Enriching a Descriptive Grammar with Treebank Queries

Gosse Bouma,[1] Marjo van Koppen,[2] Frank Landsbergen,[3]
Jan Odijk,[2] Ton van der Wouden[4] and Matje van de Camp[5]

[1]University of Groningen, [2]Utrecht University,
[3]Institute for Dutch Lexicography, [4]Meertens Institute, [5]De Taalmonsters
E-mail: g.bouma@rug.nl, j.m.vankoppen@uu.nl,
franklandsbergen@inl.nl, j.odijk@uu.nl,
ton.van.der.wouden@meertens.knaw.nl,
matje@taalmonsters.nl

Abstract

The Syntax of Dutch (SoD) is a descriptive and detailed grammar of Dutch, that provides data for many issues raised in linguistic theory. We present the results of a pilot project that investigated the possibility of enriching the on-line version of the text with links to queries that provide relevant results from syntactically annotated corpora.

1 Introduction

The Language Portal Dutch/Frisian[1] (Landsbergen et al., 2014) is an on-line resource of descriptive linguistic resources, covering syntax, morphology, and phonology of Dutch and Frisian. It contains, among others, an on-line edition of the Syntax of Dutch (SoD) (Broekhuis et al., 2012–), a descriptive grammar of Dutch that goes well beyond the level of detail provided by other sources. Although descriptive, the emphasis in the selection and presentation of phenomena is clearly guided by discussions in the theoretical literature.

In his largely positive review of the SoD volumes on NP syntax, Hoeksema (2013) points out that *"There is a growing body of work in empirical studies of judgment variation [...] that future extensions of this grammar could benefit from, especially when coupled to studies of actual usage patterns in corpus material"* and that *"This particular reader would also have welcomed to see some more lists in the book"*. By enriching the on-line version of SoD with queries over syntactically annotated corpora, the current project tries to accommodate the needs of researchers like Hoeksema.

[1] www.taalportaal.org

Creating a link between a descriptive grammar and a syntactically annotated corpus can be valuable for various reasons. Illustrating a given construction with corpus examples may help to get a better understanding of the variation of the construction and the frequency of these variants. Corpus data may also convince a reader that a given variant actually occurs in (well-formed) text, or in some cases may illustrate that examples judged ungrammatical by the authors of the descriptive grammar do occur with some frequency in actual text.

The (syntactically annotated part of the) Corpus of Spoken Dutch (manually verified, speech from various situations, 1M words) (Oostdijk, 2000), the Lassy Small treebank (manually verified, written material from various genres, 1M words, 65,200 sentences) and the Lassy Large treebank (automatically created[2], written material from various genres, 700M words, 8.6M sentences)) (van Noord et al., 2013) are all suitable corpora for our project. The first two resources provide high-quality data for a limited amount of text, while the last resource provides wide-coverage, but noisy, data. All treebanks follow (with minor modifications) the same annotation standard (Schuurman et al., 2003).

The innovative aspect of this project is the use of syntactically annotated corpora as resource. While descriptive grammars have been based on corpus research, there have been only a few attempts at documenting and extending such grammars with links to relevant examples from treebanks (but see Bender et al. (2012)). The level of annotation that is most valuable for such a resource, i.e. syntactic constituency and grammatical dependency information, does not always align well with the conceptual and ontological assumptions made in the descriptive grammar. Therefore, adding precise treebank queries to a descriptive grammar can be challenging. The goal of the current project is to investigate to what extent a fruitful combination of the two is possible and how much manual effort is required for the development of queries that illustrate phenomena discussed in the descriptive grammar.

Below we describe the treebanks and query tool used in our project. We then give some examples of phenomena that were problematic for our approach, either because annotations did not match, or because the phenomena are so rare that they are hard to find with reasonable precision in the (automatically annotated) treebank. We also give an impression of the coverage of the treebanks, and of the complexity of the queries. Next, we discuss related work and we finish with a discussion of the results.

2 Search interface

We use the web-based corpus query tool PaQu[3] in combination with the example-based query system Gretel[4] for creating and executing treebank queries. The PaQu

[2] using the Alpino parser (van Noord, 2006)

[3] `http://zardoz.service.rug.nl:8067/xpath`.

[4] `http://nederbooms.ccl.kuleuven.be/eng/gretel`.

interface returns matching sentences in the selected corpus, with the option to display the matching nodes in the syntactic dependency graph. It displays the query being executed along with a brief description. Queries are dynamic, i.e. the user can switch between treebank corpora, or substitute a given lexical item by an alternative. Furthermore, users can select up to three attributes (i.e. lemma, part of speech, dependency relation, etc.) of matching nodes to obtain a frequency distribution of the attribute-values. Advanced users can also modify the XPATH query as they see fit. Integration of queries into the electronic version of the SoD will be done by adding links (in the form of an icon) to paragraphs and examples for which queries are available.

Construction of queries can be challenging, as it is not always clear how a given constraint should be expressed in terms of XPATH, but also because it is not always clear how a given phenomenon is annotated in the treebanks. To facilitate query formulation, we have used Gretel (Augustinus et al., 2012), a corpus query tool that supports the formulation of XPATH queries that are compatible with the treebank annotation. Users can enter an example sentence, which is parsed automatically by Alpino. Next, relevant parts from the dependency tree can be selected, and a corresponding XPATH query is created. This query can be used to find similar cases in the treebank.

As an example, consider the following statement from SoD concerning the linear order of adjectives and their PP-complements:[5]

> Adjectives typically select a PP as their complement. Although this PP-complement can often either precede or follow the adjective, it is normally assumed that its base-position is the one following the adjective, whereas the pre-adjectival position is derived by leftward movement.
>
> (1) a. Jan was ⟨over deze opmerking⟩ boos ⟨over deze opmerking⟩
> Jan is about that remark angry
> b. Jan is ⟨over zijn beloning⟩ tevreden ⟨over zijn beloning⟩
> Jan is about his reward satisfied

Adjectives selecting for a PP-complement are relatively frequent, and Lassy Small contains many examples of sentences illustrating this syntactic configuration. An example is given in Figure 1. A query that searches the treebank for adjectives selecting a PP-complement is:

[5] http://www.taalportaal.org/taalportaal/topic/link/syntax__Dutch__ap__a2__a2_complementation.2.1.xml.

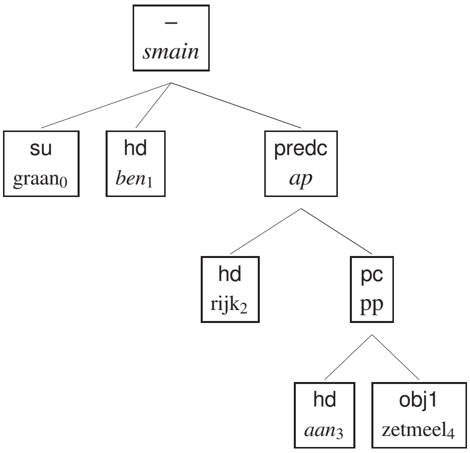

Figure 1: Dependency tree for 'Graan is rijk aan zetmeel' (*Corn is rich with starch*).

```
//node[@cat="ap"]/
        node[@rel="hd" and
             @pt="adj" and
             ../node[@rel="pc" and
                     @cat="pp"]
             ]
```

This query selects the adjectival head of a node of category AP. Furthermore, the node that matches the head has to have a sibling that is of category PP and whose dependency relation is PC (*prepositional commplement*). Here '//' matches an arbitrary position in a tree, '/' denotes the 'child of' relation and '../' denotes the sibling relation. The query below adds the constraint that the PP has to precede the adjective:

```
//node[@cat="ap"]/
        node[@rel="hd" and
             @pt="adj" and
             ../node[@rel="pc" and
                     @cat="pp"]/number(@end) = number(@begin)
             ]
```

The attributes `begin` and `end` refer to the begin and end position (in the string) of the corresponding lexical or syntactic node. Here, we require that the end position of the PP has to be equal to the begin position of the adjective.

Counts for adjectives in Lassy Small matching with the first and second query, respectively, are given in Table (2). With 1,125 hits (for 186 lemma's) PP-complements of adjectives are relatively frequent (i.e. occurring in approx. 2% of the sentences in the corpus). When we restrict attention to PP-A order, however, only 85

Adjective	A+PC	PP-A order
afhankelijk (*dependent*)	100	8
verantwoordelijk (*responsible*)	79	3
afkomstig (*originating*)	56	8
nodig (*needed*)	49	6
eens (*agreed*)	44	18
bezig (*busy*)	34	6
goed (*good*)	34	0
vergelijkbaar (*comparable*)	26	0
bewust (*conscious*)	25	0
tevreden (*content*)	25	0
...		
boos (*angry*)	2	0
total	1,125	85

Table 1: Adjectives with a PP-complement in Lassy Small (second column) and cases where the complement precedes the adjective (third column).

hits remain (for 30 lemma's), i.e. the PP-A order occurs in less than 10% of all cases where we find a PP-complement. This underlines the point made in the descriptive grammar, that A-PP orders are in some sense more basic or less 'marked' than PP-A orders. One might also wonder whether some adjectives do not allow PP-A orders at all. For instance, the adjective *boos*, used in (1-a), does not occur with this word order in Lassy Small. If we execute the same queries on Lassy Large, we find that there are 76 hits for *boos*+PC, but only one for the order PP+*boos*:

(2) Leopold II was over die aantasting van … bijzonder boos
Leopold II was over that violation of … extremely angry
Leopold II was extremly upset with that violation of…

This suggests that the PP-A order is exceptional but not impossible for the adjective *boos*.

3 Query development

The SoD uses generic linguistic concepts to present its analyses. Although there is some reference to concepts from generative linguistics, the analyses appear to be general enough to be translatable into most syntactic frameworks. The treebank annotation uses both dependency relations and constituent labels. Dependency relations are widely used in computational linguistics (e.g. see the Universal Dependency format (De Marneffe et al., 2014) that is quickly gaining popularity). The annotation style used in the Dutch treebanks follows earlier work on German (Brants et al., 2003). The dependency annotation allows for crossing branches, something that simplifies annotation of Dutch word order significantly. The preservation of constituent nodes allows a connection with analyses couched in terms of phrase structure trees.

While this set-up suggests that it should be relatively straightforward to translate analyses as formulated in the SoD into treebank terms, in practice this turned out to be challenging for a substantial number of phenomena. This can be due to principled and motivated differences in analysis between the two sources, or by the fact that one of the two sources makes a distinction that is missing in the other.

For instance, the SoD presents a (somewhat artificial) distinction between genitive (3-a) and dative (3-b) nominal complements of adjectives:

(3) a. Jan is zich dat probleem bewust
Jan is REFL that problem$_{ACC}$ aware
John is aware of that problem
b. Het probleem werd Peter niet duidelijk
the problem became Peter$_{DAT}$ not clear
The problem didn't become clear to Peter

In the treebank, the adjective *bewust* does indeed occur with a nominal complement (labeled with the dependency relation *obj1*) (Figure 2, left). Examples like (3-b) occur as well, but not as a single constituent. Instead, *duidelijk* is annotated as predicative complement of the verb *worden* and *Peter* is annotated as an indirect object (*obj2*) complement of *worden* (Figure 2, right).

The most effective method for becoming aware of such mismatches is to parse the example from the descriptive grammar with the example-based query system Gretel (Augustinus et al., 2012). Gretel uses Alpino for syntactic analysis, and thus its results are guaranteed to be consistent with data from the automatically annotated corpus Lassy Large and, given the high level of accuracy and coverage of Alpino, usually also with the manually annotated treebanks. A user can highlight relevant parts of the dependency tree, and Gretel will construct an XPATH query on the basis of this. This query can than be used to search the treebanks for more examples.

While most complementation and modification possibilities mentioned in SoD are easily found in the manually verified treebanks, this is not the case for all word order possibilities being discussed. For instance, the SoD discusses discontinuous APs like (4) in terms of 'PP-over-V', and 'topicalization'.

(4) a. Trots is Jan nooit geweest op zijn vader
Proud has Jan never been of his father
Jan has never been proud of his father
b. Op zijn vader is Jan nooit trots geweest

In the treebank, discontinuous constituents are annotated as such, i.e. as nodes in a dependency graph that dominate a discontinuous part of the sentence (see

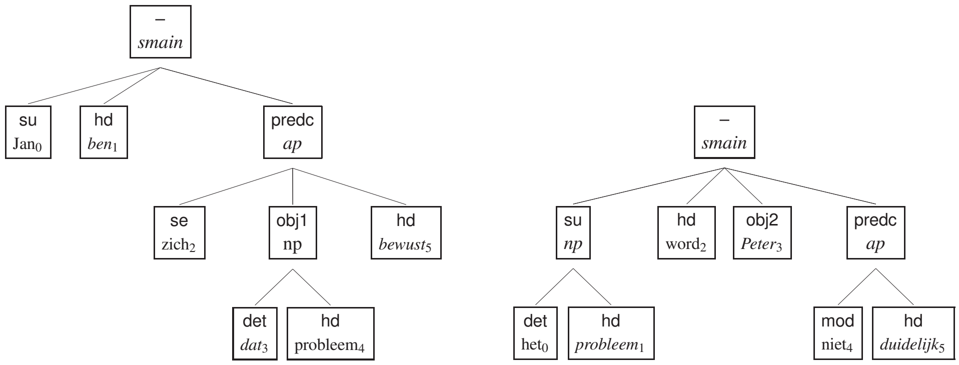

Figure 2: Treebank annotation of *Jan is zich dat probleem bewust* (*John is aware of that problem*) and *Het probleem werd Peter niet duidelijk* (*The problem did not become clear to Peter*).

Figure 3). Using the `begin` and `end` attributes of nodes, we can easily search for sentence initial adjectives that have a non-adjacent PP-complement, or, to find cases like (4-b), for sentence initial prepositional complements of adjectives. The second word order occurs with minimal frequency in our data, returning 34 hits on Lassy Small. Some examples are given in (5).

(5) a. Voor deze activiteiten is veel geld nodig
For these activities is much money needed$_{\text{ADJ}}$
These activities require a considerable amount of money

b. Vooral over Mijn vlakke land was Brel zeer tevreden
Especially about Le Plat Pays was Brel very content
Brel was especially pleased with Le Plat Pays

c. Over de oorzaak is nog niets bekend
On the cause is yet nothing known$_{\text{A}}$
Nothing is known yet about the cause

Word orders like (4-a) are far less frequent, however, and can only be found in the Lassy Large treebank. While returning 9 valid hits, search on Lassy Large also returns 11 false or debatable hits. Some examples are shown in (6) below. The last example, (6-d), is a false hit. All false hits are cases of sentences starting with an adjective and ending with a PP, where the parser erroneously prefers to analyse the PP as a complement of a distant adjective instead of attaching it as a modifier to a nearby noun. Despite the moderate accuracy of the automatic annotation on such cases, we believe the result is valuable, as it provides quick access to valid examples that are much harder to find using less sophisticated search methods (i.e. combinations of word and part-of-speech patterns).

(6) a. Verliefd$_{\text{ADJ}}$ was hij doorlopend en dan bij voorkeur [$_{\text{PP}}$ op
love was he continuously and than by preference with

jonge dames tussen 15-20 jaar]
young ladies between 15-20 years
He was continuously in love, and preferably with young ladies in the age of 15-20 years

b. Beroemd$_{\text{ADJ}}$ werd hij [PP met zijn openlijke uitspraken in de
famous became he with his public statements in the
pers over seks en drugsgebruik]
press on sexuality and drug-use
He became famous for his public statements in the press on sexuality and use of drugs

c. Enthousiast$_{\text{ADJ}}$ werd hij [PP over de muziek van de jonge
enthousiast became he over the music of the young
componist George Gershwin]
composer George Gershwin
He became enthousiastic about the music of the young composer George Gershwin

d. Beroemd$_{\text{ADJ}}$ is de eerste foto van prinses Beatrix [PP met Claus
famous is the first picture of princess Beatrix with Claus
von Amsberg]
von Amsberg
The first picture of Beatrix with Claus von Amsberg is famous

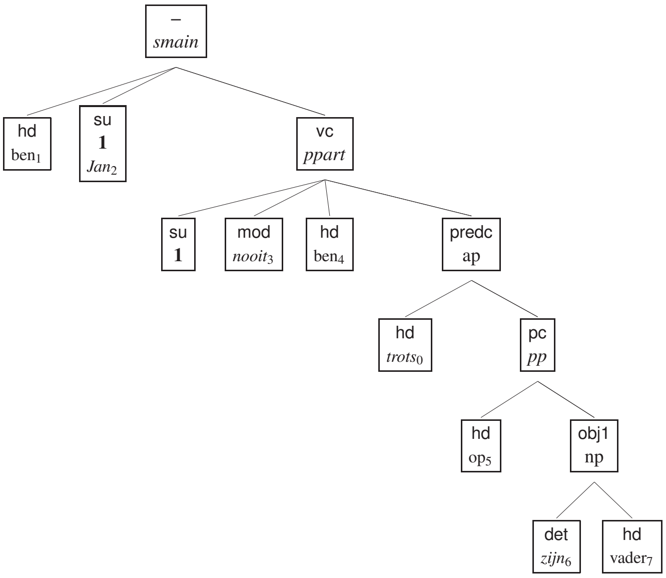

Figure 3: Dependency tree for (4-a). Note that the node dominating *trots op zijn vader* forms a discontinuous constituent.

SoD also discusses PP-A orders in sentence-initial position, like (7).

Corpus	Query type				Sum		%
	Synt (-,+w.o.)		Lex (-,+w.o)				
Lassy Small	228	(168, 60)	527	(409, 118)	755	(577, 178)	63.1
Lassy Large	45	(24, 21)	377	(260, 117)	422	(284, 138)	35.2
CGN	2	(1, 1)	18	(17, 1)	20	(18, 2)	1.7
Total	275	(193, 82)	922	(686, 236)	1,197	(879, 318)	100.0

Table 2: Distribution of queries over corpora

(7) Voor deze functie geschikt is hij niet
For this job suited is he not
He is not fit for the job

Such word orders cannot be found in the manually verified treebanks. In Lassy Large, searching for sentence-initial AP's starting with a prepositional complement also does not return any results. It turns out that the dependency treebank annotation guidelines analyse examples such as (7) as verbal constituents headed by a passive participle.[6] Searching for predicatively used sentence-initial verbal constituents containing a prepositional complement does return a small number of hits.

4 Coverage

For selected sections of the SoD, covering adjectival phrases (complementation, pronominalization, discontinuous cases, modification, and comparative constructions), and adpositions (complementation, absolute PP constructions, and modification), we have constructed almost 1,200 queries.

We assumed that most queries would be formulated over the (manually annotated) Lassy Small corpus, and that the Lassy Large and Spoken Dutch corpus would only be used if Lassy Small returned no hits. Table 2 shows that 63% of the queries indeed use the Lassy Small corpus. The Corpus of Spoken Dutch, eventhough equal in size to the Lassy Small corpus, is only rarely used.

Most queries (922, %) are 'lexical', i.e they search for a specific lexical item occurring in some syntactic context. The other 'syntactic' queries only specify a syntactic context. Queries that do not refer to word order ('-w.o.') are purely configurational. Other queries ('+w.o.') do refer to linear order. By far the most queries are anchored to some lexical item, and also most queries do not refer to linear order. The proportion of lexical queries and the proportion of word-order queries is larger in Lassy Large than in Lassy Small. This suggests that coverage of Lassy Small is sufficient to find examples for many standard syntactic configurations and frequent

[6] It should also be noted that the Alpino parser analyses *geschikt* and similar deverbal adjectives as adjectives. In the conversion step from internal parse representation to treebank annotation, the PoS tag is replaced by a verbal tag.

lexical items, while Lassy Large is used to search for infrequent combinations of a lexical item and syntactic context or word order.

The number of hits per query varies strongly. This is to be expected, as queries that search for some syntactic configuration, without imposing lexical or word order constraints, will usually return a large number of hits. Such queries will be useful mostly because they provide statistics for the syntactic heads occurring in these constructions. Queries that return only a small number of hits, are often queries anchored to a specific lexical item or searching for a non-canonical word order; these are valuable as they illustrate that such constructions do occur, though perhaps rarely, in natural text.

5 Related work

Bender et al. (2012) argue that computational grammars and treebanks can be valuable resources for documenting descriptive grammars. They demonstrate how a descriptive grammar for Wambaya (Nordlinger, 1998) could be used as starting point for the implementation of a computational grammar that covers over 90% of the example sentences in the descriptive grammar and over 75% of held out material from the same language. The computational grammar provides fully explicit analyses of sentences, something that a descriptive grammar cannot do. If the computational grammar is also used (in combination with manual disambiguation decisions to arrive at the optimal parse) to annotate a corpus fragment, a treebank results that can be used to further enrich the descriptive grammar. They argue that 'canned queries' over the treebank may be useful for users who are not familiar with the treebank design or query language, to find exemplars for given syntactic phenomena. If the treebank and query language is adequately documented, users can also formulate their own queries. Our approach provides both options. As Bender et al. (2012) we believe that preformulated queries can be important not only for non-expert users, but also as a means to document the various possibilities of obtaining results from the treebank.

Bender et al. (2012) use the query language Fangorn (Ghodke and Bird, 2012). van Noord et al. (2013) show that XPATH queries over Alpino-style dependency trees (where there is a one-to-one correspondence between linguistic dominance and embedding of elements in XML, and where word order is encoded by XML attributes that register string positions) can deal with all the cases used as test cases for linguistic query languages by Lai and Bird (2004). We therefore prefer to use XPATH, as it has the important additional advantage that it is a widely accepted standard supported by numerous XML processing tools.

Hashimoto et al. (2008) use an annotated treebank to obtain detailed syntactic information on the lexical types that occur in the treebank. Their aim is to ensure consistency both in future extensions of the treebank, as well as for computational grammars that follow the annotation guidelines underlying the treebank annotation. Flickinger et al. (2014) similarly use a treebank primarily as a means for documenting and validating their computational lexicon and grammar. Our work differs in that it uses a treebank to enrich a descriptive grammar that is completely unrelated to the treebank or the guidelines used for annotating the treebank. As a consequence, we cannot assume a transparant conceptual mapping between analyses as discussed in the descriptive grammar on the one hand and underlying the treebank annotation on the other.

6 Conclusions

After completion of approx. 1,200 queries, we have learned that creating suitable queries for a given fragment from the SoD requires creativity and careful experimentation, tuning, and documentation. Construction of queries is far from deterministic, that is, different annotators will have different opinions concerning the most suitable query for a given example or phenomenon. In a substantial number of cases, there are mismatches (in constituent structure, in part-of-speech) between the presentation in the SoD and the treebank annotation. While this makes the development of queries harder, it also underlines the value of the current project: by systematically exploring the way various linguistic examples are annotated in the treebank, we provide a starting point for further corpus exploration for users that have a general linguistic interest but who are not necessarily experts on Dutch treebank annotation.

The manually verified treebanks almost always provide sufficient examples of basic word order patterns for queries that are not restricted to a specific adjective or preposition. For queries that search for a specific lexical head or for less frequent word order patterns, the Lassy Large treebank usually has to be used. In that case, users must be prepared to see also a certain number of false hits. However, there are also examples in the SoD that cannot be found in a 700M word corpus. The conclusion that such word orders are not found in the language would be too strong, but it might be a starting point for further research (i.e. *does this construction occur only in certain registers or discourse settings?*) or for an alternative analysis (i.e. *do these cases really involve adjectives?*).

References

Liesbeth Augustinus, Vincent Vandeghinste, and Frank Van Eynde. Example-based treebank querying. In *LREC*, pages 3161–3167, 2012.

Emily M. Bender, Sumukh Ghodke, Timothy Baldwin, and Rebecca Dridan. From database to treebank: On enhancing hypertext grammars with grammar engineering and treebank search. *Language Documentation & Conservation*, Special Publication No. 4:179–206, 2012.

Thorsten Brants, Wojciech Skut, and Hans Uszkoreit. Syntactic annotation of a German newspaper corpus. In *Treebanks*, pages 73–87. Springer, 2003.

H. Broekhuis, E. Keizer, M. den Dikken, N. Corver, and R. Vos. *Syntax of Dutch*. Amsterdam University Press, Amsterdam, 2012–. (several volumes).

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of LREC*, pages 4585–4592, 2014.

Dan Flickinger, Emily M Bender, and Stephan Oepen. Towards an encyclopedia of compositional semantics: Documenting the interface of the English Resource Grammar. In *Proceedings of the Ninth International Conference of Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland, 2014.

Sumukh Ghodke and Steven Bird. Fangorn: A system for querying very large treebanks. In *COLING (Demos)*, pages 175–182, 2012.

Chikara Hashimoto, Francis Bond, Takaaki Tanaka, and Melanie Siegel. Semi-automatic documentation of an implemented linguistic grammar augmented with a treebank. *Language Resources and Evaluation*, 42(2):117–126, 2008.

Jack Hoeksema. Review of: Syntax of Dutch. Noun and Noun Phrases vols. 1 and 2. *Lingua*, 133:385–390, 2013.

Catherine Lai and Steven Bird. Querying and updating treebanks: A critical survey and requirements analysis. In *Proceedings of the Australasian language technology workshop*, pages 139–146, 2004.

Frank Landsbergen, Carole Tiberius, and Roderik Dernison. Taalportaal: an online grammar of Dutch and Frisian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2206–2210, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

Rachel Nordlinger. *A Grammar of Wambaya*. PhD thesis, Research School of Pacific and Asian Studies, The Australian National University, Canberra, 1998.

Nelleke Oostdijk. The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings of LREC 2000*, pages 887–894, 2000.

Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra, and Ton van der Wouden. CGN, an annotated corpus of spoken Dutch. In *Proceedings of 4th International Workshop on Language Resources and Evaluation*, pages 340–347, 2003.

Gertjan van Noord. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42. 2006.

Gertjan van Noord, Gosse Bouma, Frank van Eynde, Daniel de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: the STEVIN Programme*, pages 147–164. Springer, 2013.