlation between the categorial properties and the position of an element on DEPS and its dependency label, a unique mapping can be defined for almost all subcategorization types. For verbal subcategorization frames, for instance, the first element on DEPS is always is linked to the SU dependency relation, an accusative NP is always linked to OBJ1, and a verbal or clausal complement is always linked to VC (*verbal complement*). An exception is formed by PP-arguments, which can be linked to PC (*prepositional complement*) or LD (*locative or directional complement*), where the distinction between these two is primarily semantic in nature.

5 Acquisition of Dependency Frames

For lexicalist grammar formalisms, the availability of lexical resources which specify subcategorization frames is crucial. In HPSG, for instance, phrase structure schemata rely on the fact that each head contains a specification of the elements it subcategorizes for. If such specifications are missing, the grammar will wildly overgenerate.

Furthermore, to create lexical entries with dependency relations, the subcategorization information provided by the lexical database must be relatively detailed. For instance, to distinguish between a direct and indirect object, either a distinction between accusative and dative case must be made (for which there is no morphological evidence in Dutch), or the relevant dependency label must be provided explicitly. To distinguish between PP-complements with the prepositional or locative/directional complement relation, detailed semantic information or an explicit dependency label must be provided.

Lexica with subcategorization information are often not available or have very limited coverage, and therefore researchers have attempted to extract the relevant information from unannotated corpora automatically (Brent, 1993; Carroll and Rooth, 1998; Briscoe and Carroll, 1998). While this has the potential advantage of giving frequency information for subcategorization, it also has the drawback that considerable energy has to be spent on creating a (shallow) parser able to recognize with sufficient accuracy the relevant syntactic configurations. Acquisition of subcategorization information from a syntactically annotated corpus is much more straightforward, leading mainly to questions whether a dependent is to be counted as a selected argument or an adjunct (Collins, 1999; Sarkar and Zeman, 2000), but obtaining reasonable coverage requires large corpora.

11800	Total number of verbal stems
21800	Total number of dependency frames
650	Dependency frame types
300	Unique dependency frame types
6574	[SU:NP][OBJ1:NP]
4188	[SU:NP]
1161	[SU:NP][LD:PP⟨*pform*⟩]
1021	[SU:NP][PC:PP⟨*pform*⟩]
826	[SU:NP][OBJ1:NP][LD:PP⟨*pform*⟩]
549	[SU:NP][OBJ1:NP][PC:PP⟨*pform*⟩]
408	[SUP:⟨het⟩][OBJ1:NP][SU:SDAT]
341	[SU:NP][OBJ1:SDAT]
275	[SU:NP][OBJ2:NP][OBJ1:NP]
274	[SU:NP][SE:NP]

Table 1: Key figures and the 10 most frequent dependency frame types for the CGN/Celex lexical database. (*Pform* is a placeholder for various preposition forms. SUP and SE are the relation names for expletive subjects and inherently reflexive arguments, respectively. SDAT is the category for subordinate clauses introduced by the complementizer *dat*).

6 Using Existing Resources

Currently the resources required to do automatic extraction of dependency frames for Dutch are not available. However, two lexical resources exist which provide dependency frames. These have been used to create a lexicon for the Alpino Grammar with detailed subcategorization and dependency information for verbs and nouns. Below, we describe the verbal entries in both resources.

Celex (Baayen et al., 1993) is a large lexical database for Dutch, with rich phonological and morphological information. For use within the project *Corpus Spoken Dutch* (CGN), this database has been extended with dependency frames (Groot, 2000). Some key figures are given in table 1. Note that there is considerable variation in the distribution of dependency frames. A large number of frames is associated with only a few verbs, with 300 dependency frame types being associated with only a single verb.

The Dutch Parole lexicon[2] has been created as part of a project aiming at the development of uniform lexical and corpus resources for a number of European languages. The Parole lexicon comes with detailed subcategorization information, but dependency relations differ from those in the CGN proposal. Key figures are given in table 2.

While the mapping from Parole dependency frames

[2] http://www.inl.nl/corp/parole.htm

3200	Total number of verbal stems
5000	Total number of dependency frames
320	Dependency frame types
190	Unique dependency frame types
1566	[SU:NP][OBJ1:NP]
474	[SU:NP][PC:PP⟨*pform*⟩]
378	[SU:NP][ADV:PP⟨*pform*⟩]
208	[SU:NP][OBJ1:NP][OPT:PC:PP⟨*pform*⟩]
205	[SU:NP]
204	[SU:NP][ADV:ADV]
204	[SU:NP][OBJ1:NP][OPT:ADV:PP⟨*pform*⟩]
163	[SU:NP][OBJ1:NP][PC:PP⟨*pform*⟩]
107	[SU:NP][SE:NP][PC:PP⟨*pform*⟩]
101	[SU:NP][VC:S⟨subordinate,dat⟩]

Table 2: Key figures and the 10 most frequent dependency frame types for the Parole lexical database. Notation has been made conformant with the CGN/Celex notation where possible. Optional complements are marked OPT.

into the CGN dependency frames is mostly straightforward, there are also a number of problematic cases. The ADV dependency relation in Parole, for instance, has no obvious corresponding dependency relation in CGN, although manual inspection leads us to suspect that in many cases it corresponds to the LD (*locative/directional* complement) relation. Currently, verbs with dependency frames containing the ADV relation are not extracted. Another notable difference between the two sources is the relatively small number of intransitive verbs in Parole. This is partly related to the ADV dependency relation in Parole. Adverbial elements are often optional and subject to wide variation (i.e. adverbial PPs are not restricted to a small set of *pforms*, and adverbial dependents can often be both adverbs and PPs. However, even if these elements are counted as true modifiers (and thus not as part of the subcategorized-for dependents of the verb), the number of intransitives remains relatively small.

7 The Alpino Lexicon

Dependency frames for the verbal lexicon of the Alpino Grammar have been constructed using the dependency information provided by CGN/Celex, Parole, and by entering definitions by hand. The latter has been done mostly for auxiliary and modal verbs, a small class of high-frequent elements which are exceptional in a number of ways. The CGN/Celex dictionary is exceptionally large. As the Celex database comes with frequency information, we currently only include those lexical items whose frequency is above a certain threshold. For verbal stems, this means that roughly 50% of the stems in Celex is included in the Alpino lexicon. All verbal stems from the Parole lexicon with a dependency frame covered by the grammar are included.

Extraction of verbs with a specific dependency frame from Celex and Parole requires that a particular frame in the database is identified and given a definition in the Alpino Grammar. Currently, for 28 different CGN/Celex dependency frames a definition in the grammar has been provided. This covers over 80% of the verbal dependency frames in the CGN/Celex database, 10,400 of which are sufficiently frequent to be included in the Alpino lexicon. For 15 different dependency frames in the Parole lexicon a definition in Alpino is present. Using these, we extract over 4,100 dependency frames.

As CGN/Celex is the larger database, one might suspect that this database is more exhaustive than Parole. However, the union of the frames extracted from CGN/Celex and Parole contains 11,700 frames, which means that Parole contributes 13% of the frames in the Alpino lexicon. An overview of overlap and non-overlap for the most frequent frames extractable from both sources is given in table 3.

For transitive and intransitive verbs, we see that over 85% of the stems in Parole are present in CGN/Celex as well. For most other dependency frames, however, the overlap is generally much smaller, and a significant portion of the stems present in Parole is not present in Celex. This suggests that, for more specific subcategorization frames, both resources are only partially complete, and that not even the union of both provides exhaustive coverage.

As we are currently only using the most frequent 50% of the CGN/Celex database in the Alpino lexicon, we also compared Parole with the complete CGN/Celex database. Here we found that the absolute number of dependency frames goes up dramatically only for transitive and intransitive verbs, and that practically all intransitive and transitive Parole stems are included in the full CGN/Celex database. For the other dependency types, however, the figures are comparable to those given in table 3. The relatively high number of transitive and intransitive verbal stems in Parole also present in Celex is therefore probably due to the fact that in Celex these are assigned as a default to most verbal stems. This also explains why the low frequency verbs consist almost exclusively of stems with transitive or intransitive dependency frames.

A more direct method to establish coverage of the lexicon is to see to what extent the dependency frames present in a treebank are covered by the lexicon. For a small dependency treebank, annotated according to the format presented in section 2, we extracted all verbal heads, together with their non-modifier dependents. Sets of dependents were identified with specific dependency frames. For instance, if a verb occurred with an NP subject and a PP with the PC dependency relation and *prep* as head, it is assumed that this verb must be associated with the `[SU:NP][PC:PP⟨prep⟩]` dependency frame. Coverage can now be tested by counting how often a dependency frame in the treebank also occurs in

Dependency Frame	Overlap	Celex only	Parole only	Total
[SU:NP][OBJ1:NP]	1810	1211	240	3261
[SU:NP]	257	1697	42	1996
[SU:NP][PC:PP⟨*pform*⟩]	337	541	273	1151
[SU:NP][OBJ1:NP][PC:PP⟨*pform*⟩]	129	375	308	812
[SU:NP][VC:S⟨subordinate⟩]	103	136	103	342
[SUP:NP⟨het⟩][OBJ1:NP][SU:CP]	7	247	5	259
[SU:NP][OBJ2:NP][OBJ1:NP]	65	171	28	264
[SU:NP][SE:NP][PC:PP⟨*pform*⟩]	65	62	102	229
[SU:NP][SE:NP]	49	137	65	251
[SU:NP][VC:VP]	10	16	37	63

Table 3: Dependency Frames and the number of stems occurring with this frame in both resources, in CGN/Celex only, in Parole only, and the total number of stems with this dependency frame in the resulting Alpino Lexicon.

the lexicon. Extraction of dependency frames is mostly straightforward. Problematic cases are those where one dependency frame is more general than another. For instance, a verb occurring with a VP-dependent introduced by the complementizer *om* might be associated with a dependency frame selecting for an *om*-VP, but also with a more general dependency frame selecting for a VP (with or without complementizer). In such cases, we check whether at least one of the potential frames occurs in the lexicon.

We applied the evaluation method described above to a treebank, constructed for grammar evaluation purposes, consisting of 424 short sentences (up to 10 words) selected from the Eindhoven-corpus (Uit den Boogaart, 1975), with a total of just over 2,200 words. The test-set contained 473 verbal heads, 417 of which (88%) occurred in a dependency configuration which was also present in the lexicon. Although one obviously would like to obtain figures from a larger test-set, we believe that this is an encouraging result. Carroll and Briscoe (1996), for instance, report that in a small test set 12% of sentences failed to parse due to missing subcategorization information in their ANLT lexicon (which is comparable in size to our lexicon, and contains subcategorization information extracted automatically from a learners dictionary). Coverage seems higher than what can be achieved by methods based on automatic extraction of subcategorization frames. Briscoe and Carroll (1997), for instance, estimate a token recall (i.e. the percentage of true positives of the learned frames in a corpus) of 81%.

We have extracted dependency frames for nouns, but have not carried out a systematic evaluation for these dependency frames. Currently, we are extracting almost 2.000 dependency frame tokens for nouns selecting prepositional complements, more than 1.000 dependency frame tokens for nouns selecting verbal (infinitival or finite sentential) complements, and over one hundred frames for measure nouns and titles (*vice-president Jansen*).

8 Future Work

There are still a number of verbal dependency frames in the resources which are not included in the Alpino lexicon. Adding these is partially just a matter of adding the relevant definitions to the grammar. For the Parole database, the most valuable addition would be to find a method for dealing with dependency frames containing an ADV dependency relation.

The lexicon currently does not contain information on the frequency of stem/dependency frame combinations.[3] Given a sufficiently accurate parser, we could try to collect such statistics from unannotated text automatically. The fact that lexical coverage of dependency frames is relatively high suggests that this may be feasible in the near future.

Finally, we are considering methods to supplement the information in the current lexicon with dependency frames acquired automatically from corpora. The fact that we currently cover around 88% of the frames found in an annotated sample of text, suggests that it is worthwhile to look for ways of expanding the dependency frames in the lexicon.

References

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

Gosse Bouma and Gertjan van Noord. 1997. Word order constraints on Germanic verb clusters. In Erhard Hinrichs, Tsuneko Nakazawa, and Andreas Kathol, editors, *Complex Predicates in Nonderivational Syntax*, pages 43–72. Academic Press, New York.

Gosse Bouma, Rob Malouf, and Ivan Sag. 2001. Satisfying constraints on adjunction and extraction. *Natural Language and Linguistic Theory*, 19:1–65.

Michael Brent. 1993. From grammar to lexicon: Un-

[3] Note that Celex does provide frequency information, but only for word or stem forms, not for parsed text.

supervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

Ted Briscoe and John Carroll. 1998. Can subcategorization probabilities help a statistical parser? In *Proceedings of the ACL/SIGDAT workshop in Very Large Corpora*, Montreal.

John Carroll and Ted Briscoe. 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. In *Proceedings of the ACL SIGDAT Conference on Empirical Methods in Natural Language Engineering*, pages 92–100, University of Pennsylvania, Philadelphia.

Glenn Carroll and Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain.

John Carroll, Ted Briscoe, and Antonio Sanfilippo. 1998. Parser evaluation: A survey and a new proposal. In *Proceedings of the first International Conference on Language Resources and Evaluation (LREC)*, pages 447–454, Granada, Spain.

Michael Collins. 1999. *Head-driven Statistical Models for Natural Language Processing*. Ph.D. thesis, University of Pennsylvania.

Mila Groot. 2000. Lexiconopbouw: microstructuur. Internal report of the project Corpus Gesproken Nederlands.

E. Hajicova, J. Panevova, and P. Sgall. 1998. Language resources need annotations to make them really reusable: The Prague Dependency Treebank. In *Proceedings of the First International Conference on Language Resources (LREC)*, pages 713–718, Granada, Spain.

David Magerman. 1994. *Natural Language Parsing as Statistical Pattern Recognition*. Ph.D. thesis, Stanford University.

Nelleke Oostdijk. 2000. The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings of Second International Conference on Language Resources and Evaluation (LREC)*, pages 887–894.

Carl Pollard and Ivan Sag. 1994. *Head-driven Phrase Structure Grammar*. Center for the Study of Language and Information Stanford.

Anoop Sarkar and Daniel Zeman. 2000. Automatic extraction of subcategorization frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken.

Wojciech Skut, Brigitte Krenn, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC.

P. C. Uit den Boogaart. 1975. *Woordfrequenties in geschreven en gesproken Nederlands*. Oosthoek, Scheltema & Holkema, Utrecht. Werkgroep Frequentie-onderzoek van het Nederlands.

Gertjan van Noord and Gosse Bouma. 1994. Adjuncts and the processing of lexical rules. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pages 250–256, Kyoto.

Gertjan van Noord, Gosse Bouma, Rob Koeling, and Mark-Jan Nederhof. 1999. Robust grammatical analysis for spoken dialogue systems. *Journal of Natural Language Engineering*, 5:45–93.