1 LASSY: Large Scale Syntactic Annotation of written Dutch

A large corpus of written Dutch texts (1,000,000 words) is syntactically annotated (manually corrected), based on D-COI. In addition, the full D-COI corpus is syntactically annotated automatically. The project aims to extend the available syntactically annotated corpora for Dutch both in size as well as with respect to the various text genres and topical domains. In addition, various browse and search tools for syntactically annotated corpora will be further developed and made available. Their potential for applications in corpus linguistics and information extraction will be illustrated and evaluated.

2 Principal Investigator

dr. G.J.M. van Noord
Alfa-informatica
Rijksuniversiteit Groningen
Oude Kijk in ’t Jatstraat 26
Postbus 716 9700 AS Groningen
+31-50-3637811
vannoord@let.rug.nl

3 Composition of the Research Team

dr. G.J.M. van Noord, Alfa-informatica Groningen
drs. I. Schuurman, CCL Leuven
Prof.dr. F. van Eynde, CCL Leuven
dr. G. Bouma, Alfa-informatica Groningen

Both Leuven (CGN) and Groningen (Alpino Treebank) contribute experience in the construction of syntactically annotated corpora. They are also responsible for the corresponding work package of the STEVIN D-COI project. Leuven contributes experience with POS-tagging and lemmatization (they are responsible for the corresponding work package in D-COI). A consortium with both a Dutch and a Flemish partner ensures that both the northern and Flemish language variety will be accounted for appropriately. Groningen provides additional know-how with respect to the use of the Alpino parser and related tools, as well as with the use of automatically annotated corpora for various applications. Leuven provides additional know-how with respect to syntactic annotation, based a.o. on their experience in the syntactic annotation of CGN.

4 Requested budget

senior 0.1 fte	for three years	Groningen
junior 1.0 fte	for three years	Groningen
senior 0.3 fte	for two years	Leuven
student assistents 1 fte	for two years	Groningen
student assistents 1.5 fte	for two years	Leuven
hardware (HPC, hard disks)		Groningen
external validation		

The project lasts for three years. Start date: September 1, 2006 (at the end of D-COI). Below (section on work programme) we relate the budget to the work programme. The budget for hardware is requested for additional hard disks which may be required to store the very large annotated corpora. In addition, the construction of the automatically annotated treebank requires the use of the High Performance Computing cluster of the RuG. At the time of writing this service is available to us free of charge, but it might be that a small fee is required in the near future.

5 STEVIN priorities

For resources:

- richly annotated monolingual Dutch corpora

Furthermore, the present proposal will develop technology to facilitate the construction of Dutch electronic lexicons

For research:

- semantic analysis (assuming that semantic analysis will typically be performed either integrated with, or on top of syntactic analysis).
- syntactic analysis

In the area of applications:

- information extraction

Furthermore, the present proposal will develop technology to facilitate the construction of the following applications:

- semantic web
- dialogue systems and Q&A solutions
- machine translation
- educational systems

6 Description of the Proposed Research Programme

The project will result in the following resources:

- A syntactically annotated, manually corrected, corpus of Dutch written text of 1 million words (dependency structures, POS-tags, lemmatization).
- An automatically syntactically annotated corpus of Dutch written text, not manually corrected (500 million words), using the same annotation guidelines. An experiment will be conducted to estimate the accuracy of this material.
- Tools for efficient retrieval of corpus elements on the basis of linguistic queries (xml database with Xpath/Xquery support)
- A number of case studies in the area of corpus linguistics, information extraction and lexicography exemplifying and evaluating the use of the linguistic search tools both with respect to the manually annotated corpus as well as the machine annotated corpus.

To be more specific, we extend the D-COI treebank to 1 million words. For 200,000 words, D-COI will provide syntactic annotations. For 500,000 words, D-COI will provide POS-tag annotation and lemmatization. Therefore, in the current project we annotate 800,000 words with dependency structures, and we annotate 500,000 words with POS-tags as well as lemmatization.

The annotation guidelines are inherited directly from the D-COI project, and essentially consist of the dependency structures and POS-tags originally proposed for the CGN project [9], [27].

6.1 Scientific Aspects and innovative power

6.1.1 Semi-automatic Annotation

The project builds on the results of the D-COI project. In that project, the syntactic annotation guidelines, inherited from CGN [9], will be specified. In addition, various tools (based on the tools used in the development of the Alpino Treebank [21]) will be adapted to enhance the semi-automatic annotation procedures. In the proposed LASSY project we extend the available syntactically annotated corpora for Dutch both in size as well as with respect to the various text genres and topical domains. In addition, various browse and search tools for syntactically annotated corpora will be further developed and made available. Their potential for applications in corpus linguistics, information extraction and lexicography will be illustrated and evaluated.

In D-COI, a syntactically annotated corpus of 200,000 words is being constructed. In the current proposal we extend this corpus to 1,000,000 words of syntactically annotated corpus material. In addition, we will extend the text-types and topical domains of the text material that will be annotated. The choice of material will be made in close cooperation with both the various project partners in D-COI as well as members of a proposed user group (described below). For instance, GridLine has an interest in the syntactic annotation of a smaller, domain-specific corpus of financial texts, on the basis of which they can build their commercial applications.

An important tool is the Alpino wide-coverage parser for Dutch. Alpino features a large dictionary, a large feature-based grammar (HPSG style), a maximum entropy disambiguation model, a HMM trigram part-of-speech tagger, etc. The estimated accuracy of the system now is about 88% concept accuracy (proportion of correct named dependencies) for newspaper texts [13]. The speed of the system is influenced heavily by the nature of the input texts, but the system has been used to annotate (fully automatically) various large corpora, including the full CLEF (Cross Language Evaluation Forum) corpus (about 75 million words) [5][6].

For interactive annotation, Alpino provides a variety of tools, including the optional interactive assignment of lexical categories, optional interactive assignment of syntactic brackets, the potential to obtain best N or all parses, a parse selection tool to select the correct parse or the best parse from a potentially large set of parses, and access to the Thistle editor, especially adapted for dependency structures, for intuitive editing CGN-type dependency structures. In addition, a number of xml-based tools are available for automatic consistency checking of the annotations, and for browsing and searching the annotations. These tools are all available for free already.

6.1.2 Fully automatic Annotation

In recent years, it has become apparent that for a wide variety of applications in information extraction, corpus linguistics and lexicography, much larger treebanks prove useful, even if these treebanks are not manually corrected. In the current proposal, we therefore aim to provide such a large automatically annotated corpus of 500,000,000 words.

Alpino is useful also for batch processing to obtain fully automatic syntactic annotation (of lesser quality, but still useful for many applications). In this context, Alpino provides the potential to obtain the best parse efficiently. Furthermore, a number of tools are available for ”error mining”, i.e., to analyze large amounts of log-files for errors, and to correct those errors [28]. In this way, Alpino can be tuned and adapted to large corpora of various types.

One motivation for using Alpino as a tool for interactive and automated syntactic annotation is that we expect that the confrontation of large amounts of corpus data will further improve the lexicon and grammar of Alpino. The manually correct syntactically annotated data is expected to further improve the disambiguation component. As a result, the quality of Alpino as a tool for syntactic analysis of Dutch is expected to increase considerably as a side-effect of the project.

6.1.3 Corpus exploitation tools

It is crucial in order to use the syntactically annotated corpora in an interesting way that there is corpus exploitation software available. In D-COI, the corpus exploitation software (COREX) was explicitly not yet extended to include the syntactic annotation layer. Therefore, we propose that in the current project an additional focus is required on the exploitation of the syntactic annotation layer. An obvious starting point is the TigerSearch tool [12] that is used in CGN. Tools based on xml/Xpath/Xquery for browsing and searching syntactically annotated corpora have been developed in the context of the Alpino Treebank. We will explore these and other existing tools, and investigate whether there are existing tools that can be used for our purposes, taking into account constraints related to IPR and standards (for instance, TigerSearch appears to be somewhat problematic from this perspective). It appears, at first sight, that these tools must be adapted to be able to use them for the particular format as well as the enormous amounts of corpus material considered here.

In the approach foreseen in D-COI as well as in LASSY, the syntactic annotation layer consists of CGN-like dependency structures, which are straightforwardly represented in xml. Using the (ISO standard) xml search language Xpath/Xquery then provides the possibility to formulate search queries for specific syntactic configurations, which might take into account surface order, syntactic dominance, the presence of specific words or root forms, and constraints on part-of-speech tag. Indeed, we have been using Xpath/Xquery successfully for a number of years for this purpose [3]. Yet, there are improvements with respect to efficiency and expressivity that we plan to provide.

For the amount of annotated corpus material considered in the current project, it is probably necessary to integrate an Xpath/Xquery search tool with xml database technology for reasons of efficiency.

We will also consider extensions to the Xpath/Xquery search language, particularly in the area of regular expressions over strings, as well as the use of 'capturing' devices to select particular parts from a matching xml document (similar to the use of `$1 $2 $3`...variables in Perl regular expression matching). This would not only allow searching for a particular syntactic categories with particular properties and in a particular context, but also allows to extract, for instance, pairs or triples of syntactic categories that appear in particular contexts. This extended functionality is very important for more flexible applications in linguistically informed information extraction.

6.1.4 Case Studies in Corpus Exploitation

The development of the machine annotated corpus and the corpus exploitation tools should proceed in parallel with some actual case studies in which the tools are used for some specific issues in corpus linguistics, information extraction and lexicography. In this way, the tools will be evaluated on basis of the results of these case studies, and will be improved. As an additional benefit, a set of concrete example case studies of the use of the tools and the syntactically annotated corpora will be very valuable for the dissemination of the resources.

In particular, the current proposal aims at three case studies, in each of the areas Information Extraction, Corpus Linguistics, and Lexicography. Here we describe a number of smaller studies, which have been performed recently, mostly based on the 75,000,000 word CLEF corpus, as well as some similar resources.

Information Extraction and Ontology Construction The CLEF corpus consists of four years of news-paper texts, and this corpus was automatically parsed using Alpino. The resulting treebank was used for a Dutch submission to the CLEF 2005 competition (monolingual Question Answering), in which the best result for Dutch was obtained [6]. The treebank was employed both for on-line question answering, as well as off-line question answering. In the latter case, answers for typical questions are collected before the question is asked, giving rise to tables consisting of e.g. capitals, causes of deaths, functions of person names, etc. [5]. It was shown [11] that the availability of syntactic annotation improves the quality of such tables considerably.

Very similar techniques are applied for information extraction and ontology building. Van der Plas and Bouma [25] apply vector-based methods to compute the semantic similarity of words, based on co-occurrence data extracted from the CLEF treebank. Their ultimate goal is the automatic extension of Dutch EuroWordNet. They show [24] that the acquired information indeed correlates with the information in Dutch EuroWordNet, and that the performance of question answering improves with such automatically acquired lexico-semantic information.

Corpus Linguistics. Large, automatically annotated corpora are useful for applications in corpus linguistics. Bouma, Hendriks and Hoeksema study a.o. the distribution of focus particles in prepositional phrases. Their corpus study on the basis of the CLEF treebank revealed that such focus particles in fact are allowed (and fairly frequent) in Dutch, contradicting claims in theoretical linguistics. Similar techniques have been applied for the study of PP-fronting in Dutch [1], the order of noun phrases with ditransitives [19], the distribution of determinerless PPs [20], the distribution of weak pronouns, the distribution of impersonal pronouns as objects of prepositions, etc.

Lexicography. In a recent thesis [16], Villada Moirón illustrates the usefulness of huge syntactically annotated corpora for various applications in semi-automatic lexicography, in particular aiming at the identification of support verb constructions and related fixed expressions in Dutch.

In the proposed project three similar case studies will be conducted. The actual topics of these studies depend on the interests of the members of the user group (described below), and will be defined in the course of the project.

6.2 Economic Aspects

It is clear that a large syntactically annotated corpus of written Dutch is an important resource for the development of a large variety of language applications. Such a resource is not only useful for the more straightforward applications such as training and testing parsers and chunkers for Dutch, but also for the extraction of various types of ontological, linguistic and lexicographic knowledge.

In order that the project develops in a direction with economic potential, we will establish a user group. The user group will consist of an advisory committee of specialists from both industry and academia with an interest in the project. The following members of the user group have indicated their interest in the project, and their willingness to participate in it:

Pim van den Broek	Ilse Media
Tigran Spaan	GridLine
Theo Vosse	F.C. Donders Centre, Radbout Universiteit Nijmegen
Erwin Marsi	Tilburg University
Jan Kleinnijenhuis	Communicatiewetenschap, Vrije Universiteit Amsterdam
Els den Os	NWO-programme Interactive Multimodal Information Extraction
Peter Spyns	STAR-Lab Vrije Universiteit Brussel
Maarten de Rijke	Informatics Institute, University of Amsterdam
Walter Daelemans	Centrum voor Nederlandse Taal & Spraak, Universiteit Antwerpen
Marie-Francine Moens	Interdisciplinary Centre for Law & ICT, Katholieke Universiteit Leuven
Ton van der Wouden	Leiden Centre for Linguistics, Universiteit Leiden

6.3 Contribution to the STEVIN programme

One of the aims of the STEVIN programme is to realize an appropriate digital language infrastructure for Dutch. The programme also intends to stimulate strategic research in the domains of language and speech technology. The construction of a balanced 500-million-word corpus of written Dutch has been identified as one of the priorities in the programme. Such a corpus is one of the prerequisites for the development of other resources, various tools, and applications. Thus the availability of a very large corpus will give a significant boost to natural language processing involving the Dutch language.

The present proposal is directed towards making available the syntactic annotation for that corpus. In addition, the proposal focuses on the construction of tools for linguistically advanced corpus exploitation, together with a set of well-documented case studies, for reasons of dissemination, and for evaluating the usefulness of automatically syntactically annotated corpora.

6.4 IPR and standards

At the end of the project, the corpus and all other results will be made available through the HLT Agency (TST-centrale). In order to ensure that the project contributes optimally to the creation of the envisaged infrastructure, we will adopt the following principles

- relate to (the results of) other projects, past and present
- (sensible) re-use of resources and tools; as far as resources or tools exist re-use should be considered; duplication is to be avoided; however, incorporation of (parts of) existing resources is considered to the extent that a significant added value can be expected, and no problems arise with respect to IPR.
- adhere to (inter)national standards, guidelines or best practice

Note that Alpino is and will remain freely available under the GPL. In cooperation with the TST-centrale we are willing to negotiate even more liberal conditions. For instance, we can make the various tools available under LGPL if that is preferred.

The proposed project clearly is dependent on the D-COI project that started June 1 2005 (note that the parties responsible for syntactic annotation in D-COI are the same as the applicants of the LASSY proposal). We will use the material collected in that project. In the ideal case, a follow-up project to D-COI is financed in the second STEVIN round, in order to meet the explicit STEVIN goal to construct a 500 million word corpus of Written Dutch. In that case, we will furthermore use the corpus material collected in this hypothetical “D-COI-2” project.

In the unfortunate case that this material is not available in time (eg., since there is no D-COI-2 project), we will use not yet syntactically annotated material from D-COI-1 for manually corrected annotation. We will also use that material for the automatic annotation, and in addition will use easily accessible further material such as the Twente News Corpus, and the material distributed by Mediargus (texts from all Flemish newspapers and the Belga press agency).

With respect to IPR, it is clear that no issue should arise with respect to the input corpora that we will annotate, since the IPR situation of those corpora will have been settled by D-COI. The syntactic annotation that we deliver, as well as the various tools that we develop will be made available for free to the general public.

If there is a need to collect new text material, we will adapt the IPR blueprint to be proposed by D-COI (deliverable F1 of D-COI).

6.5 Coordination and project management

The management of LASSY will be the responsibility of the project coordinator, who is responsible for monitoring the overall progress on the basis of regular reports from each work package, identifying any deviations from the work plan and ensuring that suitable corrective measures are implemented.

The project partners will meet three times a year to take important design decisions, to synchronize the efforts, to discuss the project’s progress and to collaborate on the evaluation. Standard best practices (web-site, group-ware, cvs, bugzilla, etc) will be used for joint development and communication between the partners, and for dissemination of the results.

The coordinator will also organize the user group, and will ensure a number of meetings with members of the project and the user group. Those meetings will be synchronized with similar efforts in other STEVIN projects.

The coordinator will stimulate the research groups to present the project results at dedicated workshops, leading conferences, and the relevant journals.

6.6 Evaluation, validation and success criteria

The evaluation of the manually corrected syntactic annotations will be performed in terms of annotator agreement. Initial experiments in creating the Alpino Treebank suggested that an agreement of about 95% can be expected, provided there is ample opportunity for annotators to discuss and synchronize particular annotation decisions.

Once a large body of manually corrected syntactically annotated material is available, this material can be regarded the gold standard, against which we can then furthermore evaluate the Alpino parser. That evaluation will then provide a fairly accurate estimate of the accuracy of the large automatically constructed treebank. At the moment, an accuracy of about 88% can be expected, but of course we hope to be able to improve this figure somewhat.

Both evaluation schemes can exploit the concept accuracy (CA) metric, which is a metric which is used to measure similarity of dependency structures. For details, we refer to [13].

Apart from both *intrinsic* evaluation dimensions, a more extrinsic evaluation will be available as a result of the various case studies in corpus linguistics, information extraction and lexicography that we foresee. Those studies will make clear to what extent the provided resources and tools are useful for such and similar applications.

In addition, an external validation of the annotations, protocols and procedures can be performed, for instance by the same agency that is responsible for the external validation of D-COI (CST). If an external validation is requested, we will conduct further negotiations with CST or similar agencies.

7 Work Programme

7.1 Work Package 1: Management

Time Span M1-M36

Partners RuG

Responsible partner RuG

7.2 Work Package 2: Manually corrected POS-tag and lemma annotation

Time Span M1-M12

Partners KuL

Responsible partner KuL

Deliverables In this work package the following deliverables are planned:

1. 250,000 words M6
2. 500,000 words M12

IPR + standards format D-COI, IPR of corpora settled in D-COI, uses free tools as in D-COI

expected risks and alternatives No risks are foreseen

7.3 Work Package 3: Manually corrected syntactic annotation

Time Span M1-M24

Partners RuG, KuL

Responsible partner RuG

Deliverables In this work package the following deliverables are planned:

1. 400,000 words; + initial evaluation (annotator agreement) M12
2. 600,000 words M18
3. 800,000 words; + final evaluation (annotator agreement) M24

IPR + standards format D-COI, IPR of corpora settled in D-COI, uses free tools

expected risks and alternatives The amount of man power dedicated to this work package (mostly the student assistents) are fairly reliable estimates based on our experience in CGN, Alpino treebank and D-COI.

7.4 Work Package 4: Automatic syntactic annotation

Time Span M12-M24

Partners RuG (syntactic annotation), KuL (POS-tag and lemmatization annotation)

Responsible partner RuG

Deliverables In this work package the following deliverables are planned:

1. Improved and extended version of Alpino, using error mining technology M18
2. 500,000,000 words automatically POS-tagged and lemmatized M24
3. 500,000,000 words automatically annotated syntactically M24
4. evaluation on a subsection from Work Package M24

IPR + standards Uses format of D-COI, IPR of corpora settled in D-COI, uses free tools

expected risks and alternatives The main risk for this work package is the lack of available corpus material. As described above, we hope and expect to be able to use the corpus material collected and prepared by D-COI and its potential successor STEVIN project. If no such successor project is funded, we will use easily accessible alternative corpora such as the Twente News Corpus, and the material distributed by Mediargus (texts from all Flemish newspapers and the Belga press agency). In any case we will be able to use the corpus preparation tools as well as the IPR policy blueprint that will be made available by D-COI.

7.5 Work Package 5: Search / Extraction tools

Time Span M1-M36

Partners RuG

Responsible partner RuG

Deliverables In this work package the following deliverables are planned:

1. feasibility study, evaluation existing tools M12
2. specification, initial implementation M18
3. implementation M24
4. improved implementation, based on experiences in work package 5 M36

IPR + standards important criterion for deliverable 1: resulting tools should be based on standard such as Xpath/Xquery, and freely available.

expected risks and alternatives no risks are foreseen

7.6 Work Package 6: Case studies

Time Span M12-M36

Partners RuG

Responsible partner RuG

Deliverables In this work package the following deliverables are planned:

1. case study I M18
2. case study II M24
3. case study III M30
4. evaluation report integrating results of the three case studies M36

IPR + standards uses the tools and corpora for which IPR + standards has been described above.

expected risks and alternatives no risks are foreseen

Relation between work packages

Work package 1 is the responsibility of the senior in Groningen. The student assistents employed in Leuven and Groningen will carry out most of the practical work for work packages 2 and 3. This work will be supervised by the senior in Leuven as well as the senior and junior in Groningen. The junior in Groningen will in addition work on the work packages 4, 5 and 6, with supervision from senior.

There are not many dependencies between work packages. The annotation of POS-tags and lemmatization is performed independently from syntactic annotation, although we do foresee that these annotation layers eventually will be merged in the final file format.

The work package on automated syntactic annotation will benefit from the availability of the manually corrected treebank (on the basis of that material, Alpino can be re-trained and is expected to perform better on the basis of more material). However, work for this work package can proceed even in the absence of that treebank.

The work packages on the search/extraction tools and the case studies obviously are interdependent, by design. The case studies are meant to provide feedback for the work on the search and extraction tools. Since the work on those packages is performed at a single location, we are confident that this interdependence will not cause any practical problems.

8 International Perspective

In the last 15 years, the Penn Treebank has become without any doubt the most widely used resource in computational linguistics. Its influence and importance for the development of an enormous variety of tools and techniques can hardly be overestimated. Although a small number of treebanks of moderate size are now available for Dutch (Alpino treebank, CGN treebank, D-COI treebank), it is obvious that a treebank of more serious proportions is a necessity for the further development of Dutch computational linguistics, in particular in an international perspective.

Traditionally, the Dutch CL community has been quite strong in parsing. The international research community has been moving into a direction in which corpora, machine learning, and statistical techniques have become more and more important. The resources that we intend to build in this project are a necessity for the Dutch CL community to take part in this development.

In more recent years, applications in information extraction and question answering are hot topics in international research. Research groups such as those in Amsterdam (Maarten de Rijke) and Groningen (Gosse Bouma) have successfully applied IE and QA techniques to Dutch. The information extraction tools that we intend to provide in the current project will be important to further extend these initiatives.

9 Short CV Principal Applicant

Gertjan van Noord is an associate professor (UHD) at the University of Groningen. He was theme-group leader of the NWO Priority Programme on Language and Speech Technology. In

1999, he received an NWO Pionier grant for research on ’Algorithms for Linguistic Processing’. He is the key architect of the Alpino parser for Dutch. Currently, van Noord is the chair of the EACL (European Chapter of the Association for Computational Linguistics).

10 Literature

10.1 Selection of Publications

[28], [13], [23], [7], [17], [26], [2], [18], [4], [29]

10.2 International Literature

We refer to [12], [10], [15], [14], [8], [9]

11 Project Budget Details

These are appended below.

References

[1] Gosse Bouma. Treebank evidence for the analysis of pp-fronting. In S. Kubler, J. Nivre, E. Hinrichs, and H. Wunsch, editors, *Third Workshop on Treebanks and Linguistic Theories*, pages 15–26, Seminar fr Sprachwissenschaft, Tbingen, 2004.

[2] Gosse Bouma, Petra Hendriks, and Jack Hoeksema. Focus particles inside prepositional phrases: A comparison between Dutch, English and German. *Journal of Comparative Germanic Linguistics*, to appear.

[3] Gosse Bouma and Gert Kloosterman. Querying dependency treebanks in XML. In *Proceedings of the Third international conference on Language Resources and Evaluation (LREC)*, pages 1686–1691, Gran Canaria, Spain, 2002.

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

[5] Gosse Bouma, Jori Mur, and Gertjan van Noord. Reasoning over dependency relations for QA. In Farah Benamarah, Marie-Francine Moens, and Patrick Saint-Dizier, editors, *Knowledge and Reasoning for Answering Questions*, pages 15–21, 2005. Workshop associated with IJCAI 05.

[6] Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann. Question answering for Dutch using dependency relations. In *Proceedings of the CLEF2005 Workshop*, 2005.

[7] Jan Daciuk and Gertjan van Noord. Finite automata for compact representation of tuple dictionaries. *Theoretical Computer Science,* 313(1):45–56, 2004.

[8] Martin Forst, Nuria Beromeu, Berthold Crysmann, Frederik Fouvry, Silvia Hansen-Schirra, and Valia Kordonia. Towards a dependency-based gold standard for german parsers – the tiger dependency bank. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, Geneva, 2004.

[9] Heleen Hoekstra, Michael Moortgat, Bram Renmans, Machteld Schouppe, Ineke Schuurman, and Ton van der Wouden. *CGN Syntactische Annotatie*, December 2003.

[10] N. Ide, P. Bonhomme, and L. Romary. XCES: An XML-based standard for linguistic corpora. In *Proceedings of LREC 2000,* pages 825–30, Athens, Greece, 2000.

[11] V. Jijkoun, J. Mur, and M. de Rijke. Information extraction for question answering: Improving recall through syntactic patterns. In *COLING 2004*, Geneva, 2004.

[12] Esther König and Wolfgang Lezius. A description language for syntactically annotated corpora. In *COLING2000*, pages 1056–1060, Saarbrücken, 2000.

[13] Robert Malouf and Gertjan van Noord. Wide coverage parsing with stochastic attribute value grammars. In *Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*, Hainan China, 2004. IJCNLP. IJCNLP-workshop; an improved version is available as http://www.let.rug.nl/~vannoord/papers/wcpsavg.pdf.

[14] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[15] Andreas Mengel and Wolfgang Lezius. An XML-based encoding format for syntactically annotated corpora. In *Proceedings of LREC 2000*, pages 121–126, Athens, Greece, 2000.

[16] Begoña Villada Moirón. *Data-driven identification of fixed expressions and their modifiability*. PhD thesis, University of Groningen, 2005.

[17] Robbert Prins and Gertjan van Noord. Reinforcing parser preferences through tagging. *Traitement Automatique des Langues,* 44(3):121–139, 2003.

[18] Ineke Schuurman, Wim Goedertier, Heleen Hoekstra, Richard Piepenbrock, and Machteld Schouppe. Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again .... In *Proceedings of the IV International Conference on Language Resources and Evaluation Volume I*, pages 57–60, Lisbon Portugal, 2004.

[19] Leonoor van der Beek. Argument order alternations in Dutch. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG'04 Conference*. CSLI Publications, 2004.

[20] Leonoor van der Beek. The extraction of Dutch determinerless PPs. In *Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, University of Essex, Colchester, 2005.

[21] Leonoor van der Beek, Gosse Bouma, Robert Malouf, and Gertjan van Noord. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands*, 2002.

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

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

[24] Lonneke van der Plas and Gosse Bouma. Automatic acquisition of lexico-semantic knowledge for QA. In *Proceedings of the IJCNLP workshop on Ontologies and Lexical Resources*, 2005. to appear.

[25] Lonneke van der Plas and Gosse Bouma. Syntactic contexts for finding semantically related words. In *CLIN 2004*, 2005. to appear.

[26] Frank van Eynde. Minor adpositions in Dutch. *Journal of Comparative Germanic Linguistics*, 7:1–58, 2004.

[27] Frank van Eynde. Part of speech tagging en lemmatisering van het D-COI corpus, 2005. Intermediate, project-internal version.

[28] Gertjan van Noord. Error mining for wide-coverage grammar engineering. In *ACL2004*, Barcelona, 2004. ACL.

[29] Gertjan van Noord and Dale Gerdemann. Finite state transducers with predicates and identities. *Grammars*, 4:263–286, 2001.