Progress Report STEVIN Projects

Project Name	Large Scale Syntactic Annotation of Written Dutch
Project Number	STE05020
Reporting Period	October 2008 - March 2009
Participants	KU Leuven, University of Groningen
Start date	November 2006
End date (original)	November 2009
End date (requested)	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

Deliverable 1.1 (Specification of Lassy Small) has been finished. It is available on the Lassy website (`http://www.let.rug.nl/~vannoord/Lassy`) and the Stevin WIKI site.

Deliverable 5.1 (Feasibility study Search / Extraction Tools) has been completed and is available on the Lassy website (`http://www.let.rug.nl/~vannoord/Lassy`) and the Stevin WIKI site.

1.3 Changes requested

Due to a number of startup problems, explained in the previous progress report, we ask permission to move the date associated with the deliverable 5.2 (specification XML tools) forward in time to April 2009. Motivation for the delay: we were not able to fill the proposed Postdoc position in Groningen in time. The actual work for this deliverable has been finished, but the report will be available only in a few weeks.

Based on the feedback in the validation report of D-Coi, we have once again critically reviewed the existing annotation guidelines. We propose to add a new deliverable for the project: a revised and extended version of the syntactic annotation manual.

1.4 Employee involvement in relation to the original plan

The involvement of employees is in accordance to the original plan, with one exception. The three year post-doc position in Groningen could only be filled recently. For this reason, contributions by other members of the research group in Groningen (in particular Gosse Bouma, Geert Kloosterman and Gertjan van Noord) have been intensified. As of February 1st, 2008, Erik Tjong Kim Sang has been working as a post-doc for Lassy. As a consequence, we are somewhat behind with respect to work-package 5.

For the annotation work, which relies mostly on student assistants, both in Groningen and in Leuven we could not hire a student assistant, and therefore this part of the work has not made much progress. In Groningen, a student assistant has started as of March 1, 2009. In Leuven, student assistants can only be employed within terms: this will happen in the summer of 2009.

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 January 2009, the TLT conference (Treebanks and Linguistic Theory) has been organized by the Lassy consortium. The conference took place in Groningen in conjunction with the 19th Meeting of Computational Linguistics in the Netherlands. More information can be found at the website http://www.let.rug.nl/tlt. The proceedings of the workshop are available from http://lotos.library.uu.nl/.

Invited keynote speakers at TLT were Robert Malouf (San Diego), and Adam Przepiórkowski (Warsaw).

At the TLT conference, we ensured Lassy visibility by means of a number of accepted presentations by researchers from the Lassy team. The following presentations constituted session B of the programme, a session specifically dedicated to Lassy:

- Gertjan van Noord. Huge Parsed Corpora in LASSY
- Ineke Schuurman, Veronique Hoste and Paola Monachesi. Cultivating Trees: Adding Several Semantic Layers to the Lassy Treebank in SoNaR
- Gosse Bouma and Jennifer Spenader. The Distribution of Weak and Strong Object Reflexives in Dutch

This session was followed by the poster session which included:

- Erik Tjong Kim Sang. To Use a Treebank or Not - Which Is Better for NLP Tasks? (Poster)

In addition, there was a Lassy poster with accompanying demonstrations by the full Lassy team.

In June 2009, Lassy is sponsoring one of the invited speakers for the 30th TaBu meeting, organized by the Center for Language and Cognition, Groningen. We are very proud to announce that with the financial contribution of Lassy, Ken Church (Microsoft Research) will be the keynote speaker of this event. Ken Church has accepted the invitation. The contribution of Lassy consists of the travel costs and hotel costs of Ken Church.

1.5.1 Publications

- Gertjan van Noord. Huge Parsed Corpora in LASSY. In: Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7), Groningen, The Netherlands, 2009. pp 115-126.
- Ineke Schuurman, Veronique Hoste and Paola Monachesi. Cultivating Trees: Adding Several Semantic Layers to the Lassy Treebank in SoNaR. In: Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7), Groningen, The Netherlands, 2009. pp. 135-146.

- Gosse Bouma and Jennifer Spenader. The Distribution of Weak and Strong Object Reflexives in Dutch. In: Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7), Groningen, The Netherlands, 2009. pp. 103-114.

- Erik Tjong Kim Sang. To Use a Treebank or Not - Which Is Better for Hypernym Extraction? In: Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7), Groningen, The Netherlands, 2009. 171-176.

- Frank van Eynde, Anette Frank, Koenraad de Smedt, Gertjan van Noord (editors). Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7), Groningen, The Netherlands, 2009. 197 pages.

1.5.2 Presentations

- Gertjan van Noord. Huge Parsed Corpora in LASSY. Presentation at TLT 7, The Seventh International Workshop on Treebanks and Linguistic Theories. January 23, 2009, Groningen.

- Ineke Schuurman, Veronique Hoste and Paola Monachesi. Cultivating Trees: Adding Several Semantic Layers to the Lassy Treebank in SoNaR. Presentation at TLT 7, The Seventh International Workshop on Treebanks and Linguistic Theories. January 23, 2009, Groningen.

- Gosse Bouma and Jennifer Spenader. The Distribution of Weak and Strong Object Reflexives in Dutch. Presentation at TLT 7, The Seventh International Workshop on Treebanks and Linguistic Theories. January 23, 2009, Groningen.

1.5.3 Posters

- Erik Tjong Kim Sang. To Use a Treebank or Not - Which Is Better for NLP Tasks? Poster at TLT 7, The Seventh International Workshop on Treebanks and Linguistic Theories. January 23, 2009, Groningen.

- Gertjan van Noord, Gosse Bouma, Ineke Schuurman, Vincent Vandeghinste, Frank van Eynde, Erik Tjong Kim Sang. Lassy Demos. Poster and demos at TLT 7, The Seventh International Workshop on Treebanks and Linguistic Theories. January 23, 2009, Groningen.

1.6 Exploitation of the results

For a number of initiatives refer to the section *Deliverables 6* below.

material	size	IPC
D-Coi minus Wikipedia, Europarl	22M	settled in D-Coi
XMLWiki Wikipedia 2008	110M	GFDL
Europarl version 3	38M	public
TwNC newspaper material	531M	ok for research; unclear otherwise
Mediargus newspaper material	1397M	ok for research; unclear otherwise
European Medicines Agency	14M	public

Table 1: Corpus selection Lassy Large. Information about the copyright status of the material from the European Medicines Agency can be obtained from `http://www.emea.europa.eu/htms/technical/dmp/copyrite.htm#copyright`

2 Progress per deliverable

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

- Organisation of the TLT conference, including presentations, posters and demonstrations of Lassy participants
- Revision and extension of syntactic annotation manual
- XML Technology

We now describe progress per deliverable as follows.

2.1 Deliverables 1: Corpus Selection

Deliverable 1.2 (contents of Lassy Large) has not been completed yet, due to the fact that D-Coi's follow-up project SoNaR did not start as early as we had hoped.

The current corpus selection can be summarized as in table 1.

The selection in this overview should be regarded as potential components of a fall-back option if material from SoNaR is not available in time. For TwNC and Mediargus data, this would involve further negotiations with content providers.

The EMEA texts are selected, cleaned, and tokenized in collaboration with the Paco-MT STEVIN project.

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, April 1, 2009, manual annotation has progressed in table 2. As can be seen from this table, annotation for Lassy Small progresses according to schedule. Please note that the numbers in this table refer to the number of words to be annotated

layer	annotated (oct 2008)	annotated (april 2009)	target	to do
lemmatization	720	720	820	100
POS-tagging	720	720	820	100
Syntactic	810	840	900	60

Table 2: Progress of Annotation Efforts. All numbers are Kilo-words.

in Lassy. Together with the amount of data annotated in D-Coi, the resulting Lassy Small corpus contains 1.000.000 words.

In the table, it can be observed that the amount of data that we set out to annotate has now been annotated. However, because we included some new datasets in Lassy Small (in particular the inclusion of material from DPC and Wikipedia material from WikiXML), this implied we had to invest in additional annotation work for lemmatization and POS-tagging.

A lot of effort has been spent on the improvement and extension of the syntactic annotation manual. Based on the feedback in the validation report of D-Coi, we have once again critically reviewed the existing annotation guidelines. We have reduced the amount of linguistic argumentation, and we have extended the manual rather drastically by adding many more examples, and by filling in various omissions in the manual. The manual also considers some of the issues in automatically annotated corpora. The main purpose of the manual is the documentation of the treebanks (both manually verified, and automatically constructed). A draft of this manual, which now contains over 200 pages, is available.

We propose to add the revised syntactic annotation manual as an additional deliverable (3.5) of the Lassy project.

In addition to the completion of the manual annotation task, a further task concerns the integration of the POStag and lemma annotations layers on the one hand with the syntactic annotation layer on the other hand. We have developed scripts to integrate the two formats, but the actual integration has not yet been completed.

Finally we foresee, once the various annotation layers are integrated, a phase of consistency checking and error corrections.

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. There now is an experimental version of Alpino available for the Windows platform.

In a paper accepted for the main session of the EACL 2009 in Athens, we describe a corpus-based technique to improve the efficiency of wide-coverage high-accuracy parsers. By keeping track of the derivation steps which lead to the best parse for a very large collection of sentences, the parser learns which steps in the parser can be filtered without significant loss in parsing accuracy. It turns out that the increased efficiency of the parser is helpful for our efforts to parse large amounts of corpus material, reducing CPU-requirements to 25%.

Preparations are underway to be able to employ grid computing for parsing large amounts of data 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 working on the initial version of a tool which translates the easily specified common corpus queries to XQuery automatically.

2.5 Deliverables 6: Case Studies

This set of deliverables is due at a later phase. We list a number of initiatives that members of the Lassy consortium were involved in, where syntactically annotated corpora comparable to Lassy Large were used for tasks of the type foreseen here. These initiatives constitute potential candidate applications to be worked out in full detail as one of the three case studies foreseen here.

2.5.1 Information Extraction

In a cooperation with Katja Hofmann (University of Amsterdam), we have been investigating two preprocessing methods for automatically extracting semantic information from text: shallow parsing and dependency parsing. We are particularly interested in whether the richer annotation produced by dependency parsing allows for a better performance of subsequent information extraction work. We evaluate extraction approaches for hypernym information and conclude that application of dependency patterns outperforms application of shallow parsing patterns, albeit at a considerable extra processing cost. This suggests that the construction of Lassy Large can indeed be a useful resource for applications in information extraction. Furthermore, the availability of a large parsed corpus can be advantageous to alleviate the observed efficiency bottle-neck for on-line application of a dependency parser.

2.5.2 Corpus Linguistics

In a cooperation with Bastiaanse (University of Groningen), we have performed a corpus linguistics study on the basis of a very large corpus of automatically syntactically annotated sentences (this resource can be regarded an initial version of Lassy Large). The corpus study resulted in corpus frequency data for constructions that have previously been used to show the influence of linguistic complexity on Dutch agrammatic speech production.

There is a long standing debate between aphasiologists with a linguistic and a psychological background on the essential factor that constitutes the behavioral patterns of loss and preservation in agrammatic Broca's aphasia. Generally speaking, linguists attempt to describe these patterns in terms of linguistic complexity, whereas psychologists prefer an explanation in terms of processing. In the latter, frequency plays a large role. The idea is that the more frequent a phenomenon is, the easier it is to process for aphasic patients. Frequency may play a role at several levels. For agrammatic patients, for example, the frequency of sentence constructions may be crucial, whereas for fluent aphasic speakers word frequency influences performance.

We compared the data of our corpus research with the performance of agrammatic speakers on the construction. These are data on: (1) verb movement; (2) object scrambling; and (3) verbs with alternating transitivity.

The conclusion is that frequency cannot account for the data.

2.5.3 Bilexical Preferences

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.

2.5.4 Question Answering

A prototype question answering system, based on Alpino and called *Joost* has been implemented in the context of the NWO IMIX programme. The system is extended with various techniques to create, enhance and exploit semantic ontologies and pronoun resolution. Joost takes part in the European CLEF evaluation platform since 2005, and obtained the best results for Dutch each year it participated. This initiative is linked with Lassy, because Joost assumes access to syntactic analyses of all of the sentences of its corpus. Last year, the corpus of CLEF was extended beyond the four years of newspaper texts from previous years, to include the full Dutch Wikipedia (58 million words). The full text collection was parsed and the resulting Lassy dependency structures were stored in XML. In 2008, Joost was the only participant of the Dutch monolingual QA task at CLEF.