Algorithms for

LINGUISTIC PARSING

NWO PIONIER

Final Report

October 2005

Algorithms for Linguistic Processing
NWO PIONIER
Final Report

Leonoor van der Beek
Gosse Bouma
Jan Daciuk
Tanja Gaustad
Robert Malouf
Mark-Jan Nederhof
Gertjan van Noord
Robbert Prins
Begoña Villada Moirón

Alfa-informatica RuG

October, 2005

Contents

Chapter 1

Summary Project Goals

Algorithms for Linguistic Processing is a research proposal in the area of *computational linguistics*. The proposal focuses on problems of *ambiguity* and *processing efficiency* by investigating *grammar approximation* and *grammar specialization* techniques.

Theoretical linguistics has developed extensive and precise accounts of the grammatical knowledge implicit in our use of language. It has been able to adduce explanations of impressive generality and detail. These explanations account for speakers' discrimination between different linguistic structures, their ability to distinguish well-formed from ill-formed structures, and their ability to assign meaning to such well-formed structures. *Grammars* are hypothesised which model the well-formed utterances of a given natural language and the meaning representations which correspond with these utterances.

The smaller and younger field of computational linguistics has also been successful in obtaining results about the computational processing of language. These range from descriptions of dozens of concrete algorithms and architectures for understanding and producing language (parsers and generators), to careful theoretical analysis of the underlying algorithms. The theoretical analyses classify algorithms in terms of their applicability, and the time and space they require to operate correctly. The scientific success of this endeavor has opened the door to many new opportunities for applied linguistics.

However a number of important research problems have not been solved. An important challenge for computational accounts of language is the observed *efficiency* and *certainty* with which language is processed. The efficiency challenge is both theoretical and practical: grammars with transparent inspiration from linguistic theory cannot be processed efficiently. This can be demonstrated theoretically, and has been corroborated experimentally. In current practice, such grammars are recast into alternative formats, and are restricted in implementation. Effectively, large areas of language are then set aside.

The *certainty* with which language is processed is not appreciated generally. But careful implementation of wide-coverage grammars inevitably results in systems which regard even simple sentences as grammatically ambiguous, even to a high degree. The computational challenge is to incorporate disambiguation into processing.

There are two central leading hypotheses of the project. We shall explore *approximation* techniques which recast theoretically sound grammars automatically into forms which allow for efficient processing. The hypothesis is that processing models of an extremely simple type, namely finite automata, can be employed. The use of finite automata leads to interesting hypotheses about language processing, as we will argue below.

Second, we test the hypothesis that *certainty* can be accounted for—at least to some extent—by incorporating the results of language experience into processing. This will involve the application of machine learning techniques to grammars in combination with large samples of linguistic behavior, called corpora. Such techniques will ensure that a given utterance, which receives a number of competing analyses if considered in isolation, will receive a single analysis if the relevant context and situation are taken into account.

The project aims furthermore at significant partial results. In order to test its processing claims, large scale grammars of some theoretical ambition must be tested. While these exist now for English, the project will devote resources to extending existing Dutch grammars to further test the claims. An extensive Dutch grammar in the public domain would be a major contribution to Dutch computational linguistics and to the international community. Second, the processing techniques and concrete implementations are technology which directly enables a number of interesting applications in spoken language information systems, language instruction, linguistic research, grammar checking, and language aids to the disabled.

Chapter 2

Adaptations in Project Goals

The general project goals have been kept quite stable during the project. A number of changes have been made with respect to details of implementation of the original work plan.

In the original project proposal, the following sub-topics were identified:

1. Finite-state Language Processing
2. Grammar specialization (disambiguation)
3. Grammar development for Dutch
4. Linguistically-informed search tool

The first two topics were large and the last two topics were smaller, in the sense of manpower assigned to them.

It turned out that one of the reviewers for the proposed project was skeptical about the fourth item (the linguistically-informed search tool). In addition, a search tool of the type described in the original project proposal became available before the start of the project: *gsearch* (Corley et al., 2001). Furthermore, if syntactic analysis structures are annotated in XML, standard XML search languages such as XPATH are applicable, so there is no need to implement such a specialized query language. These considerations led to the conclusion that it would be better to reduce the efforts dedicated to this fourth theme, and instead focus on a different contribution to Dutch corpus linguistics. In particular, by providing corpora of syntactically annotated Dutch sentences, we provide a resource for theoretical linguists, corpus linguists and computational linguists that is at least as useful as the originally proposed tool. In addition, we provide XML-based tools to query and browse the various corpora.

Chapter 3

Results

3.1 Alpino

A natural language analysis system of Dutch has been constructed which integrates various components designed in many of the sub-projects. We also use the system to construct various training sets which are then used to improve some of the sub-components of the system (*bootstrapping*). The system constructs dependency structures as proposed by the NWO project *Corpus Gesproken Nederlands*. The system now constitutes a de facto standard for parsing of Dutch, and is freely available at `http://www.let.rug.nl/~vannoord/alp/Alpino/`.

3.2 Computational Grammar

A large, detailed, computational grammar and lexicon of Dutch has been implemented as part of the Alpino system. Aspects of the grammar are described in a number of articles ((Bouma, Malouf, and Sag, 2001), (van der Beek, Bouma, and van Noord, 2002), (Bouma, Hendriks, and Hoeksema, 2005)), a large number of contributions to conferences and workshops, and the Ph-D dissertations of Leonoor van der Beek and Begoña Villada Moirón ((van der Beek, 2005b), (Villada Moirón, 2005)).

(Villada Moirón, 2005) explores methodological and empirical issues that the computational linguist encounters when attempting to extend a computational grammar with fixed expressions. This thesis investigates two properties of fixed expressions: automatic identification and the establishment of their potential for internal modification (and other sorts of variation). These two issues are pertinent to building and updating computational lexica and also to augmenting grammars in the process of expanding the coverage of a computational parser. The experimental work applies to Dutch fixed expressions, in particular collocational prepositional phrases and support verb constructions.

(van der Beek, 2005b) presents four studies in Dutch syntax. The first study presents an analysis of the Dutch it-cleft construction. It is shown that the Dutch it-cleft construction in fact consists of two distinct constructions. The first is analyzed as a transitive construction with a final relative clause. The second construction is analyzed as an intransitive construction with an expletive pronoun in subject

position and a final complementizer clause. The account conforms to the rules of canonical word order without violating the principle of subject-verb agreement.

The second study investigates if and how the factors that are claimed to influence the English dative alternation also influence the Dutch construction. In English, the two possible realizations differ with respect to both the order of the arguments and the syntactic category of the recipient. In Dutch, word order and recipient category may vary independently. The study shows that the choice of verb lexeme influences the syntactic category of the recipient (some verbs prefer PPs, other verbs prefer NPs). The surface order, on the other hand, is determined by factors such as pronoun type and definiteness.

The third study shows that the syntactically marked combination of a preposition and a bare count noun (determinerless PP or PP-D) may be the result of various different syntactic constructions. These constructions differ in productivity and modifiability. We indicated how each of these constructions could be accounted for in a grammar, given the information about which preposition and which noun may participate in a PP-D and to what extend the combination allows modification. However, this information is generally not available. It is then shown that with the help of an automatically parsed corpus and various simple statistic measures, we can extract lists of PP-Ds of particular types and their modification potential semi-automatically. The quality of this extraction and classification method heavily depends on the availability of accurate noun countability information. In the final study, a number of experiments are performed to extract such countability information automatically from existing resources.

The general trend that can be observed in these contributions on computational grammar is the combination of linguistic detail (traditionally associated with theoretical linguistics) on the one hand, and the exploitation of corpus material on the other hand. Both manually syntactically annotated corpora, as well as automatically syntactically annotated corpora are used in these studies.

3.3 Annotation Efforts

A number of tools have been implemented to facilitate the construction of various corpora of syntactically annotated Dutch sentences, using the syntactic annotation guidelines of the Corpus Gesproken Nederlands (CGN). The various tools were described in a number of workshop and conference papers. The various treebanks as well as the tools remain freely available at `http://www.let.rug.nl/~vannoord/trees/`.

In addition, a huge treebank has been automatically constructed using the Alpino parser (four years of newspaper text), and we have started to use such machine-annotated treebanks for applications in information extraction, corpus linguistics, and lexicography.

3.4 Finite State Language Processing

We have worked on a variety of topics in the context of finite-state natural language processing. In a series of papers, including an article in Computational Linguistics (Daciuk et al., 2000), Jan Daciuk has established a number of results in the area of efficient construction of compact finite state automata. These results are relevant for the construction and use of very large dictionaries and language models, as described in an article in *Theoretical Computer Science* (Daciuk and van Noord, 2004).

Robbert Prins established that a finite-state preprocessing stage can greatly improve parsing effiency. He uses a POS-tagger to filter unlikely lexical categories. Although the tagger is trained on annotated data, the approach does not require manual annotation: all annotated data is constructed by the parser itself, giving rise to a successfull example of bootstrapping. The POS-tagger is an integrated part of the Alpino system, and essentially makes it possible to use Alpino in practice. Prins showed that the use of this filter is extremely effective: average parse times are up to twenty times shorter (!), whereas in addition a small increase in accuracy is observed. The main results are described in a journal article (Prins and van Noord, 2003) and Robbert's Ph.D dissertation.

We also continued our research on the finite-state calculus, using and extending the Finite State Utilities toolbox. (van Noord and Gerdemann, 2001) presents a generalization of finite state automata in which labels represent predicates over symbols. This generalization is motivated by problems in natural language processing.

A finite-state implementation of *Optimality Theory* phonology is described in (Gerdemann and van Noord, 2000). The paper was presented as a keynote lecture at the SIGPHON Finite State Phonology workshop in 2000.

(Bouma, 2003) describes finite-state solutions to the problem of hyphenation.

Mark-Jan Nederhof established a number of theoretical results on specific finite-state formalisms, including IDL expressions (Nederhof and Satta, 2004a) and non-recursive context-free grammars (Nederhof and Satta, 2004b).

With Lauri Karttunen and Kimmo Koskenniemi, Gertjan van Noord chaired the ESSLLI workshop on *Finite State Methods in Natural Language Processing*. In addition, a related special event was organized entitled *20 years of two-level morphology*. The workshop resulted in a special issue of the journal *Natural Language Engineering* (Karttunen, Koskenniemi, and van Noord, 2003).

3.5 Disambiguation

Tanja Gaustad investigated the influence of a variety of linguistic features for word sense disambiguation. Her Ph.D thesis shows, a.o., that the exploitation of automatically extracted syntactic dependencies improves word sense disambiguation systems (Gaustad, 2004).

The Alpino system features a maximum-entropy based disambiguation component. The system is trained on the Alpino treebank, using an implementation of various log-linear model training algorithms developed by Rob Malouf (Malouf,

2002). Malouf showed that certain more general methods that are being employed in other areas of computing perform much better than the traditional IIS algorithm. The disambiguation component is embedded in a beam search algorithm for selecting the best parse for a given sentence (Malouf and van Noord, 2004).

Mark-Jan Nederhof investigated properties of statistical parsing in a series of papers and journal articles ((Nederhof, 2003), (Nederhof, 2005)).

3.6 Quantitative Evaluation of Alpino

The availability of annotated corpus material enables us to monitor the progress of the Alpino system in a quantitative way. In the first published result of the Alpino system (Bouma, van Noord, and Malouf, 2001), f-scores (combining precision and recall of dependency relations) were reported over all sentences of up to twenty words of the corpus (in as far as these were annotated at the time). The best reported score for this sub-corpus was 75%. These experiments were performed in the spring of 2001.

In the mean-time we discovered that *concept accuracy* is a somewhat more reliable metric. Such concept accuracy scores are always somewhat lower than f-scores. Even so, the version of Alpino that we reported on in the mid-term evaluation report obtained an accuracy score of 82.5% (using the default settings) on a random sample of (annotated) sentences of up to twenty words.

The current version of Alpino obtains an accuracy score of 88% for sentences of all lengths (result on the full Alpino treebank; f-score: 88.4%). For a random sample of annotated sentences of up to twenty words, the system obtains an accuracy of over 91.8% (f-score: 92.2%).

3.7 Dissemination

In the summer of 2001, Gertjan van Noord initiated and co-organized with Kimmo Koskenniemi and Lauri Karttunen the Finite State Methods in Natural Language Processing workshop, at the occassion of ESSLLI in Helsinki. The workshop included presentations by project members Gosse Bouma and Jan Daciuk. On the basis of the workshop, a special issue of *Natural Language Engineering*, edited by the workshop chairs, with a selection of extended versions of the workshop papers appeared in 2003.

The thirteenth 'CLIN' meeting, the yearly meeting of computational linguists in the Netherlands, was organized by Tanja Gaustad in 2002. It turned out to be a very succesful meeting (more than 100 participants) which included in the programme an invited presentation by computational linguist Hugo Brandt Corstius. During the CLIN meeting, van Noord handed the first copy of the Alpino Treebank CDROM to Hugo Brandt Corstius; all other participants of the CLIN symposium received a copy of the CDROM too.

The end of the PIONIER project was celebrated with a special one-day workshop in Groningen on Friday November 11, 2005.

A number of international researchers were invited for presentations in the CLCG Linguistic Colloquium series. These included John Carroll, Rob Koeling, Franck Thollard, Nathan Vaillete, Giorgio Satta and Annie Zaenen.

The automatic syntactic analysis methods develoded in the project were the topic of one edition of *Adams Appel*, a television programme for regional (Groningen, Drenthe) television. The broadcast can be viewed on-line at `mms://wmvideo.service.rug.nl/adamsappel/050514.wmv`.

Chapter 4

Outlook

4.1 New Opportunities with Huge Machine-annotated Corpora

Using the insights and techniques developed in the *Algorithms for Linguistic Processing* project, it is now possible to construct syntactic annotations for huge corpora fully automatically. The general quality of such huge machine-annotated resources appears to be adequate for many applications in information extraction, automatic ontology construction, corpus linguistics, and lexicography.

In such a set-up, Alpino is used for batch processing to obtain fully automatic syntactic annotation. 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 (van Noord, 2004). In this way, Alpino can be tuned and adapted to large corpora of various types.

We describe a small number of recent studies mostly based on the 75,000,000 word CLEF corpus. These studies illustrate and motivate our claim that such huge machine-annotated resources provide for new exciting opportunities.

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 (Bouma et al., 2005). The treebank was employed both for online 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. (Bouma, Mur, and van Noord, 2005). It was shown (Jijkoun, Mur, and de Rijke, 2004) 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 (van der Plas and Bouma, 2005b) 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 au-

tomatic extension of Dutch EuroWordNet. They show (van der Plas and Bouma, 2005a) 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 (Bouma, 2004), the order of noun phrases with ditransitives (van der Beek, 2004), the distribution of determinerless PPs (van der Beek, 2005a), the distribution of weak pronouns, the distribution of impersonal pronouns as objects of prepositions, etc.

Lexicography. In her thesis (Villada Moirón, 2005), 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.

4.2 Follow-up projects

The linguistic analysis technology developed in the *Algorithms for Linguistic Processing* project is now being employed directly in a number of recent projects. The most important of these projects are *IMIX* and *STEVIN*.

In IMIX Alpino is being employed in two sub-projects. Erwin Marsi uses Alpino in his work on sentence fusion in the context of answer generation. Alpino is also used in the QADR (Question Anwering with Dependency Relations) sub-project. Here, Alpino is used in a question-answering system both for analysing the question, as well as analysing candidate answers. The system is described in (Bouma, Mur, and van Noord, 2005). In the CLEF 2005 Question Answering competition, the system performed very well (3rd overall, best result for Dutch) (Bouma et al., 2005).

The Alpino parser is also an important ingredient in some of the STEVIN projects. In the D-COI annotation project, syntactic annotations are constructed using the tools developed in the *Algorithms for Linguistic Processing* project. Alpino is also part of the IRME project on identification and representation of multi-word-units, extending the techniques pioneered by Villada Moirón in her thesis. Finally, the COREA project develops a coreference resolution system for Dutch, based on syntactic information provided by Alpino.

The faculty of arts of the RuG promised to fund two researchers, at the end of the ALP project, who will continue to do research in the direction of ALP. The two researchers (an AIO and a PostDoc) have not yet been funded, due to financial problems of the faculty. It is hoped that in 2006 this promise can be implemented with only a slight delay.

References

[van der Beek2004] van der Beek, Leonoor. 2004. Argument order alternations in Dutch. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG'04 Conference*. CSLI Publications.

[van der Beek2005a] van der Beek, Leonoor. 2005a. 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.

[van der Beek2005b] van der Beek, Leonoor. 2005b. *Topics in Corpus Based Dutch Syntax*. Ph.D. thesis, University of Groningen.

[van der Beek, Bouma, and van Noord2002] van der Beek, Leonoor, Gosse Bouma, and Gertjan van Noord. 2002. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7(4):353–374.

[Bouma2003] Bouma, Gosse. 2003. Finite state methods for hyphenation. *Natural Language Engineering*, 9(1).

[Bouma2004] Bouma, Gosse. 2004. 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 für Sprachwissenschaft, Tübingen.

[Bouma, Hendriks, and Hoeksema2005] Bouma, Gosse, Petra Hendriks, and Jack Hoeksema. 2005. Focus particles inside prepositional phrases. *Journal of Comparative Germanic Linguistics*. to appear.

[Bouma, Malouf, and Sag2001] Bouma, Gosse, Rob Malouf, and Ivan Sag. 2001. Satisfying constraints on adjunction and extraction. *Natural Language and Linguistic Theory*, 19:1–65.

[Bouma, Mur, and van Noord2005] Bouma, Gosse, Jori Mur, and Gertjan van Noord. 2005. 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. Workshop associated with IJCAI 05.

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

[Bouma, van Noord, and Malouf2001] Bouma, Gosse, Gertjan van Noord, and Robert Malouf. 2001. Wide coverage computational analysis of Dutch. In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, editors, *Computational Linguistics in the Netherlands 2000*.

[Corley et al.2001] Corley, Steffan, Martin Corley, Frank Keller, Matthew W. Crocker, and Shari Trewin. 2001. Finding syntactic structure in unparsed corpora. *Computers and the Humanities*, 35(2):81–94.

[Daciuk et al.2000] Daciuk, Jan, Stoyan Mihov, Bruce W. Watson, and Richard E. Watson. 2000. Incremental construction of minimal acyclic finite-state automata. *Computational Linguistics*, 26(1):3–16.

[Daciuk and van Noord2004] Daciuk, Jan and Gertjan van Noord. 2004. Finite automata for compact representation of tuple dictionaries. *Theoretical Computer Science*, 313(1):45–56.

[Gaustad2004] Gaustad, Tanja. 2004. *Linguistic Knowledge and Word Sense Disambiguation*. Ph.D. thesis, University of Groningen.

[Gerdemann and van Noord2000] Gerdemann, Dale and Gertjan van Noord. 2000. Approximation and exactness in finite state optimality theory. In Jason Eisner, Lauri Karttunen, and Alain Thériault, editors, *Finite-State Phonology. Proceedings of the Fifth Workshop of the ACL SPecial Interest Group in Computational Phonology*, pages 34–45, Luxembourg.

[Jijkoun, Mur, and de Rijke2004] Jijkoun, V., J. Mur, and M. de Rijke. 2004. Information extraction for question answering: Improving recall through syntactic patterns. In *COLING 2004*, Geneva.

[Karttunen, Koskenniemi, and van Noord2003] Karttunen, Lauri, Kimmo Koskenniemi, and Gertjan van Noord. 2003. Special issue: Finite state methods in language language processing. *Natural Language Engineering*, 9(1).

[Malouf2002] Malouf, Robert. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, Taiwan.

[Malouf and van Noord2004] Malouf, Robert and Gertjan van Noord. 2004. Wide coverage parsing with stochastic attribute value grammars. In *Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*, Hainan China. IJCNLP. IJCNLP-workshop; an improved version is available as http://www.let.rug.nl/~vannoord/papers/wcpsavg.pdf.

[Nederhof2003] Nederhof, Mark-Jan. 2003. Weighted deductive parsing and knuth's algorithm. *Computational Linguistics*, 29(1):135–143.

[Nederhof2005] Nederhof, Mark-Jan. 2005. A general technique to train language models on language models. *Computational Linguistics*, 31(2):173–185.

[Nederhof and Satta2004a] Nederhof, Mark-Jan and Giorgio Satta. 2004a. Idl-expressions: A formalism for representing and parsing finite languages in natural language processing. *Journal of Artificial Intelligence Research*, 21:287–317.

[Nederhof and Satta2004b] Nederhof, Mark-Jan and Giorgio Satta. 2004b. The language intersection problem for non-recursive context-free grammars. *Information and Computation*, 192(2):172–184.

[van Noord2004] van Noord, Gertjan. 2004. Error mining for wide-coverage grammar engineering. In *ACL2004*, Barcelona. ACL.

[van Noord and Gerdemann2001] van Noord, Gertjan and Dale Gerdemann. 2001. Finite state transducers with predicates and identities. *Grammars*, 4:263–286.

[van der Plas and Bouma2005a] van der Plas, Lonneke and Gosse Bouma. 2005a. Automatic acquisition of lexico-semantic knowledge for QA. In *Proceedings of the IJCNLP workshop on Ontologies and Lexical Resources*. to appear.

[van der Plas and Bouma2005b] van der Plas, Lonneke and Gosse Bouma. 2005b. Syntactic contexts for finding semantically related words. In *CLIN 2004*. to appear.

[Prins and van Noord2003] Prins, Robbert and Gertjan van Noord. 2003. Reinforcing parser preferences through tagging. *Traitement Automatique des Langues*, 44(3):121–139.

[Villada Moirón2005] Villada Moirón, Begoña . 2005. *Data-driven identification of fixed expressions and their modifiability*. Ph.D. thesis, University of Groningen.

Appendix A

Publications

A.1 Journal Articles

Leonoor van der Beek, Gosse Bouma, Gertjan van Noord. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde* 7 (2), 2002, pp. 353–374.

E. Bertsch and Mark-Jan Nederhof. Fast parallel recognition of LR language suffixes. *Information Processing Letters*, 92:225-229, 2004.

Gosse Bouma. Finite state methods for hyphenation. *Natural Language Engineering* 9 (1) 2003.

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.

Gosse Bouma and Ineke Schuurman. Naar een digitale bibliotheek voor de taalkunde. *Nederlandse Taalkunde* 5 (2). 2000. pp. 177–180.

Gosse Bouma and Ineke Schuurman. De digitale infrastructuur van het Nederlands. *Nederlandse Taalkunde* 5 (2). 2000. pp. 90–94.

Gosse Bouma, Robert Malouf and Ivan Sag, Satisfying Constraints on Extraction and Adjunction, In: *Natural Language and Linguistic Theory*, 19, 2001, pp. 1–65.

Jan Daciuk and S. Mihov and B. Watson and R. Watson, Incremental Construction of Minimal Acyclic Finite-state Automata. *Computational Linguistics*, 26 (1). 2000. pp. 3–16.

Jan Daciuk and Gertjan van Noord. Finite Automata for Compact Representation of Tuple Dictionaries. *Theoretical Computer Science* 313 (1). 2004. pp. 45–56.

Mark-Jan Nederhof. Weighted deductive parsing and Knuth's algorithm. *Computational Linguistics* 29 (1). 2003. pp. 135–143.

Mark-Jan Nederhof. A general technique to train language models on language models. *Computational Linguistics*. 31 (2). pp. 173–185, 2005.

Mark-Jan Nederhof and Giorgio Satta. The language intersection problem for non-recursive context-free grammars, *Information and Computation*, 192 (2), pp. 172–184, 2004.

Mark-Jan Nederhof and Giorgio Satta. IDL-expressions: A formalism for representing and parsing finite languages in natural language processing. *Journal of Artificial Intelligence Research*, 21, pp. 287–317, 2004.

Gertjan van Noord, Treatment of Epsilon Moves in Subset Construction. *Computational Linguistics* 26 (1). 2000. pp. 61–76.

Gertjan van Noord and Dale Gerdemann. Finite State Transducers with Predicates and Identity. *Grammars* 4 (3). 2001. pp. 263–286.

Robbert Prins and Gertjan van Noord. Reinforcing Parser Preferences through Tagging. *Traitement Automatique des Langues* 44 (3) 2003, pp 121–139.

A.2 Ph.D. Theses

Leonoor van der Beek, Topics in Corpus Based Dutch Syntax. Ph.D. Thesis. University of Groningen 2005.

Tanja Gaustad, Linguistic Knowledge and Word Sense Disambiguation. Ph.D. Thesis. University of Groningen 2005.

Begoña Villada Moirón, Data-driven Identification of Fixed Expressions and Their Modifiability. Ph.D. Thesis. University of Groningen 2005.

Tony Mullen, An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection. Ph.D. Thesis. University of Groningen 2002.

Robbert Prins, Finite-State Pre-Processing for Natural Language Analysis. Ph.D. Thesis. University of Groningen 2005.

A.3 Edited Volumes

Tanja Gaustad (editor). Computational Linguistics in the Netherlands 2002. Selected Papers from the Thirteenth CLIN Meeting. Number 47 in "Language and Computers: Studies in Practical Linguistics", Rodopi: Amsterdam. 2003.

Jean-Claude Junqua and Gertjan van Noord (editors), *Robustness in Language and Speech Technology.* Kluwer. 2001. ISBN 0-7923-6790-1

Lauri Karttunen, Kimmo Koskenniemi, Gertjan van Noord (editors), *Finite State Methods in Natural Language Processing. FSMNLP 2001.* Extended Abstracts. ESSLLI Workshop, Helsinki. 2001.

Lauri Karttunen, Kimmo Koskenniemi and Gertjan van Noord, Special issue: Finite State Methods in Natural Language Processing.*Natural Language Engineering.* 9 (1). 2003.

A.4 Reviewed Book Chapters

Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger and Ivan A. Sag, 2005, In Search of a Systematic Treatment of Determinerless PPs, in Patrick Saint-Dizier (ed.) *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*, Springer, 2005.

Leonoor van der Beek and Timothy Baldwin, Crosslingual Countability Classification with EuroWordNet, in *Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands* (CLIN 2003), Antwerp, Belgium, 2004.

Leonoor van der Beek, Gosse Bouma, Robert Malouf, Gertjan van Noord. The Alpino Dependency Treebank. *Computational Linguistics in the Netherlands 2001*.

Gosse Bouma. A Corpus Investigation of PP-fronting in Dutch, In: Decadt, B. , Hoste, V. , Pauw, G. de (red.), *Computational Linguistics in the Netherlands 2003*, Universiteit Antwerpen, Antwerpen, 2004, pp. 15-30

Gosse Bouma. Verb clusters and the scope of adjuncts in Dutch. In Pieter A. M. Seuren and Gerard Kempen, editors, *Verb Constructions in German and Dutch*. Benjamins, Amsterdam and Philadelphia, 2003.

Gosse Bouma and Frank van Eynde and Dan Flickinger, Constraint-based Lexica. In F. van Eynde and D. Gibbon (eds), *Lexicon Development for Speech and Language Processing*. Kluwer 2000. pp. 43–76.

Gosse Bouma, Gertjan van Noord, Robert Malouf. Alpino: Wide Coverage Computational Analysis of Dutch. In: W. Daelemans, K. Sima'an, J. Veenstra, J. Zavrel (eds), *Computational Linguistics in the Netherlands CLIN 2000*. Rodopi, Amsterdam, 2001. pp. 45–59.

Gosse Bouma and Begoña Villada Moirón. Corpus-based acquisition of collocational prepositional phrases. *Computational Linguistics in the Netherlands CLIN 2001*. To appear.

Gosse Bouma. Argument realization and Dutch R-Pronouns: Solving Bech's problem without movement or deletion. In Ronnie Cann, Claire Grover, and Philip Miller (eds), *Grammatical Interfaces in Head-driven Phrase Structure Grammar*. CSLI Publications, 2000.

Jan Daciuk, Treatment of Unknown Words, In: O. Boldt, and H. Jörgensen, *Automata Implementation*, 4th International Workshop on Implementing Automata, WIA '99 Potsdam, Germany, July 1999, Revised Papers, Springer Verlag, Berlin-New York, etc. 2001, pp. 71–80.

Tanja Gaustad and Gosse Bouma. Accurate Stemming of Dutch for Text Classification. *Computational Linguistics in the Netherlands CLIN 2001*. To appear.

Robert Malouf, Cooperating Constructions. In: E. Francis and L. Michaelis (eds.), *Linguistic Mismatch: Scope and Theory*. CSLI Publications. 2002.

Mark-Jan Nederhof and Giorgio Satta. Tabular Parsing. In C. Martin-Vide, V. Mitrana, and G. Paun, editors, *Formal Languages and Applications, Studies in Fuzziness and Soft Computing 148*, pages 529-549. Springer, 2004.

Gertjan van Noord, Dale Gerdemann, An Extendible Regular Expression Compiler for Finite-state Approaches in Natural Language Processing. In: O.Boldt, H.Juergensen (eds), *Automata Implementation. 4th International Workshop on Implementing Automata, WIA '99, Potsdam Germany, July 1999, Revised Papers*. Springer Lecture Notes in Computer Science 2214. 2000.

Gertjan van Noord, Robust Parsing of Word Graphs. In: Jean-Claude Junqua and Gertjan van Noord (editors), *Robustness in Language and Speech Technology.* Kluwer. 2001. ISBN 0-7923-6790-1

A.5 Reviewed Conference Proceedings

Timothy Baldwin and Leonoor van der Beek. The Ins and Outs of Dutch Noun Countability Classification, in *Proceedings of the 2003 Australasian Language Technology Workshop* (ALTW2003), Melbourne, Australia, pp. 33-40, 2003.

Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger and Ivan A. Sag, In Search of a Systematic Treatment of Determinerless PPs, in *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Toulouse, France, 2003.

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, UK.

Leonoor van der Beek, Argument Order Alternations in Dutch, in Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG'04 Conference*, CSLI Publications, 2004.

Leonoor van der Beek, The Dutch cleft constructions, in Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG '03 Conference*, CSLI Publications, 2003.

Leonoor van der Beek and Gerlof Bouma, The Role of the Lexicon in Optimality Theoretic Syntax, in Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG'04 Conference*, CSLI Publications, 2004.

Gosse Bouma, A finite state and data-oriented method for grapheme to phoneme conversion. In: *NAACL 2000*, pp 303–310. 2000.

Gosse Bouma, Extracting Dependency Frames from Existing Lexical Resources, In: Moldovan, D , e.a., (red.), *WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh 3-4 June, 2001, NAACL 2001 Workshop, Association for Computational Linguistics, 2001, pp. 65-70

Gosse Bouma, Finite State Methods for Hyphenation, In: Lauri Karttunen, Kimmo Koskenniemi, Gertjan van Noord (eds), *Finite State Methods in Natural Language Processing*, Extended Abstracts, ESSLLI Workshop, August 20-24 2001, Helsinki, University of Helsinki, Helsinki, 2001, pp. 29-33.

Gosse Bouma. Treebank evidence for the analysis of PP-fronting, In: Kubler, S. , Nivre, J. , Hinrichs, E , Wunsch, H. (red.), Third Workshop on Treebanks and Linguistic Theories, Seminar fr Sprachwissenschaft, Tbingen, 2004, pp. 15-26

Gosse Bouma and Geert Kloosterman. Querying dependency treebanks in XML. In *Proceedings of the Third international conference on Language Resources and Evaluation* (LREC). Gran Canaria. 2002.

Gosse Bouma, Jori Mur and Gertjan van Noord. Reasoning over Dependency Relations for QA. In: Farah Benamarah, Marie-Francine Moens, and Patrick Saint-Dizier (eds.), Knowledge and Reasoning for Answering Questions, Workshop associated with IJCAI 05, pp. 15-21.

Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann. Question Answering for Dutch using Dependency Relations. In: CLEF2005 workshop.

Jan Daciuk and Gertjan van Noord, Finite Automata for Compact Representation of Language Models in NLP, In: Bruce Watson and Derick Wood (eds.), *Proceedings of the 6th Conference on Implementations and Applications on Automata*, 23–25 July 2001, Pretoria, Republic of South Africa, University of Pretoria, Dept. of Computer Science, Pretoria, 2001, pp. 45–55.

Jan Daciuk, Computer-Assisted Enlargement of Morphological Dictionaries, In: Lauri Karttunen, Kimmo Koskenniemi and Gertjan van Noord (eds.), *Finite State Methods in Natural Language Processing,* Extended Abstracts, ESSLLI Workshop, August 20-24 2001, Helsinki, University of Helsinki, Helsinki, 2001, pp. 23-27

Jan Daciuk, Experiments with Automata Compression. In: M. Daley, M. Eramian and S. Yu (eds), *CIAA*, University of Western Ontario, London, Canada. 2000. pp. 113–119.

Jan Daciuk. Finite State Tools. In: R. Zajac (ed), COLING Workshop *Using Toolsets and Architectures to build NLP Systems*. 2000. pp. 34–37.

Jan Daciuk, Comparison of Construction Algorithms for Minimal, Acyclic, Deterministic, Finite-State Automata from Sets of Strings, *Seventh International Conference on Implementation and Application of Automata* CIAA '2002, Tours, France, 2002.

Frederic Fouvry, Valia Kordoni and Gertjan van Noord. Object-to-Subject Raising: An Analysis of the Dutch Passive. In: HPSG2005, Lisbon 2005.

Tanja Gaustad, Statistical Corpus-Based Word Sense Disambiguation: Pseudowords vs. Real Ambiguous Words, In: Miltsakaki, E , e.a., (eds.), *39th Annual Meeting aand 10th Conference of the European Chapter, Companion Volume to the Proceedings of the Conference: Proceedings of the Student Research Workshop and Tutorial Abstracts*, CNRS , Toulouse, 2001, pp. 61–66.

Tanja Gaustad, Extraktion und Verifikaton von Subkategorisierungsmustern für Französische Verben. In: U. Heid, S. Evert, E. Lehmann, C. Rohrer (eds), *Proceedings of the Ninth Euralex International Congress*. University of Stuttgart. 2000. pp. 611–617.

Tanja Gaustad. A Lemma-Based Approach to a Maximum Entropy Word Sense Disambiguation System for Dutch. *Proceedings of the 20th International Conference on Computational Linguistics* (Coling 2004), Geneva, Switzerland, pp. 778-784. 2004.

Tanja Gaustad. The Importance of High Quality Input for WSD: An Application-Oriented Comparison of Part-of-Speech Taggers. *Proceedings of the Australasian Language Technology Workshop* (ALTW 2003), Melbourne, Australia, pp. 65-72. 2003.

Dale Gerdemann and Gertjan van Noord. Approximation and Exactness in Finite State Optimality Theory. In: Jason Eisner, Lauri Karttunen, Alain Thriault (editors), *SIGPHON 2000, Finite State Phonology. Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology.* August 2000, Luxembourg.

Robert Malouf. Markov models for language-independent named entity recognition. In *Proceedings of the Sixth Conference on Natural Language Learning* (CoNLL-2002). Taiwan. To appear.

Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning* (CoNLL-2002). Taiwan. To appear.

Robert Malouf and Gertjan van Noord, Wide Coverage Parsing with Stochastic Attribute Value Grammars. In: IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses. Hainan 2004.

Tony Mullen, Robert Malouf, and Gertjan van Noord, Statistical Parsing of Dutch using Maximum Entropy Models with Feature Merging, In: Tsujii, J (eds.), *NLPRS2001, Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, November 27–30, 2001, University of Tokyo Press, Tokyo, 2001, pp. 481–486.

Mark-Jan Nederhof and Giorgio Satta. IDL-expressions: a compact representation for finite languages in generation systems. *Proceedings of FG Trento*, The 7th Conference on Formal Grammar, pages 125-136, Trento, Italy, 2002.

Mark-Jan Nederhof and Giorgio Satta. Probabilistic parsing strategies. In J. Dassow et al., editors, *Descriptional Complexity of Formal Systems* (DCFS), Pre-Proceedings of a Workshop, pages 216-230, London, Canada, August 2002.

Mark-Jan Nederhof and Giorgio Satta. Parsing non-recursive context-free grammars. *40th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 112-119, Philadelphia, Pennsylvania, USA, July 2002.

Mark-Jan Nederhof and Giorgio Satta. Probabilistic parsing strategies. *Proceedings of the 3rd AMAST Workshop on Algebraic Methods in Language Processing* (AMiLP 2003), pages 19-32, Verona, Italy, 2003.

Mark-Jan Nederhof and Giorgio Satta. Probabilistic parsing as intersection. *8th International Workshop on Parsing Technologies*, pages 137-148, Nancy, France, April 2003.

Mark-Jan Nederhof and Giorgio Satta. Kullback-Leibler distance between probabilistic context-free grammars and probabilistic finite automata. *20th International*

Conference on Computational Linguistics Computational Linguistics, Proceedings of the Conference, volume 1, pages 71-77, Geneva, Switzerland, August 2004.

Mark-Jan Nederhof and Giorgio Satta. An alternative method of training probabilistic LR parsers. *42th Annual Meeting of the Association for Computational Linguistics*, Proceedings of the Conference, pages 551-558, Barcelona, Spain, July 2004.

Mark-Jan Nederhof and Giorgio Satta. Probabilistic Parsing Strategies. *42th Annual Meeting of the Association for Computational Linguistics*, Proceedings of the Conference, pages 543-550, Barcelona, Spain, July 2004.

Mark-Jan Nederhof, Giorgio Satta and Stuart Shieber. Partially ordered multiset context-free grammars and free-word-order parsing. *8th International Workshop on Parsing Technologies*, pages 171-182, Nancy, France, April 2003.

Gertjan van Noord. Error Mining for Wide-Coverage Grammar Engineering. ACL 2004, Barcelona.

Robbert Prins. Beyond N in N-gram Tagging. *Proceedings of the ACL 2004 Student Research Workshop*, pages 61–66, Barcelona, Spain, 2004

Robbert Prins and Gertjan van Noord, Unsupervised pos-tagging improves parsing accuracy and parsing efficiency, In: Bunt, H (eds), *Proceedings of the Seventh International Workshop on Parsing Technologies* - IWPT - 2001, 17–19 October, 2001 Peking University, Beijing, China, Tsinghua University Press, Beijing, 2001, pp. 154–165.

Begoña Villada Moirón. Distinguishing prepositional complements from fixed arguments. *Proceedings of the 11th EURALEX International Congress*. Vol. III, pp. 935-942. Lorient, France. 2004.

Begoña Villada Moirón. Discarding noise in an automatically acquired lexicon of support verb constructions. *Proceedings of the 4th International Conference on Language Resources and Evaluation* LREC 2004. Vol. V, pp 1859–1862. Lisbon, Portugal. 2004.

Begoña Villada Moirón, Linguistically enriched corpora for establishing variation in support verb constructions. *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora (LINC'05) held at The 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*. Jeju Island, Republic of Korea. 2005.

Begoña Villada Moirón and Gosse Bouma. A corpus-based approach to the acquisition of collocational prepositional phrases. In: EURALEX 2002, Copenhagen, August 2002. To appear.

A.6 Other Publications

Gosse Bouma. Unification: Classical and Default. In Keith Brown (ed.), Encyclopedia of Language and Linguistics, 2nd edition. Elsevier. 2005.

Gertjan van Noord, Finite State Language Processing. In Lynn Nadel (editor-in-chief), Encyclopedia of Cognitive Science. Nature Publishing Group. pp 130–134.

Appendix B

Presentations

Leonoor van der Beek, Gosse Bouma, Robert Malouf and Gertjan van Noord. The Alpino Dependency Treebank. *Empirical methods in the new millennium: Linguistically Interpreted Corpora* (LINC 2001), Leuven, August 29 2001.

Leonoor van der Beek, The Alpino Dependency Treebank; three tools for treebanking, CLIN 2001, Twente University, Enschede, November 30 2001.

Leonoor van der Beek, Cleft Sentences. BCN Poster Day. February 2002, Groningen.

Leonoor van der Beek, Dutch pronouns in copular constructions, WOTS7 (conference/workshop), Nijmegen, October 27 2003.

Leonoor van der Beek, Crosslingual countability classification with EuroWordnet, CLIN 2003 (conference), Antwerp, December 19 2003

Leonoor van der Beek, Argument Order Variations in Dutch. LFG 2004, Christchurch, New Zealand. July 10, 2004.

Beek, L J van der, The Dutch Cleft Constructions, LFG 2003, Saratoga Springs, NY, USA, July 17 2003

Gosse Bouma, Finite State and Data-Oriented Methods for Grapheme to Phoneme Conversion, NAACL 2000, Seattle. May 2000.

Gosse Bouma, Gertjan van Noord, Alpino: A Wide Coverage Computational Grammar of Dutch. CLIN, Tilburg, November 3 2000.

Gosse Bouma, Alpino, A Wide Coverage Computational Grammar for Dutch, LOT winterschool, University of Amsterdam, January 15 2001 [invited].

Gosse Bouma, Extracting Dependency Frames from Existing Lexical Resources, WordNet and Other Lexical Resources Workshop, Pittsburgh, USA, June 3 2001.

Gosse Bouma, Finite State Methods for Hyphenation, Finite State Methods in Natural Language Processing, Helsinki, Finland, August 20-24 2001.

Gosse Bouma, The Alpino Parser and Treebank, Invited Lecture, NLP course, Master of AI programme, Catholic University of Leuven, February 8 2003 [invited]

Gosse Bouma, De Alpino Parser en Treebank, Workshop Corpus Gesproken Nederlands, University of Nijmegen, May 16 2003 [invited]

Gosse Bouma, Introduction to Computational Linguistics, Linguistic Society of America Linguistic Institute, University of Michigan, East-Lansing, MI, June 30–July 18 2003 [invited]

Gosse Bouma, A corpus-investigation of PP-fronting in Dutch, Computational Linguistics in the Netherlands, University of Antwerp, December 19 2003

Gosse Bouma, Word Order and Scope of Adjuncts in Dutch, 10th Conference on Head-driven Phrase Structure Grammar, University of Michigan, East-Lansing, MI, July 18 2003

Gosse Bouma, Treebank Evidence for the Analysis of PP fronting. The 4th Workshop on Treebanks and Linguistic Theories. University of Tbingen, Germany. December 11, 2004.

Gosse Bouma and Geert Kloosterman. Querying dependency treebanks in XML. Poster at the Third international conference on Language Resources and Evaluation (LREC), Gran Canaria, 2002.

Jan Daciuk, Finite Automata for Compact Representation of Language Models in NLP, Six Internations Conference on Implementation and Application of Automata, University of Pretoria, South-Africa, July 23-25 2001.

Jan Daciuk, Computer-assisted Enlargement of Morphological Dictionaries, Finite State Methods in Natural Language Processing, Helsinki, Finland, August 13–24 2001.

Jan Daciuk, Gertjan van Noord A Finite-State Library for NLP, CLIN 2001, University of Twente, Enschede, November 30 2001.

Jan Daciuk. Computer-Aided Enlargment of Morphological Dictionaries. Presented at the Natural Language Processing Seminar, The Linguistic Engineering / Formal Linguistics Group, Linguistic Engineering Group at the Department of Artificial Intelligence, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland, June 25th, 2001 [invited].

Jan Daciuk. Incremental Construction of Minimal, Deterministic, Acyclic, Finite-State Automata. Presented at the Seminar für Sprachwissenschaft, Tübingen University, Germany, May 24th, 2000 [invited].

Jan Daciuk. Construction of Guessing Automata for Morphological Analysis and Morphological Descriptions. Presented at the Seminar für Sprachwissenschaft, Tübingen University, Germany, May 22nd, 2000 [invited].

Tanja Gaustad, The Best of two Worlds: Word Sense Disambiguation Using Statistics and Linguistics. BCN Ph.D. Retreat, Doorwerth, April 2002.

Tanja Gaustad, Word Sense Disambiguation as Classification Problem. Potchefstroomse Universiteit vir Christelike Hoër Onderwys, Potchefstroom, South Africa. March 2002.

Tanja Gaustad, The Importance of High Quality Input for WSD: An Application-Oriented Comparison of Part-of-Speech Taggers, Australasian Language Technology Workshop 2003, University of Melbourne, Australia, December 10 2003,

Tanja Gaustad, Gosse Bouma, Accurate Stemming and Email Classification. CLIN 2001, Enschede, November 30 2001.

Tanja Gaustad, Statistical Corpus-Based Word Sense Disambiguation: Pseudowords vs. Real Ambiguous Words, ACL 2001, Toulouse, France, July 9–11 2001.

Tanja Gaustad, Extraction and Verification of Subcategorization Patterns for French Verbs. EURALEX 2000, Stuttgart, August 2000

Tanja Gaustad, Word Sense Disambiguation Using a Naive Bayes Classification Algorithm and Pseudowords. BCN Poster Day, January 2001, Groningen.

Tanja Gaustad, A Lemma-Based Approach to a Maximum Entropy Word Sense Disambiguation System for Dutch. International Conference on Computational Linguistics (CoLing 2004), Geneva, Switzerland. August 26 2004.

Tanja Gaustad, Linguistic Knowledge and Word Sense Disambiguation. CLCG Linguistic Colloquium, Groningen. October 8 2004.

Robert Malouf, Tony Mullen, Gertjan van Noord, Probabilistic parsing with the Alpino grammar, CLIN 2001, Enschede, November 30 2001.

Robert Malouf and Miles Osborne. A toolkit for robust and efficient maximum entropy language modeling. CLIN 2000, Tilburg, November 2000.

Robert Malouf, Stochastic Head-Driven Phrase Structure Grammar. Department of computational linguistics, University of Saarbrücken, May 2002. [invited]

Robert Malouf. Mixed categories in constraint-based grammars. Emanuel Vasiliu Lectures in Formal Grammars, University of Bucharest, April 2002. [invited]

Robert Malouf, Stochastic Head-Driven Phrase Structure Grammar. Deparment of Linguistics and Oriental Languages, San Diego State University, December 2001. [invited]

Robert Malouf, Stochastic Head-Driven Phrase Structure Grammar, Human Communication Research Centre, University of Edinburgh, June 2001. [invited]

Robert Malouf, Practical and efficient default unification. Microsoft Research, Redmond, Washington. December 2000. [invited]

Mark-Jan Nederhof, Probabilistic parsing as intersection, 8th International Workshop on Parsing Technologies, Nancy, France, April 23 2003

Mark-Jan Nederhof, Partially ordered multiset context-free grammars and free-word-order parsing, 8th International Workshop on Parsing Technologies, Nancy, France, April 23 2003

Nederhof, M J, Probabilistic parsing strategies, University of Edinburgh, May 9 [invited]

Gertjan van Noord Grammar Engineering Using Very Large Corpora, Symposium, University of Edinburgh, March 7 2003 [invited]

Gertjan van Noord, Finite State Language Processing, ACL Tutorial, Sapporo, July 7 2003 [invited]

Gertjan van Noord, Finite State Transducers with Predicates and Identity, CLIN Tilburg, November 3, 2000.

Gertjan van Noord, Alpino: Wide Coverage Computational Analysis of Dutch, Computing with LLL Seminar, University of Amsterdam, June 15 2001 [invited].

Gertjan van Noord, with Dale Gerdemann, Invited Speaker at *SIGPHON 2000, Finite State phonology*, Approximation and Exactness in Finite State Optimality Theory. August 2000, Luxembourg.

Gertjan van Noord, Wide Coverage Computational Analysis of Dutch. University of Sussex. Brighton. February 21 2002 [invited].

Gertjan van Noord, Wide Coverage Computational Analysis of Dutch. Johns Hopkins University. Baltimore. April 9 2002 [invited].

Gertjan van Noord, Alpino: Wide Coverage Parsing with SAVGs. CLIN, Leiden. December 17 2004.

Gertjan van Noord, Error Mining for Wide-Coverage Grammar Engineering. Saarbrcken, Germany. 24th June, 2004. [invited]

Gertjan van Noord, Error Mining for Wide-Coverage Grammar Engineering. ACL 2004 Main Programme, Barcelona, Spain. 23rd July, 2004.

Gertjan van Noord, Wide Coverage Parsing with Stochastic Attribute Value Grammars. IJCNLP Workshop Beyond Shallow Analyses. Hainan, China. 21st March, 2004.

Robbert Prins and Gertjan van Noord. Unsupervised POS-Tagging Improves Parsing Accuracy and Parsing Efficiency. BCN Poster Day. February 2002, Groningen.

Begoña Villada Moirón, Gosse Bouma. A corpus-based approach to the acquisition of collocational prepositional phrases. CLIN 2001, Enschede. November 30 2001.

Begoña Villada Moirón. Extraction of collocational prepositions. BCN Poster Day. February 2002, Groningen.

Begoña Villada Moirón. Computational aspects of (semi-)fixed +expressions: acquisition and modif ication. KNAW visitatie, Groningen. December 15, 2003.

Begoña Villada Moirón. Corpus-based acquisition and classification of idiomatic expressions. Doorwerth, The Netherlands. April 29, 2004.

Begoña Villada Moirón. Computational issues of multi-word expressions: +automatic identification and m odifiability. Trinity College Dublin, Dublin January 21, 2005.

Begoña Villada Moirón. Linguistically enriched corpora for establishing variation in support verb constructions. 6th International Workshop on Linguistically Interpreted Corpora (Linc'05). Jeju Island, R. of Korea. October 14, 2005.

Begoña Villada Moirón. Discarding noise in an automatically acquired lexicon of support verb constructions. LREC, Lisbon Portugal. May 26 2004.

Begoña Villada Moirón. Distinguishing prepositional complements from fixed arguments. EURALEX, Lorient France. July 6, 2004.

Begoña Villada Moirón. Establishing variation and modifiability. CLIN 2004, December 17 2004.

Begoña Villada Moirón. Gaining computational ground on fixed expressions. CLCG Linguistic Colloquium, Groningen. February 25 2005.

Appendix C

Other Research Activities

Leonoor van der Beek:

- member of the board of the EACL Student Board
- co-chair of the ACL Student Research Workshop, Barcelona 2004.

Gosse Bouma:

- program committee *Learning Language in Logic*, Lisbon, 2000.
- program committee *International Conference on HPSG*, Berkely, 2000.
- member *Corpusannotatie* NWO project Corpus Gesproken Nederlands.
- member coordinating committee *Elektronisering van de ANS*, Nederlandse Taalunie.
- member *Platform for Taal- en Spraaktechnologie*, Nederlandse Taalunie.
- editorial board *Computational Linguistics*
- program committee *Formal Grammar* 2002.
- program committee *LREC Workshop Beyond Parseval: towards Improved Evaluation Measures of Parsing Systems*.
- program committee ESSLLI 2003.

Jan Daciuk:

- reviewer for Computational Linguistics and Natural Language Engineering
- reviewer for workshop *Finite State Methods in Natural Language Processing*. Helsinki. 2001

Tanja Gaustad:

- organiser TABU-day, one-day conference on general linguistics, University of Groningen, June 22 2001.
- organiser 13th CLIN Meeting (Computational Linguistics in the Netherlands), 29 november 2002, University of Groningen.

Robert Malouf:

- one week lecture at ESSLLI (with Miles Osborne), An Introduction to Stochastic Attribute-Value Grammars. Helsinki august 2001.
- one week lecture at HPSG Summer School, Statistics for Linguists, Trondheim, Summer 2001.
- KNAW Research Fellow as of January 1, 2002.
- HPSG-L electronic mailing list manager
- Reviewer for Computational Linguistics, Language and Computation, Natural Language Engineering, Natural Language and Linguistic Theory

Mark-Jan Nederhof:

- member of programme committee of Mathematics of Language, 2003.
- member of programme committee of Formal Grammar, 2003
- member of programme committee of Algebraic Methods in Language Processing, 2003

Gertjan van Noord:

- area chair, *COLING*, Saarbrücken. 2000.
- tutorial chair *EACL/ACL* 2001 Toulouse, 2000.
- editorial board *Computer Speech and Language*.
- editorial board of *WEB-SLS*, The European Student Journal of Language and Speech.
- program committee *Workshop Using Toolsets and Architectures to build NLP Systems*. Luxembourg, 2000.
- program committee *Workshop Efficiency in Large-scale Parsing Systems*, Luxembourg. 2000.
- program committee *TAG+ Workshop*, Paris. 2000.
- program committee *TAG+ Workshop*, Venice. 2001.

- program committee *International Conference on HPSG*, Norway. 2001.
- program committee *International Workshop on Naturl Language Understanding and Logic Programming*. Copenhagen, 2002.
- co-chair workshop *Finite State Methods in Natural Language Processing*. Helsinki. 2001
- co-chair *20 years of two-level morphology*. August 2001. Helsinki. 2001.
- one week lecture at the LOT summerschool, Tilburg, June 2000 entitled *Finite State Language Processing*.
- chair-elect EACL 2003-2004.
- chair EACL 2005-2006.
- member programme committee NWO IMIX programme.
- member programme committee IJCNLP 2004.
- member programme committee ACL Student Research Workshop.
- chair International Workshop on Parsing Techonologies (IWPT), Nancy 2003.
- Area chair ACL 2003.

Villada Moirón

- reviewer for International Journal of Lexicography
- reviewer for ESSLI 2005 PhD Student Session

Appendix D

Software and other resources

A number of software packages as well as a number of other resources are maintained by members of the Pionier group. These resources are freely availabe to other members of the research community (the detailed conditions of usage may vary).

Adfa Adfa is a program for testing various acyclic automata construction methods. `http://www.eti.pg.gda.pl/~jandac/adfa.html`

Alpino Treebanks Collection of corpora annotated with CGN dependency structures. Includes both manually corrected (small) treebanks as well as machine-annoated (huge) treebanks `http://www.let.rug.nl/~vannoord/trees/`

Estimate Estimate is a program for parameter estimation of maximum entropy models. `http://www.let.rug.nl/~malouf/maxent/`

Fadd Fadd is a library accessing dictionaries in form of finite-state automata, finite-state perfect hashing functions, and compressed finite-state language models (as produced by the s_fsa program). `http://www.eti.pg.gda.pl/~jandac/fadd.html`

FSA Utilities This is a collection of utilities to manipulate regular expressions, finite-state automata and finite-state transducers. Manipulations include automata construction from regular expresssions, determinization (both for finite-state acceptors and finite-state transducers), minimization, composition, complementation, intersection, Kleene closure, etc. `http://www.let.rug.nl/~vannoord/Fsa`

Minim A set of programs for testing automata minimization algorithms, and in particular Daciuk's version of the incremental algorithm by Bruce Watson. `http://www.eti.pg.gda.pl/~jandac/minim.html`

S_FSA A package of programs for construction and use of finite-state automata for morphological analysis, spelling correction, restoration of diacritics, and perfect hashing. `http://www.eti.pg.gda.pl/~jandac/fsa.html`

Hdrug Hdrug is an environment to develop logic grammars / parsers / generators for natural languages. `http://www.let.rug.nl/~vannoord/Hdrug/`

Stemmer/Lemmatiser A dictionary-based stemmer/lemmatiser for Dutch based on CELEX. `http://www.let.rug.nl/~tanja/code.html`

Error Mining Implementation of the technique described in (van Noord, 2004) to compute Ngram frequencies for very large corpora and very large N. Uses the fadd library mentioned above, and contains a variant of the suffix array construction implementation of Ken Church. `http://www.let.rug.nl/~vannoord/SuffixArrays.tgz`

Appendix E

List of Project Members

Leonoor van der Beek AIO, started April 1 2001. End-date July 1 2005. Project was extended three months for her work in GAIOO.

Gosse Bouma UD/UHD at Alfa-informatica RuG. Some of his teaching is taken over by replacements which are being financed from the Pionier budget.

Jan Daciuk Postdoc. Started February 1 2000. End-date February 1 2003.

Tanja Gaustad AIO, started April 1 2000. Between april 1 2001 and october 1 2001, Tanja worked for an email classification project (*Kennisontwikkeling in Partnerschap* with Bussiness Support Center, Groningen). Therefore, planned end-date October 1 2004.

Robert Malouf Postdoc. From July 1 2001 until January 1 2002. Both before and after this period, Malouf participated in the project.

Tony Mullen Researcher. After his Ph.D. project, Mullen worked for three months for Pionier (January 1 2002 - April 1 2002). During this period he finished his Ph.D.

Robbert Prins AIO, started November 1 2000. Planned end-date November 1 2004.

Gertjan van Noord UHD at Alfa-informatica RuG. Some of his teaching is taken over by replacements which are being financed from the Pionier budget.

Begoña Villada Moirón AIO, started November 1 2000. Planned end-date November 1 2004.

Mark-Jan Nederhof UD, 2003-2005.

John Nerbonne Promotor of Van der Beek, Gaustad, Mullen, Prins, and Villada Moirón.

Appendix F

Financial overview

The following table gives the financial situation of the project. All amounts in EURO.

	2000	2001	2002	2003	2004	2005	2006	total
Personnel:								
Daciuk	50k	57k	61k	5k				175k
Gaustad	19k	14k	31k	38k	31k			134k
Prins	5k	27k	29k	34k	35k			132k
Villada	5k	27k	30k	35k	33k	13k		145k
Malouf		20k						20k
Mullen			11k					11k
vd Beek		19k	26k	28k	32k	9k		116k
Nederhof				21k	51k	16k		98k
Teaching	11k	50k	52k	50k	35k	20k	15k	219k
Further Costs	88k	10k	6k	4k	5k	20k		124k
Reserved								20k
Total								1204k

- *Teaching* refers to a number of teachers we have employed to take over most of the teaching obligations of Gosse Bouma and Gertjan van Noord.
- *Further Costs* include travel money (for conference visits, etc.), non-standard hardware (in the first year of the project we invested in a cluster of 7 Alpha 64-bit Unix machines), and the costs for dissemination, including the final workshop.