Progress Report STEVIN Projects

Project Name	Large Scale Syntactic Annotation of Written Dutch
Project Number	STE05020
Reporting Period	October 2009 - March 2010
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, no deliverables have been completed.

1.3 Changes requested

In the reporting period, much of the project resources have been aimed at improving the quality of the manually annotated data. The integration of the POStag and lemma information with the syntactic annotations gave new possibilities for semi-automatic error disovery. Thousands of errors have been corrected.

Since the manual annotation efforts have not been finalized, the related report (deliverable 3.4) has not yet been delivered. It is expected for June 2010.

Similarly, since the Lassy Small corpus has been undergoing changes, we delayed the deliverable 4.2 (evaluation of Alpino parser on Lassy Small corpus). The tools and scripts are in place, and at any moment we can generate the required numbers for this report. The deliverable is delayed until June 2010.

Parsing the Lassy Large corpus is underway. As indicated in the previous deliverable, we have re-parsed most of the material, in order that the resulting treebank is more consistent with the annotation manual and with the manual annotations. We still 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/`

1.5.1 Presentations

- Erik Tjong Kim Sang, LASSY for Beginners. Talk presented at CLIN 2010, Utrecht, 05-02-2010. http://ifarm.nl/erikt/talks/clin2010.pdf

2 Progress per deliverable

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

- Manual correction of Lassy Small, both with respect to POStag and lemma annotation, and with respect to the syntactic dependency annotations.
- With respect to the XML tools for treebank exploitation, we have developed a web interface to Lassy Small, which provides a quick and easy access to the LASSY corpus for a general audience.
- Improvements to the Alpino parser
  - Many small improvements based on error mining of Lassy Large corpora
  - port to SWI-Prolog (open source prolog)
  - UTF-8 encoded input and output
  - Alpino can now be incorporated as a Prolog library in other software programs

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

The manual annotation of LEMMA, POSTAG and SYNTAX has been completed, but as indicated in the previous report, the quality of some parts of the corpus appeared to be disappointing, and we have invested additional effort in manual verification of the corpus. This entailed both going over some parts of the corpus systematically, as well as using tools to discover inconsistencies and unexpected patterns of annotation. Thousands of errors have been corrected in the reporting period.

A further change in Lassy Small is the conversion to UTF-8 encoded XML files, rather than latin1 encoding that we used previously.

However, we have made an update of the beta-release of the Lassy Small corpus available. This beta-release is now also uploaded to the TST-Centrale, and is available for interested parties via the TST-Centrale.

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.

More recently parsed parts of the Lassy Large corpus will be available as (compressed sequences of) UTF-8 encoded XML files.

In the context of this deliverable, it is worth mentioning that Alpino no longer requires SICStus Prolog. In a cooperation with Jan Wielemakers, a fully functional version of Alpino can be built from sources using the open source SWI-Prolog.

Since the Lassy Small corpus is undergoing further manual correction, we have not yet finalized the report of the evaluation of Alpino on this corpus. However, we can give the accuracy obtained by Alpino on the current version of Lassy Small in table 1. These results indicate which accuracy can be expected for the automatically annotated Lassy Large corpus.

2.4 Deliverables 5: XML Technology

A web interface to the Lassy Small corpus has been developed which provides a quick and easy access to the LASSY corpus for a general audience. Users can enter queries consisting of individual words or phrases. The interface will provide them with sentences from the corpus that contain these words or phrases. For example, the query “leek” (seemed) will provide results like:

1. Biobrandstof **leek** zo’n goed alternatief om onze afhankelijkheid van fossiele brandstoffen te verminderen .
    - leek:WW:hd::
2. Het **leek** zo’n goed middel in onze strijd tegen de uitstoot van broeikasgassen .
    - leek:WW:hd::

An advantage over web engine search is that our data also includes syntactic annotation. A summary of the annotation details of the query parts are made visible in the search results (below each sentence). In this example, both query words are verbs (WW) and are the grammatical head (hd) of their sentence. The syntactic annotation also makes possible queries for words with a specific part-of-speech class, like in ”word=’leek’ and postag=’n’”:

1. Er werd een **leek** aangesteld van de school van Bain en Spencer .
    - leek:N:obj1:WW:aangesteld

sub-corpus	CA%	msec/s	#s	#w	mean l
dpc-bal-	92.55	1787	620	8825	14.2
dpc-bmm-	87.27	4683	794	15589	19.6
dpc-cam-	91.41	3184	508	9961	19.6
dpc-dns-	90.25	1209	264	3833	14.5
dpc-eli-	89.62	4756	603	11309	18.8
dpc-eup-	88.29	10148	233	6085	26.1
dpc-fsz-	84.38	4793	574	10967	19.1
dpc-gaz-	88.03	3953	210	3806	18.1
dpc-ibm-	89.93	5426	419	8473	20.2
dpc-ind-	90.99	4416	1650	33928	20.6
dpc-kam-	89.16	5326	52	1329	25.6
dpc-kok-	88.28	2926	101	1846	18.3
dpc-med-	90.27	4200	650	13575	20.9
dpc-qty-	89.74	8689	618	13720	22.2
dpc-riz-	86.21	5381	210	4217	20.1
dpc-rou-	91.52	2495	1356	22640	16.7
dpc-svb-	89.72	2037	478	7570	15.8
dpc-vhs-	90.53	1862	461	6649	14.4
dpc-vla-	90.84	2765	1915	32157	16.8
wiki	89.20	2116	7343	98110	13.4
WR-P-E-C	83.36	1956	1014	12239	12.1
WR-P-E-E	82.65	5003	90	1813	20.1
WR-P-E-H	87.41	2281	2832	32222	11.4
WR-P-E-I	87.29	4469	9785	199150	20.4
WR-P-E-J	86.71	5484	699	15015	21.5
WR-P-P-B	92.97	322	275	2008	7.3
WR-P-P-C	87.23	2229	5648	83590	14.8
WR-P-P-E	87.45	4020	306	5808	19.0
WR-P-P-F	81.74	5117	397	6499	16.4
WR-P-P-G	78.52	11906	279	6468	23.2
WR-P-P-H	90.89	2279	2267	37241	16.4
WR-P-P-I	89.96	3745	5789	115933	20.0
WR-P-P-J	86.02	7034	1264	30021	23.8
WR-P-P-K	88.99	3732	351	6982	19.9
WR-P-P-L	87.68	3661	1115	20662	18.5
WS	90.36	1799	14032	205944	14.7
total	88.81	3115	65202	1096184	16.8

Table 1: Evaluation results of the Alpino parser on the various parts of Lassy Small, as well as the total corpus. We list the name of the sub-part of the corpus, the accuracy (as usual in terms of named dependencies), the required parse-time in milliseconds CPU-time, the number of sentences, words, and average word length.

It is also possible to search for a specific relation between pairs of words, for example for all sentences that mention a situation that seemed something: ”word=’situatie’ and hword=’leek’ and rel=’su’”:

1. De **situatie** leek immers hopeloos toen op 10 mei 1940 een oppermachtig Nazi-Duitsland België binnenviel .

    - situatie:N:su:WW:leek

Users have access to more information than displayed here, for example the complete parse trees of the sentences in the search results and statistics for the relations found by the search. The web interface was presented in a talk at CLIN 2010. It was also used in an event for high school students, which applied it for finding relevant corpus material without any problems.

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.