Progress Report STEVIN Projects

Project Name	Large Scale Syntactic Annotation of Written Dutch
Project Number	STE05020
Reporting Period	April 2009 - September 2009
Participants	KU Leuven, University of Groningen
Start date	November 2006
End date (original)	November 2009
End date (extended)	September 2010

1 Summary of the project

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.

1.1 Deliverables

Deliverable 1.1 Planned after 3 months.

Specification of the 1 million word corpus (Lassy Small) that will be annotated syntactically.

Deliverable 1.2 Planned after 18 months.

Specification of the 500 million word corpus that will be automatically parsed in Lassy.

Please note that the numbers in Deliverable 2.1 – 3.4 refer to the total number of words, which include the portions that were already annotated in D-Coi (200.000 words syntactically annotated, and 500.000 words annotated with POS-tag and lemma). Therefore, although the resulting Lassy Small corpus contains 1.000.000 words, in Lassy we must annotate 800.000 words syntactically, and 500.000 words with respect to POS-tag and lemma.

Deliverable 2.1 Planned after 6 months.

250.000 words annotated and verified for POS-tag and lemma. In total, 750.000 words (75% of Lassy Small) is now annotated for POS and lemma.

Deliverable 2.2 Planned after 12 months.

250.000 words annotated and verified for POS-tag and lemma. In total, 1.000.000 words (100% of Lassy Small) is now annotated for POS and lemma.

Deliverable 3.1 Planned after 12 months.

400.000 words syntactically annotated. In total, 600.000 words (60% of Lassy Small) is now syntactically annotated.

Deliverable 3.2 Planned after 18 months.

600.000 words syntactically annotated. In total, 800.000 words (80% of Lassy Small) is now syntactically annotated.

Deliverable 3.3 Planned after 24 months.

1.000.000 words syntactically annotated. In total, 1.000.000 words (100% of Lassy Small) is now syntactically annotated.

Deliverable 3.4 Planned after 24 months.

Report on annotation (including manual verification) of Lassy Small.

Deliverable 3.5 New deliverable: revised and extended syntactic annotation manual. Planned after 24 months.

Deliverable 4.1 Planned after 18 months.

Improved version of Alpino, based on initial experiments with Lassy Large.

Deliverable 4.2 Planned after 24 months.

Report on formal quantitative evaluation of annotation on Lassy Small, in order to estimate quality of Lassy Large.

Deliverable 4.3 Planned after 24 months.

POS-tags and Lemma annotation for Lassy Large. Not manually verified.

Deliverable 4.4 Planned after 24 months.

Syntactic annotation for Lassy Large. Not manually verified.

Deliverable 5.1 Planned after 12 months.

Feasibility study on information extraction from resources such as Lassy Large, i.e., large collections of XML-encoded dependency structures.

Deliverable 5.2 Planned after 18 months.

Specification of XML tools for information extraction from large XML-encoded syntactic corpora.

Deliverable 5.3 Planned after 24 months.

First release of XML tools for information extraction from large XML-encoded syntactic corpora.

Deliverable 5.4 Planned after 36 months.

Final release of XML tools for information extraction from large XML-encoded syntactic corpora.

Deliverable 6.1 Planned after 18 months.

Report on case study 1.

Deliverable 6.2 Planned after 24 months.

Report on case study 2.

Deliverable 6.3 Planned after 30 months.

Report on case study 3.

Deliverable 7 Planned after 36 months.

Final report

1.2 Previously completed deliverables

In the reporting period, the following deliverables have been completed: 1.2, 2.2, 3.3, 3.5, 5.2, 5.3, 6.1 and 6.2.

1.3 Changes requested

At the moment of writing, we have integrated the results of deliverable 2.2 (POS and lemma annotation) and 3.3 (syntactic annotation) into a single XML-format. This work has been completed only very recently.

As a result, the deliverable 3.4 (report of 3.3) and 4.2 (evaluation of Alpino parser on Lassy Small corpus) are delayed a few months. These are expected for December 2009.

Parsing the Lassy Large corpus is underway. In fact, most of the material has been parsed already (and has been used in the various case studies). For the final deliverable we plan to parse most of the material again, in order that the resulting treebank is more consistent with the annotation manual and with the manual annotations. Moreover, the latest version of Alpino includes many small improvements which were implemented by error mining the earlier parsing results. We expect to finish this work (deliverable 4.4) in May 2010.

1.4 Employee involvement in relation to the original plan

The involvement of employees is in accordance to the original plan.

1.5 Dissemination of the results

There is a web-page dedicated to Lassy with links to all available resources: `http://www.let.rug.nl/~vannoord/Lassy/`

In June 2009, Lassy sponsored one of the invited speakers for the 30th TaBu meeting, organized by the Center for Language and Cognition, Groningen. We were very proud that with the financial contribution of Lassy, Ken Church (formerly Microsoft Research; now Johns Hopkins University) was the keynote speaker of this event. The contribution of Lassy consisted of the travel costs and hotel costs of Ken Church.

1.5.1 Publications

- Danil de Kok, Jianqiang Ma and Gertjan van Noord, A generalized method for iterative error mining in parsing results. In: ACL2009 Workshop Grammar Engineering Across Frameworks (GEAF), Singapore, 2009.

- Kostadin Cholakov and Gertjan van Noord. Combining Finite State and Corpus-based Techniques for Unknown Word Prediction. In: Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria, 2009.

- Anna Lobanova, Jennifer Spenader, Tim van de Cruys, Tom van der Kleij and Erik Tjong Kim Sang. Automatic Relation Extraction - Can Synonym Extraction Benefit from Antonym Knowledge?, In: Proceedings of WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies (NODALIDA2009 workshop), Odense, Denmark, May 2009.

- Erik Tjong Kim Sang and Katja Hofmann, Lexical Patterns or Dependency Patterns: Which Is Better for Hypernym Extraction? In: Proceedings of CoNLL-2009, Boulder, CO, USA, June 2009.

1.5.2 Presentations

- Gertjan van Noord. Overzicht Lassy Project. STEVIN programmadag, 2 september 2009, Tilburg.

- Danil de Kok, Jianqiang Ma and Gertjan van Noord, A generalized method for iterative error mining in parsing results. In: ACL2009 Workshop Grammar Engineering Across Frameworks (GEAF), Singapore, 2009.

- Kostadin Cholakov and Gertjan van Noord. Combining Finite State and Corpus-based Techniques for Unknown Word Prediction. In: Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria, 2009.

- Anna Lobanova, Jennifer Spenader, Tim van de Cruys, Tom van der Kleij and Erik Tjong Kim Sang. Automatic Relation Extraction - Can Synonym Extraction Benefit from Antonym Knowledge?, In: Proceedings of WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies (NODALIDA2009 workshop), Odense, Denmark, May 2009.

- Erik Tjong Kim Sang and Katja Hofmann, Lexical Patterns or Dependency Patterns: Which Is Better for Hypernym Extraction? In: Proceedings of CoNLL-2009, Boulder, CO, USA, June 2009.

layer	annotated (april 2009)	annotated (october 2009)	target	to do
lemmatization	720	820	820	0
POS-tagging	720	820	820	0
Syntactic	840	900	900	0

Table 1: Progress of Annotation Efforts. All numbers are Kilo-words.

2 Progress per deliverable

In the reporting period, most of the work has gone into:

- Manual correction of Lassy Small
- Integration of POStag and lemma annotation on the one hand, with the syntactic annotation on the other hand
- Finalizing two of the case studies

We now describe progress per deliverable as follows.

2.1 Deliverables 1: Corpus Selection

Deliverable 1.2 (contents of Lassy Large) has now be completed. We have included SONAR release 1. If further SONAR data is available in time, it will also be added to Lassy Large.

2.2 Deliverables 2 and 3: Manual Annotation Efforts

We summarize the progress with respect to the manual annotation efforts here for both lemmatization, POS-tagging and syntactic annotation.

As per the end of the reporting period, the manual annotation of LEMMA, POSTAG and SYNTAX has been completed.

This means that we now have the Lassy Small corpus of about 1,000,000 words annotated with the LEMMA, POSTAG and SYNTAX layers. Each of the sentences and layers has at least been once manually verified.

In addition, we have developed tools to check for consistency and for frequent mistakes automatically. This has resulted in thousands of errors which were detected and subsequentially corrected. The speed with which new erros were identified suggested that the quality of the corpus can still be improved quite drastically. For this reason we decided to continue manually verifying the corpus. This work will take place in Groningen over the next couple of months. However, we have make a first beta-release version of the corpus available per November, 2, 2009.

This version of the corpus will contain a single XML file format which includes both the syntactic dependency annotation as well as the POSTAG and LEMMA annotation.

Once again, effort has been spent on the improvement and extension of the syntactic annotation manual. The main purpose of the manual is the documentation of the treebanks (both manually verified, and automatically constructed). The manual is now considered finished.

2.3 Deliverables 4: Automatically parsed treebanks

Recent version of Alpino which include improvements based on error mining on large corpora, is available on the Alpino website. Alpino is now available for the Linux (32 and 64 bit), Windows and OS/X (32 and 64 bit) platforms.

We employ the High Performance Computing cluster of the University of Groningen, as well as the European GRID to parse large sets of sentences with Alpino.

2.4 Deliverables 5: XML Technology

The LASSY postdoc in Groningen, Erik Tjong Kim Sang, has performed a evaluation of tools used for searching in syntactically annotated corpora (deliverable 5.1). He found that XQuery is the most suitable method for specifying complex syntactic queries. However, XQuery is a programming language which makes it unsuitable for users without programming skills. Currently, we are still working on the of a tool which translates the easily specified common corpus queries to XQuery automatically. In addition, a library of common XQuery functions for accessing the Lassy treebanks is being (further) developed.

The most recent implementations of the XML tools is part of the Alpino distribution.

2.5 Deliverables 6: Case Studies

The first two case studies have now been finished.

The third case study, on bilexical preferences, has been carried out, but the report describing the study in detail is still missing.

In a paper presented at IWPT 2007, van Noord describes a method to incorporate bilexical preferences between phrase heads, such as selection restrictions, in a Maximum-Entropy parser for Dutch. The bilexical preferences are modeled as association rates which are determined on the basis of a very large parsed corpus (about 500M words). We show that the incorporation of such self-trained preferences improves parsing accuracy significantly.

More recently, we have attempted to use the same method for different corpora and for parsing in other domains.