Progress Report STEVIN Projects

Project Name	Large Scale Syntactic Annotation of Written Dutch
Project Number	STE05020
Reporting Period	April 2008 - September 2008
Participants	KU Leuven, University of Groningen

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.

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 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 will be made available on the Stevin WIKI site soon.

1.3 Changes requested

Due to a number of startup problems, explained in the previous progress report, we ask permission to move the dates associated with the deliverables 5.1 and 5.2 (feasability and specification XML tools) forward in time. In particular, we propose to move this date to January 2009. Motivation for the delay: we were not able to fill the proposed Postdoc position in Groningen in time. We have chosen to focus initially on completion of the actual annotation work. The actual annotation work has been mostly completed. During the coming months there is time to focus on the work associated with work package 5.

1.4 Employee involvement in relation to the original plan

The involvement of empoyees 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 workpackage 5.

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) will be organized by the Lassy consortium. The conference takes place in Groningen in conjunction with the 19th Meeting of Compuational Linguistics in the Netherlands. More information can be found at the website `http://www.let.rug.nl/tlt`.

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

At the TLT conference, we ensure Lassy visibility because of a number of accepted presentations by researchers from the Lassy team:

- Ineke Schuurman, Veronique Hoste and Paola Monachesi. Cultivating Trees: Adding Several Semantic Layers to the Lassy Treebank in SoNaR

- Gertjan van Noord. Huge Parsed Corpora in LASSY

- Gosse Bouma and Jennifer Spenader. The Distribution of Weak and Strong Object Reflexives in Dutch
- Erik Tjong Kim Sang. To Use a Treebank or Not - Which Is Better for NLP Tasks? (Poster)

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.

1.5.1 Publications

- Nelleke Oostdijk, Martin Reynaert, Paola Monachesi, Gertjan van Noord, Roland Ordelman, Ineke Schuurman, Vincent Vandeghinste. From D-Coi to SoNaR: A reference corpus for Dutch. In: Proceedings of LREC 2008. Marrakecht, Morocco.
- Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas, Jörg Tiedemann. Question Answering with Joost at CLEF 2008. CLEF 2008 Working Notes. Aarhus Denmark.
- Barbara Plank and Gertjan van Noord. Exploring An Auxiliary Distribution based approach to Domain Adaptation of a Syntactic Disambiguation Model. In: Coling Workhop 'Cross Framework and Cross Domain Parser Evaluation'.
- Pamela Forner, Anselmo Peñas, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe and Erik Tjong Kim Sang, "Overview of the CLEF 2008 Multilingual Question Answering Track. In "Working Notes for the CLEF 2008 Workshop", Aarhus, Denmark, 2008.

1.5.2 Presentations

- Gertjan van Noord. Self-trained Bilexical Preferences for Improved Syntactic Disambiguation. Invited lecture at CoLi department, University of Saarland, Saarbruecken. 10 July 2008.
- Gertjan van Noord. Large Scale Syntactic Annotation for Dutch. Amsterdam. 25 April 2008. Invited lecture at Q-go, Diemen.
- Erik Tjong Kim Sang and Katja Hofmann, "Shallow Parsing and Full Parsing: Which is the Better Preprocessor?". Presented at ATILA 2008, Antwerp, 15-16 October 2008.

1.5.3 Posters

- Lassy was presented by poster at the STEVIN mid-term review meeting, Rotterdam. 27 June 2008.
- Lassy was presented by poster at the Stevin Programmadag, Hoeven. 11 September 2008.

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`

1.6 Exploitation of the results

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

2 Progress per deliverable

2.1 Deliverables 1: Corpus Selection

Deliverable 1.1 (contents of Lassy Small) has been completed.

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 2.

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.

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, October 1, 2008, manual annotation has progressed in table 3. As can be seen from this table, annotation for Lassy Small progresses according to schedule.

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 further point worth mentioning is that our software for editing XML-encoded dependency structures has been improved quite substantially. It was decided that the original tool that was based on Thistle is not supported any longer. We now exclusively use the TrEd editor (developed by Petr Pajas in Prague, and available from `http://ufal.mff.cuni.cz/~pajas/`

layer	annotated	original target	current target	todo
lemmatization	720	500	820	100
POS-tagging	720	500	820	100
Syntactic	810	800	900	90

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

tred/). This tool allows for the inclusion of platform specific extensions (defined in Perl). We have extended the Alpino specific parts of the editor considerably, using wishes from the annotators as input. Both TrEd as well as the Alpino specific extension modules constitutes free software, available under the GPL - The General Public Licence. The Alpino specific extension modules are distributed with Alpino, under the LGPL license. Alpino is available from http://www.let.rug.nl/~vannoord/alp/Alpino/.

Based on the feedback in the validation report of D-Coi, we have started to extend and improve the syntactic annotation manual. We plan to have a new version ready with the completion of the manual annotation phase. This version of the manual will of course be less useful during actual annotation work (which is finished by then), but is still important to document the annotated material.

2.3 Deliverables 4: Automatically parsed treebanks

2.3.1 Increased amounts of automatically parsed data

In recent months, we have greatly extended our collection of automatically annotated syntactic material. This can be taken as the preliminary activities for the deliverables 4. In the past few months, new annotations were constructed for the 2008 dump of the Dutch Wikipedia, (110 million words), newspapers from 2005 of the TwNC (Dutch newspapers) and material from the European Medicines Agency (14 million words).

Based on careful inspection of the parse results as well as the various log files, we have been able to spot many detailed inconsistencies and errors in various components of Alpino, as well as in initial steps (corpus cleanup, tokenization). This has led to a long list of detailed changes to Alpino, most of which have been implemented. A new version of Alpino including further improvements is already available on-line, and can be seen as the current version of deliverable 4.1. Depending on decisions concerning the final composition of Lassy Large, we will make additional improvements to Alpino available.

2.3.2 Increased Parsing Efficiency

In a recently submitted paper, 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.

2.4 Deliverables 5: XML Technology

Due to the delay in finding a suitable post-doc candidate in Groningen, work for this deliverable is still behind schedule.

We have investigated the use of XPATH and XQUERY for exploiting large annotated corpora. Initial results were reported in a paper by Gosse Bouma and Geert Kloosterman, presented at the ACL workshop on Linguistic Annotation, entitled *Mining Syntactically Annotated Corpora using XQuery*.

As described in that paper, users have taken quite different approaches to corpus exploration and data extraction.

- For corpus exploration, Alpino `dtsearch` is the most widely used tool. It allows XPath queries to be matched against trees in a treebank. The result can be a visual display of trees with matching nodes highlighted, but alternative outputs are possible as well. Examples of how XPath can be used for extraction are presented in the next section.
- For relation extraction (for instance, finding symptoms of diseases, or finding capitals of countries), the Alpino system itself has been used. It provides functionality for converting dependency trees in XML into a Prolog list of dependency triples. The full functionality of Prolog can then be used to do the actual extraction.
- Alternatively, one can use XSLT to extract data from the XML directly. As XSLT is primarily intended for transformations, this tends to give rise to very complex code. More complicated extraction patterns are almost impossible to implement in this way.
- Alternatively, a general purpose scripting or programming language such as Perl or Python, with suitable XML support, can be used. As in the Alpino/Prolog case, this has the advantage that one has a full programming language available. A disadvantage is that there is no specific support for working with dependency trees or triples.

None of the approaches listed above is optimal. XPath is suitable only for identifying syntactic patterns, and does not offer the possibility of extraction of elements (i.e. it has no capturing mechanism). The other three approaches do allow for both matching and extraction, but they all require skills that go considerably beyond conceptual knowledge of the treebank and some basic knowledge of XML.

Another disadvantage of the current situation is that there is little or no sharing of solutions between users. Yet, different applications tend to encounter the same problems. For instance, multiword expressions (such as Alan Turing or 7 juni 1954) are encoded as trees, dominated by a cat='mwu' node. An extraction task that requires names to be extracted must thus take into account the fact that names can be both nodes with a label pos='name' as well as cat='mwu' nodes (dominating a pos='name'). There are a large number of similar issues that complicate the task of formulating extraction patterns.

Bouma and Kloosterman conclude that XPATH (and the Alpino/D-Coi/Lassy tool which uses it, `dtsearch`) essentially is appropriate for search, whereas for extraction application, they illustrate that XQuery could be a suitable candidate. Moreover, they provide an XQuery library consisting of a collection of high-level constructs specifically for the CGN/Alpino/D-Coi/Lassy dependency structures. The availability of such a library facilitates the specification of extraction patterns from Lassy corpora considerably.

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 consitute potentical 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 modelled 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 strutures were stored in XML. In 2008, Joost was the only participant of the Dutch monolingual QA task at CLEF.