

ru | STEVIN

Dutch-Flemish research programme for
Dutch Language and Speech Technology

STEVIN Project Progress Reports



May, 2008

ñ | STEVIN

INTRODUCTION

In this report the progress reports of the STEVIN R&D projects have been gathered. In total at the moment nineteen R&D projects have been funded by the Dutch-Flemish STEVIN programme.

The five R&D projects that were funded in the first open call have now (almost) finished. Final status reports for these projects have been included in this report. As some of the projects have not yet delivered all their results, some of these report still need to be revised and completed in the next few months.

The six R&D projects funded in the second open call and two tender projects are well on their way and most of them will finish within the next 12 months. Of these projects some mid term progress reports have been included in this report. Unfortunately, lack of time prevented us from unifying the layout of these progress reports.

Finally, five projects were granted in the third and final open call, these projects have only just started and so has the start-up phase of the SoNaR project. For these six therefore, no progress reports are available as yet. Summaries of the aims of these projects are available in the STEVIN mid-term Fact File.

Some additional comments on the progress reports by members of the STEVIN PC:

STE04017 JASMIN

- The descriptions of the planned content of the corpus (after the first table) mentions 88hrs of speech, yet on page 1 the report mentions 95hrs. This looks like a contradiction. Also it appears that in the end 111hrs have been recorded, but it is unclear how many hours have been produced with the required annotations.
- There is no information on the external validation by BAS as to whether the data were sent to BAS and when they expect to receive the report.

STE05024 DAESO

- On page 9 the report states that the sentence splitter and tokenizer of the D-COI project appears "more error prone than expected". This statement must not be interpreted (now) as a validated critique since it is possible that DAESO members in Tilburg have themselves collaborated on the D-COI software. More clarification is needed.

STE05026 DPC

- It is still unclear from the report what the chances are that the project will be completed as planned.

STE05030 MIDAS

- It would be better to mention the company Nuance instead of SSFT (Scansoft) as a partner in all the relevant tables.

STE05035 STEVINcanPRAAT

- The report does not fully meet the qualitative requirements. Some parts have been assembled in a less than structured form.

STE05038 SPRAAK

- On page 4 of the report the authors mention the STEVIN project DISCO as a STEVIN 'Demonstration' project. This is incorrect. DISCO is a 'regular' R&D project.

Proposals funded in the 1st Call for Proposals for strategic research proposals and HLT resources (data & tools)

<i>acronym</i>	<i>coordinating institute and other academic partners</i>	<i>industrial partners</i>	<i>STEVIN priorities addressed (subject)</i>	<i>planned duration</i>	<i>funding</i>
COREA STE04005	Groningen University (Gosse Bouma) Antwerpen University	Language and Computing	Language resources Language research (semantic annotation)	24 mnths	€ 353.875
D-coi STE04008	Radboud Univ. Nijmegen-CLST (Nelleke Oostdijk) Tilburg University Antwerpen University Twente University Utrecht University Groningen University Leuven University	Polderland	Language resources Speech resources (Corpus written Dutch protocols)	14 mnths	€ 566.531
AUTONOMATA STE04014	Ghent University (Jean-Pierre Martens) Radboud Univ. Nijmegen Utrecht University	TeleAtlas Scansoft	Speech resources (speech synthesis)	24 mnths	€ 322.848
JASMIN-CGN STE04017	Radboud Univ. Nijmegen- CLST (Catia Cucchiarini) Leuven University	TalkingHome	Speech resources (speech corpus)	24 mnths	€ 419.471
IRME STE04019	Utrecht University (Jan Odijk) Groningen University	Van Dale Lexicografie	Language resources Language research (semantic and syntactic annotation)	24 mnths	€ 389.500

Project name	COREA: Coreference Resolution for Extracting Answers
Project number	STE04005
Planned starting date project	01-05-2005
Real starting date project	01-08-2005
Planned end of project date	30-04-2007
Real end of project date	31-10-2007

Consortium partners

Information Science, University of Groningen, the Netherlands
Center for Dutch Language and Speech, University of Antwerp, Belgium
Language and Computing, Sint-Denijs-Westrem, Belgium

Names of participating researchers per partner

University of Groningen

- Gosse Bouma (coordinator)
- Anne-Marie Mineur (researcher)
- Geert Kloosterman (programmer)

University of Antwerp

- Walter Daelemans (coordinator)
- Veronique Hoste (researcher)
- Iris Hendrickx (researcher)

Language and Computing

- Jean-Luc Verschelde (coordinator)
- Frederik Coppens (researcher)
- Joeri Van Der Vloet (researcher)

1. Final report

1.1. Summary of the project

Coreference resolution is a key ingredient for the automatic interpretation of text. It has been studied mainly from a linguistic perspective, with an emphasis on establishing potential antecedents for pronouns. Practical applications, such as Information Extraction (IE), summarization and Question Answering (QA), require accurate identification of coreference relations between noun phrases in general. Computational systems for assigning such relations automatically, require the availability of a sufficient amount of annotated data for training and testing. For Dutch, annotated data is scarce and coreference resolution systems are lacking.

In this COREA project, a two-year project which started in July 2005, we aim to develop a robust system for assigning such relations automatically, and we will investigate the effect of making coreference relations explicit on the accuracy of systems for for IE and QA. We will annotate a limited amount of application-specific corpus material, which is required for the evaluation of the coreference resolution system in the context of IE and QA.

1.2. Overview deliverables (+time of delivery: planned and real)

Deliverable	Planned	Realized
WP 1 Guidelines	nov 2005	dec 2005 (1 st version) -- jul 2007 (final version)
WP 2 Annotation Tool	feb 2006	feb 2006(1 st version) - dec 2006 (final version)
WP 3 Corpus Annotation	may 2006	sep 2007
WP 4 Coreference Resolution Tool	jul 2007	jul 2007
WP 5 Application Dependent Evaluation	oct 2007	oct 2007

1.3. Changes in content of deliverables and motivation for those changes

All deliverables were carried out according to the project proposal.

The corpus contains 50% more material than planned. The corpus visualisation tools we developed were not foreseen in the original proposal.

1.4. Problems and solutions

The actual annotation task took more time than planned. This hindered the development of the resolution tool. A solution was found by including data from a previous project by the University of Antwerp as well. These data are not included in the final results due to IPR restrictions.

The medical corpus includes material from a Dutch medical encyclopedia. This prevents us from making the results available to all third parties (almost) free of charge. A IPR-agreement was realized between Spectrum publishers, the Stevin "*TST-centrale*" and the University of Groningen, which allows the TST-centrale to distribute the material to third parties. Use for non-research (commercial) purposes is charged extra.

1.5. Recommendations for future research

In the Corea project, the coreference resolution module follows the approach of (Soon et al, 2001) and checks for coreference relations between pairs of Nps. For future research we would propose a more globally oriented method that takes in to account previous decisions and found coreferential relations.

1.6. Dissemination of results

Publications

- Iris Hendrickx et al, Coreference Resolution for Extracting Answers in Dutch, submitted for LREC 2008.
- Véronique Hoste, Iris Hendrickx and Walter Daelemans, Disambiguation of the neuter pronoun and its effect on pronominal coreference resolution In: Lecture Notes in Artificial Intelligence. Text, Speech and Dialogue. Proceedings of the 10th International Conference}, Volume 4629, pp.48-55, Plzen, Czech Republic, 2007 [[pdf](#)]

- Iris Hendrickx, Véronique Hoste and Walter Daelemans,
Semantic and Syntactic features for Anaphora Resolution for Dutch
In: Springer LNCS proceedings of the CICLing-2008 conference, Haifa, Isreal, 2008, to appear
- Gosse Bouma and Geert Kloosterman,
Mining Syntactically Annotated Corpora using XQuery
In: Proceedings of the Linguistic Annotation Workshop (held in conjunction with ACL 2007), pp. Prague, Czech Republic, 2007 [[pdf](#)]
- Véronique Hoste, Iris Hendrickx and Lieve Macken,
The Referential versus Non-referential Use of the Neuter Pronoun in Dutch and English
In: Proceedings of Corpus Linguistics 2007, Birmingham, England, 2007 [[pdf](#)]
- Iris Hendrickx, Walter Daelemans
Adding Semantic Information: Unsupervised Clusters for Coreference Resolution,
Workshop on Machine Learning for Natural Language Processing, Amsterdam, Nederland, 2007 (poster) [[pdf](#)]
- Iris Hendrickx, Veronique Hoste and Walter Daelemans,
Evaluating hybrid versus data-driven coreference resolution, In: Lecture Notes in Artificial Intelligence. Anaphora: Analysis, Algorithms and Application, Volume 4410, pp. 137-150, 2007. (DAARC 2007) [[pdf](#)]
- Véronique Hoste and Antal van den Bosch,
A Modular Approach to Learning Dutch Co-reference Resolution. In: Proceedings of the First WAR Colloquium 2005. Cambridge Scholars Press. [still not officially published] [[pdf](#)]
- Véronique Hoste and Walter Daelemans,
Comparing Learning Approaches to Coreference Resolution. There is More to it Than 'Bias'.
In: Proceedings of the Workshop on Meta-Learning (held in conjunction with ICML-2005), pp. 20-27, Bonn, Germany, 2005.

Presentations

- Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Anne-Marie Mineur, Joeri Van Der Vloet, Jean-Luc Verschelde
COREA Wrap up: the corpus and evaluation
CLIN, 7 December 2007, Nijmegen
- Iris Hendrickx
COREA: Coreference Resolution for Extracting Answers
Stevin-dag 23 November 2007, Antwerpen
- Gosse Bouma and Geert Kloosterman,
Mining Syntactically Annotated Corpora using XQuery
Linguistic Annotation Workshop, ACL 2007, June, Prague

- Veronique Hoste
Framing discourse as a classification approach
Workshop on Machine Learning for Natural Language Processing, 16 mei 2007, Amsterdam
- Veronique Hoste, Iris Hendrickx, Walter Daelemans
The automatic resolution of "het" in a machine learning approach to Dutch coreference resolution
CLIN meeting, Computational Linguistics in the Netherlands, 12 January 2007, Leuven
- Iris Hendrickx, Veronique Hoste, Walter Daelemans
Evaluating hybrid versus data-driven coreference resolution
CLIN meeting, Computational Linguistics in the Netherlands, 12 January 2007, Leuven
- Iris Hendrickx
COREA: project on coreference for Dutch
ISO/TC 37/SC 4 WG Meeting, 9 January 2007, Tilburg
- Gosse Bouma
COREA: Coreferentie en informatie-extractie
TST-dag, 30 november 2006, Rotterdam
- Iris Hendrickx
COREA: Coreference Resolution for Extracting Answers
ATILA meeting, 15 november 2006, Corsendonk
- Gosse Bouma
Corea,
Stevin-dag, 11 september 2006, Antwerpen

Outreach activities

- Gosse Bouma,
COREA: Coreference Resolution for Extracting Answers, Stevin-themanummer of DIXIT 2006
(Dutch journal for language and speech professionals)

1.7. Exploitation of results

- The COREA-corpus has been used to develop and evaluate the coreference resolution module of the Question-Answering system *Joost* of the University of Groningen. Coreference resolution is used in the QA-system as a means to increase recall of the QA-system.
- Antwerpen participates in the Stevin DAESO project and focuses on the development of an automatic multi-document summarization application. One of the problems in automatic summarization is 'dangling anaphora', and we plan to use the COREA coreference resolution software to tackle this problem.

2. External validation

No external validation was carried out. In our project proposal, external validation was not foreseen.

It should be noted, however, that we did carry out an evaluation of inter-annotator agreement for the corpus annotation task, the performance of the coreference resolution component, and the contribution of coreference resolution in two applications (information extraction and question answering).

Project name	Dutch Language Corpus Initiative (D-Coi)
Project number	STE04008
Planned starting date project	1 June 2005
Real starting date project	1 August 2005
Planned end of project date	31 May 2006
Real end of project date	31 December 2006

Consortium partners

Centre for Language and Speech Technology (CLST), Radboud University Nijmegen
Induction of Linguistic Knowledge (ILK), Tilburg University
Polderland Language & Speech Technology bv, Nijmegen
Human Media Interaction (HMI), Twente University
Utrecht Institute of Linguistics (UiL-OTS), Utrecht University
Alfa-Informatica, Groningen University
Centre for Computational Linguistics (CCL), Katholieke Universiteit Leuven

Names of participating researchers per partner

CLST: Nelleke Oostdijk, Olga van Herwijnen, Lou Boves
ILK: Martin Reynaert, Antal van den Bosch
Polderland: Wilco Apperloo, Peter Beinema, Theo van den Heuvel
HMI: Roeland Ordelman, Hendri Hondorp, Thijs Verschoor, Franciska de Jong, Arjan van Hessen
UiL-OTS: Paola Monachesi, Jantine Trapman
Alfa-Informatica: Gertjan van Noord et al.
CCL: Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde

1. Final report**1.1. Summary of the project**

The project can be characterized as a preparatory project and aims to produce a blueprint for the construction of a 500-million-word corpus of contemporary written Dutch. This will entail the design of the corpus and the development (or adaptation) of protocols, procedures and tools that are needed for sampling data, cleaning up, converting file formats, marking up, annotating, postediting, and validating the data. In order to support these developments, a 50-million-word pilot corpus will be compiled, parts of which will be enriched with linguistic annotations. The pilot corpus is intended to demonstrate the feasibility of the approach. It will provide the necessary testing ground on the basis of which feedback can be obtained about the adequacy and practicability of various annotation schemes and procedures, and the level of success with which tools can be applied. Moreover, it will serve to establish the usefulness of this type of resource and annotations for different types of HLT research and the development of applications. The Center for Sprogteknologi (CST) will undertake the evaluation of the protocols and procedures. At the end of the project, the pilot corpus together with all other results obtained within the project will be made available through the Dutch HLT Agency (TST-Centrale).

1.2. Overview deliverables

Deliverables for the D-Coi project are the following:

1. Report corpus design, sampling and metadata
 - Oostdijk, N. 2006. *A Reference Corpus of Written Dutch. Corpus design* (TR D-COI-06-01)
2. Technical specification basic format and accompanying validation tools
 - Beinema, P. 2006. *A Reference Corpus of Written Dutch. Technical specification of file formats and validation tools used* (TR D-COI-06-02)
 - Reynaert, M. 2006. *Addendum to D-Coi Technical Report TR2*
 - Ordelman, R., H. Hondorp, T. Verschoor, F. de Jong. 2007. *Final Report D-Coi data processing, format conversion and corpus management*
3. Technical report on the evaluation of available tokenizers and sentence splitters
 - Reynaert, M. 2006. *Sentence splitting and tokenisation* (TR D-COI-06-03)
4. Technical report CICL (Corpus-induced Corpus Clean-up)
 - Reynaert, M. 2006. *Text normalisation and correction in D-Coi* (TR D-COI-06-04)
5. POS tagging and lemmatisation: Protocol. Documented version of POS tagger-lemmatiser with accompanying tools for consistency checking
 - Van Eynde, F. 2005. *Part of Speech Tagging en Lemmatisering van het D-Coi Corpus*. Centrum voor Computerlinguïstiek, KU Leuven. 88 pp. (TR D-COI-05-01)
6. Syntactic annotation: Protocol. Documented version of the Alpino parser
 - Noord, G. van. 2006. *Annotation manual D-Coi version of CGN manual* (TR-D-COI-06-06a)
 - Noord, G. van. 2006. *Alpino User Guide* (TR-D-Coi-06-06b)
 - Noord, G. van. 2006. *Manual for syntactic annotation* (TR-D-COI-06c)
 - Noord, G. van. 2006. *Alpino dtd* (TR-D-Coi-06d)
 - Kloosterman, G. 2006. *An overview of the Alpino Treebank tools* (TR-D-COI-06e)
 - Noord, G. van. 2006. *Automatically assigned syntactic constructions* (TR-D-COI-06f)

For the Alpino parser and data (both the data that were manually verified and the data that were automatically annotated), see <http://www.let.rug.nl/~vannoord/DCOI/>

7. Semantic annotation: The results of two pilot studies (one concerning the annotation of semantic roles, the other investigating the annotation of spatio-temporal aspects). The results comprise the protocols that were developed and the data that were annotated.
 - Trapman, J. and P. Monachesi. 2006. *Report on the Annotation of Semantic Roles* (TR DCOI-06-07a)
 - Trapman, J. and P. Monachesi. 2006. *Manual for Semantic Annotation in D-Coi*
For the data, see [gelabeldCorpus_31jan06.zip](#)
 - Schuurman, I. 2007. *A Reference Corpus of Written Dutch. MiniSTEx. Protocol. Version 01* (TR-D-COI-06-07b). For the data, see [corp-stex.xml.tar.gz](#)
8. COREX: Adapted version of the COREX software which originally had been developed for use with the Spoken Dutch Corpus (Corpus Geschreven Nederlands, CGN). The adapted version should

make it possible to exploit the data available from the D-Coi project. Documentation. The COREX software has already been delivered to the Dutch HLT Agency.

9. Pilot corpus with documentation about the composition, available annotation, protocols used, formats and tools. The corpus is available in XML format. Corrections with respect to the original source texts are always available in the form of separate annotations. The entire corpus has been tagged for part of speech, lemmatized and syntactically annotated. For part of the corpus the POS tagging, lemmatisation and syntactic annotation has been manually verified (POS tagging and lemmatisation 500,000 words; syntactic annotation 200,000 words). All data can be accessed by means of the adapted COREX software.

- Oostdijk, N. 2006. *Dutch Language Corpus Initiative Pilot Corpus. Corpus description* (TRD-COI-06-09)

10. Final report

- Oostdijk, N. 2007. Dutch Language Corpus Initiative (D-Coi). *Eindverslag*.
- Ordelman, R., H. Hondorp, T. Verschoor, F. de Jong. 2007. *Eindverslag D-Coi dataverwerking, formaatconversie en corpusbeheer*.

The evaluation of the results obtained in the D-Coi project is carried out by CST, Copenhagen. The results are expected to be available shortly.

All deliverables except for the tools (e.g. validation tools, POS tagger-lemmatiser) are available at the STEVIN wikisite. Parties responsible for various tools and data have negotiated the delivery of these resources directly with the HLT Agency.

1.3. Changes with respect to the original project proposal and motivation for those changes

The project proposal was awarded funding under the following conditions: (1) a substantial reduction of the original budget should be effected (the project was awarded funding in the amount of € 566,531 whereas total project costs for the project as originally planned amounted to € 721,531) and (2) the work packages involving the annotation of co-reference and multi-word units were to be cancelled, while € 40,000 should be spent on other types of semantic annotation. This led to a number of adaptations in the project plans, most notably

- the total duration of the project was reduced from 18 to 14 months (as a side-effect of which the management costs were reduced);
- the annotation of co-reference and multi-word units was cancelled; instead two pilot studies were defined, one of which aimed to investigate the annotation of semantic roles, while the other aimed to investigate the annotation of spatio-temporal relations.

Starting date and end of project date

The formal starting date of the D-Coi project was 1 June 2005, i.e. the date mentioned in the letter awarding funding to the project and actually 5 months later than anticipated when the proposal was submitted. This meant that not all partners were able to make the necessary personnel available: in view of the delay they had meanwhile been assigned to other tasks. In view also of the summer holidays that were approaching it was decided to consider 1 August 2005 as the actual starting date. Towards the end of the project a request was put up for an extension. The end of project date was subsequently moved to 31 December 2006.

In their last consortium meeting, on 20 December 2006 the partners agreed to deliver all results by 31 January 2007 at the latest. The deadline was met by all partners except one: Polderland failed

to deliver the integrated data. This was due to a serious skiing accident in which Wilko Apperloo was severely injured. The project manager duly informed the STEVIN programme office and negotiated a different time plan for the evaluation by CST. Subsequent delays at Polderland have continued to frustrate the progress of the evaluation. The final report is expected shortly.

Project management

The project manager was responsible for all communication with the STEVIN programme office, drafting the consortium agreement, and managing the project's finances. On various occasions the project manager on behalf of the consortium presented the project to a wider audience. The project manager was also responsible for convening project meetings which aimed to discuss progress and possible problems/solutions.

On two occasions the project manager decided to transfer money from Nijmegen to Tilburg. Thus money was made available to Tilburg for (1) the conversion of pdf-files and (2) extending Martin Reynaert's contract. As regards the conversion of pdf-files: in Tilburg students were available that could work directly under the supervision of Martin Reynaert. As regards Martin Reynaert's contract: the one-year D-Coi contract was due to end by 1 October 2006. By then Tilburg would have spend all of the money allocated to Tilburg. Since Martin Reynaert held a crucial role in the project (he was responsible for all the work pertaining to normalisation of the data, spelling checking, tokenisation and sentence splitting) it was decided to re-allocate some of the Nijmegen money for this purpose. Consequently, Nijmegen's role in collecting metadata was largely cancelled.

1.4. Summary of the D-Coi project: achievements, problems and solutions

Internal communication

For purposes of project internal communication a mailing list (SURFNet) was set up. In addition a wiki-site was hosted by ILK Tilburg. Project meetings were convened on a regular basis. At these meetings overall progress was discussed. Minutes of these meetings were placed on the project's wiki-site. Where specific tasks were concerned (such as the specification of protocols and procedures) meetings were convened which involved only the relevant partners.

Website

An (English) website was launched and hosted by Nijmegen which aimed to inform the general public about the project. (<http://lands.let.ru.nl/projects/d-coi>)

Corpus design

A design was made for a reference corpus of written Dutch. To this end the available literature was consulted and experiences obtained in other large scale corpus projects (e.g. BNC, ANC, CGN) were taken into account. In addition a user requirements study was conducted. The design is ambitious, not only because at the moment there is no comparable corpus in terms of size and composition that we are aware of that is accessible in the fashion that we envisage. Apart from more conventional genres the corpus is to include also texts from new media. The motivation for the design and the choices that have been made have been described in a report entitled A Reference Corpus of Written Dutch. Corpus design (TR-D-COI-06-01).

Acquisitie en IPR

Throughout project text acquisition has been hindered by the fact that content owners do not seem to be very keen on making their texts available for inclusion in a corpus that is to be distributed. In some cases (such as Mediargus) it appears that parties have very large quantities of data without

owning the rights. While we were informed that they would consider making these data available for the duration of the project/a specific project (the data must be discarded after the project ends), they refused to make the data available for use and distribution as envisaged in the DCoi/ SoNaR projects.

Initially in the D-Coi project the strategy adopted was to start work on data only if IPR had been properly arranged. However, as we found that this kept us from making any serious progress, we decided to focus on data for which IPR was either unproblematic or negotiable in due course. As a result, the D-Coi corpus comprises texts from internet, for instance texts from government and other public websites. While in these cases we expect IPR to be rather unproblematic, it is not at all clear who holds the rights and therefore should be approached to obtain permission.

It is expected that in due course proper IPR arrangements can be made for the news texts (newspapers and autocues) from the Twente Nieuwscorpus and also the KNACK data (made available by the COREA project). Unproblematic are the texts that are available under GPL (Wikipedia, JRC and Europarl), while for texts that were obtained from Neder-L, Dedicon (previously FNB), NTU and Uitgeverij de Harmonie IPR has been settled. Permission was obtained from DARENet for the inclusion in the corpus of theses, abstracts etc. This, however, was too late in the project to be further explored. Franciska de Jong has established various contacts that may be further pursued in the SoNaR project.

The establishment of the STEVIN IPR committee which is concerned with IPR issues, including those relating to data acquisition, has proven to be useful. The draft contracts on behalf of the Dutch Language Union make it easier to convince content owners of the importance of the compilation of resources like the Dutch reference corpus. Unfortunately, this development has occurred fairly late in the D-Coi project so that the effect was limited.

Data processing, format conversion and corpus management

All D-Coi data were stored centrally on a computer that had been purchased and installed in Twente especially for this purpose. For corpus management a simple database was installed. Various processing stages and annotation activities were carried out using local copies. Once completed the results were then uploaded on the central computer. Until the end of the project there were problems relating to the status of certain files. This underlines the need to have a protocol in place right from the very start in which clear procedures are laid down pertaining to the download and upload of files (see also the report by Ordelman et al.).

Data conversion has appeared to be one of the most annoying and persistent problems experienced in the D-Coi project. The effort that was required had clearly been underestimated. As a result, considerable time and manpower was spent on this task.

Sentence-splitting and tokenisation

Available tokenizers were evaluated (see TR-D-COI-06-03). From the tokenizers owned by the partners only two were included in the evaluation, since only these (the Alpino tokenizer and the ILK tokenizer) could also do sentence-splitting. Although both tokenizers performed equally well, it was decided to use the ILK tokenizer for sentence-splitting and tokenisation in D-Coi project because the people who were going to use the tool were already familiar with it. The tool was adapted slightly for use in D-Coi. The adapted version has been made available to the HLT Agency.

Text normalisation and spelling correction

In order to increase the usability of the corpus it was considered desirable to remove any irregularities from the texts. This process was sometimes called 'corpus clean-up', at other times

'normalisation'. It included also the correction of certain types of spelling errors. More details can be found in the document entitled Text normalization and correction in D-Coi (TR-D-Coi-06-04). It should be pointed out that it was not the ambition of the D-Coi project to have all text data conform to the current official spelling guidelines (NTU 2005). Therefore users of the D-Coi corpus will find that they may come across different spelling variants, including those that date from earlier periods.

Logistics

The shortened duration of the project combined with the fact that Leuven was the one party that actually started work on the D-Coi project on 1 June 2005 gave rise to problems with the availability of data. While initially work focused on adaptation of the protocol for POS tagging and lemmatisation, it soon became clear that as data acquisition did not run as smoothly as foreseen and data conversion turned out to be more problematic than anticipated there was going to be a shortage of data for Leuven to manually verify. Therefore it was decided to have Leuven (and later also Groningen) work on raw data, that would later be 'upgraded' to meet the project's standards.

POS tagging and lemmatisation

Both the protocol and the tagger-lemmatiser that had been used in the Spoken Dutch Corpus project were adapted for use with written data (Van Eynde 2005: Part of Speech Tagging en Lemmatisering van het D-Coi Corpus). For approx. 500,000 words the POS tagging and lemmatisation were manually verified. Unlike the procedure adopted in the Spoken Dutch Corpus project, in the D-Coi project we developed an approach such that only certain tokens were inspected and if necessary corrected when the POS tagger-lemmatiser indicated that the probability of a tag/lemma was below a certain threshold. At the end of the project all the data were tagged and lemmatised automatically, without manual verification.

Syntactic annotation

For syntactic annotation the Alpino parser was used. A subset of 200,000 words was manually verified, partly in Flanders (Leuven) and partly in the Netherlands (Groningen). The entire corpus was automatically parsed. Adaptations to the Alpino parser and accompanying tools, the format used have been extensively documented. For users a manual is available. Syntactic annotation is continued within the Lassy project which aims to produce a Treebank for Dutch.

Semantic annotation

Two pilot studies were carried out, one investigating the annotation of semantic roles, the other the annotation of spatio-temporal relations. In both cases an annotation scheme has been developed. The schemes were put to the test while manually annotating modest quantities of data. The money allocated to the task did not permit the development of tools.

Pilot corpus

IPR issues have had a strong impact on the compilation of the pilot corpus. As priority was given to the collection of data that would support the overall progress of the project, data were collected opportunistically. Consequently, the D-Coi corpus does not in its composition reflect to the full the design and selection criteria that have been proposed for the large reference corpus. Certain types of data (including SMS and email) were not available to the D-Coi project. No experience has been gained in how to process these types, nor do we know at this point what difficulties to expect if we attempt to annotate these data.

COREX

It had been envisaged that in the D-Coi project the COREX software would be adapted so that it would be possible to use the same software for the storage and exploitation of the written data and the spoken data (from the CGN project). With additional funding that was available from NWO the plans for the adaptation of COREX were changed in the sense that the adapted version should also cater for other STEVIN corpus projects (such as Jasmin-CGN). The new version of COREX was delivered to the HLT Agency early March 2007.

External evaluation

CST, Copenhagen is responsible for the external validation of the results. Originally it was envisaged that the evaluation would take place between February and May 2007. In the light of the events, half January 2007 it was decided to adapt the time schedule. Although CST received materials batchwise from 1 February 2007 onwards, the deadline for the delivery of the integrated data was not met and further delay in doing so has prevented CST from completing the evaluation for several months.

1.5. Dissemination of results

Publications

The following publications have appeared pertaining to (various aspects of) the D-Coi project:

Conference proceedings (7):

1. Bosch, A. van den, I. Schuurman and V. Vandeghinste. Transferring POS tagging and lemmatization tools from spoken to written Dutch corpus development. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa - Italy, 22-28 May 2006. Genua, 2006.
2. Noord, G. van. I. Schuurman, and V. Vandeghinste. Syntactic annotation of large corpora in STEVIN. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa - Italy, 22-28 May 2006. pages 1811-1814. Genua, 2006.
3. Oostdijk, N. and L. Boves. User requirement analysis for the design of a reference corpus of written Dutch. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa - Italy, 22-28 May 2006, pages 106-1211. Genua, 2006.
4. Reynaert, M. Corpus-induced corpus clean-up. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa - Italy, 22-28 May 2006. pages 87- 92. Genua, 2006.
5. Schuurman, I. and P. Monachesi. The Contours of a Semantic Annotation Scheme for Dutch. In Proceedings 16th Meeting of Computational Linguistics in the Netherlands. Edited by K. Sima'an, M. de Rijke, R. Scha and R. van Son. pages 67-82. Amsterdam: University of Amsterdam. 2006.
6. Trapman, J. and P. Monachesi. Where FrameNet meets the Spoken Dutch Corpus: in the middle. In Proceedings 16th Meeting of Computational Linguistics in the Netherlands. Edited by K. Sima'an, M. de Rijke, R. Scha and R. van Son. pages 99-116. Amsterdam: University of Amsterdam. 2006.
7. Oostdijk, N., M. Reynaert, P. Monachesi, G. van Noord, R. Ordeman, I. Schuurman, and V. Vandeghinste. From D-Coi to SoNaR: A reference corpus for Dutch. In Proceedings LREC 2008.

Workshop proceedings (1):

1. Monachesi, P. and J. Trapman. Merging FrameNet and PropBank in a corpus of written Dutch. In Proceedings of the workshop Merging and layering linguistic information. Workshop held in conjunction with LREC 2006, Genoa - Italy, 23 May 2006. pages 32-39. Genua, 2006.

Other (1):

1. Oostdijk, N. Aanzet tot een Nederlandstalig tekstcorpus. D-Coi (Dutch Language Corpus Initiative). In Dixit. Tijdschrift voor toegepaste taal- en spraaktechnologie. Jaarboek. Jaargang 4, nummer 2, page 25. 2006

Presentations

The D-Coi project has been presented at various conferences, workshops and seminars:

Taal in bedrijf, 22 November 2005, Eindhoven

Poster presentation D-Coi project and demo syntactic annotation in D-Coi (Alpino parser)

CLIN-16, Computational Linguistics in The Netherlands. 16 December 2005, Amsterdam

Oral presentations (4):

1. Paola Monachesi and Ineke Schuurman: The contours of a semantic annotation scheme for Dutch
2. Nelleke Oostdijk: Dutch Language Corpus Initiative (D-Coi)
3. Martin Reynaert: Anagram-key based tokenizer evaluation
4. Jantine Trapman and Paola Monachesi: Where FrameNet meets the spoken Dutch Corpus: in the middle

Corpusdag, seminar organised by the Dutch HLT Agency. 23 March 2006, Rotterdam

Oral presentation (1):

1. Gertjan van Noord: Syntactische annotatie in D-COI en Lassy

Workshop Merging and layering linguistic information, workshop in conjunction with LREC2006

Oral presentation (1):

1. Paola Monachesi and Jantine Trapman: Merging FrameNet and PropBank in a corpus of written Dutch

LREC2006, International Conference on Language Resources and Evaluation. 22-28 Mei, Genua (Italy)

Oral presentations (4):

1. Antal van den Bosch, Ineke Schuurman and Vincent Vandeghinste: Transferring POS tagging and lemmatization tools from spoken to written Dutch corpus development
2. Gertjan van Noord, I. Schuurman and Vincent Vandeghinste: Syntactic annotation of large corpora in STEVIN
3. Nelleke Oostdijk and Lou Boves: User requirement analysis for the design of a reference corpus of written Dutch
4. Martin Reynaert: Corpus-induced corpus clean-up

STEVIN workshop, 11 september 2006, Antwerp

Presentation (1):

Nelleke Oostdijk: Dutch Language Corpus Initiative (D-Coi)

CLIN-17, Computational Linguistics in the Netherlands. 12 January 2007, Leuven

Oral presentations (4):

5. Wilko Apperloo and Remco van Veenendaal: COREX 7.0: corpus exploration software revisited
2. Paola Monachesi and Jantine Trapman: Annotating semantic roles in the Dutch Language Corpus
3. Martin Reynaert: TICCL and Typos

4. Ineke Schuurman: Spatiotemporal annotation for Dutch

STEVIN-programmadag, 21 september 2007, Hoeven
Poster presentation D-Coi

LREC2008, International Conference on Language Resources and Evaluation. 28-30 May, Marrakech (Morocco)

Presentations (2):

1. Oostdijk, N., M. Reynaert, P. Monachesi, G. van Noord, R. Ordelman, I. Schuurman, and V. Vandeghinste. From D-Coi to SoNaR: A reference corpus for Dutch. Poster presentation.
2. M. Reynaert. From D-Coi to SoNaR: A reference corpus for Dutch. Oral presentation.

Outreach activities

A website was launched to inform the wider public of the initiative:

<http://lands.let.ru.nl/projects/d-coi>

1.6. Exploitation of results: New project proposals

The creation of a treebank for Dutch will be further pursued within the Lassy project. The compilation of the planned 500 MW reference corpus is foreseen within the SoNaR project. Both projects receive funding from the STEVIN programme.

2. External validation

2.1. Name organisation that performed the external validation

CST, Copenhagen.

2.2. Delivery date external validation report

No exact date has been given.

2.3. Summary conclusions external validation report

Project name	Autonomata
Project number	STE04014
Planned starting date project	01-06-2005
Real starting date project	01-06-2005
Planned end of project date	31-05-2007
Real end of project date	31-05-2007

Consortium partners

1. Electronics & Information Systems (ELIS), Gent
2. Centre for Language and Speech Technology (CLST), Radboud Universiteit Nijmegen
3. Instituut voor Linguïstiek –OTS (UiL-OTS), Universiteit Utrecht
4. Nuance Communications International bvba (NCI) (voorheen ScanSoft), Merelbeke
5. Tele Atlas (TA), Gent

Names of participating researchers per partner

Ghent University

Jean-Pierre Martens, ELIS (Project coordinator)
Kristof D'Hanens, ELIS
Qian Yang, ELIS (until 31/5/2006)

Radboud University

Henk van den Heuvel, CLST (Project leader)
Nanneke Konings, CLST
Corina Koolen, CLST

University Utrecht

Gerrit Bloothoofd, UiL-OTS (Project Leader)
Michiel Hildebrand, UiL-OTS
Gerwert Stevens, UiL-OTS

Nuance Communications

Jan Verhasselt (Project Leader)
Robrecht Comeyne
Sigrid Falley
Bart Baeyens
Thomas Kuehnel
Johan Smolders
Qian Yang (from 1/6/2006 on)

TeleAtlasLuc Peirlinckx (Project Leader)

Lieven Luypaert
Mieke Verheye
Karen Windey

1. Final report

1.1. Summary of the project

In many modern applications such as directory assistance, name dialing, car navigation, etc. one needs a speech recognizer and/or a speech synthesizer. The former to recognize spoken user commands and the latter to pronounce information found in a database. Both components make use of phonetic transcriptions of the words to recognize/pronounce. In order to develop an application, the developer needs a tool that accepts words/sentences and that returns the phonetic transcriptions of these words/sentences. The first goal of this project was to develop such a tool that incorporates a state-of-the-art grapheme-to-phoneme converter (the one from Nuance), as well as a dedicated phoneme-to-phoneme (p2p) post-processor which can automatically correct some of the mistakes which are being made by the standard g2p. Dedicated post-processors were developed for person names and geographical names.

A problem, especially in view of the recognition of names, is the existence of different pronunciations for the same name. These pronunciations often depend on the background (mother tongue) of the user. Typical examples are the pronunciation of foreign city names, foreign proper names, etc. The second goal of the project was to collect a large number of name pronunciations and to provide manually corrected phonetic transcription of these name utterances. Together with meta-data on the speakers, this corpus is expected to become a valuable resource in the research towards a better name recognition.

1.2. Overview deliverables (+time of delivery: planned and real)

D1: transcription tool (g2p) that can hold p2p converters	T15	T18
D2: p2p converters for person and place names	T15	T18
D3: report on inductive en deductive approach to p2p learning	T15	T18
D4: software for p2p learning	T15	T18
D5: report on internal validation of p2p learning software	T15	T18
D6: corpus of spoken person and place names	T24	T24
D7: report on internal validation during corpus development	T24	T24
D8: report on external validation of transcription tool	T24	T24
D9: report on external validation of spoken name corpus	T24	T24

1.3. Changes in content of deliverables and motivation for those changes

All deliverables were delivered according to plan, with this exception that it was possible to create a manually verified transcription of **all** the spoken names, which is more than was promised.

1.4. Problems and solutions

Although the deliverables were physically transferred to the NTU, the official transfer has not yet been realized. There was general agreement among the consortium partners on the underlying principles of this transfer for a long time, but the STEVIN-IPR commission and the NTU had to agree as well. In the mean time the negotiations have converged to a license agreement that will soon be ready for signing by the different parties.

1.5.Recommendations for future research

On the basis of the external validation of the spoken name corpus, there is evidence that the quality of the manually verified phonetic transcriptions of the foreign names occurring in the Flemish part of the corpus could be further improved by performing an additional check round.

The external validation report also showed that a small fraction of the recordings show strong wind noise. It would be possible to extend the metadata with information on the presence of wind noise in the recordings (per speaker).

1.6.Dissemination of results

Publications (all on public website)

- Q. Yang, J.-P. Martens, N. Konings, H. van den Heuvel (2006). "Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names",. Proceedings LREC (Genua), 287-292.
- H. van den Heuvel, J.P. Martens, N. Konings (2007). "G2P conversion of names. What can we do (better)?", Proceedings Interspeech (Antwerp), 1773-1776.
- H. van den Heuvel, J.P. Martens, B. D'hoore, K. D'hanens, N. Konings (2008). "The Autonomata Spoken Name Corpus. Design, recording, transcription and distribution of the corpus" to appear in Proceedings LREC (Marrakech).
- H. van den Heuvel, J.P. Martens, N. Konings (2008). "Fast and easy development of pronunciation lexicons for names", Proceedings LangTech(Rome), 117-120.

Presentations

- Lecture for the Dutch Society of Phonetic Sciences (H. van den Heuvel, Amsterdam, June 9, 2006)
- Lecture at first STEVIN program day: AUTONOMATA (J.P. Martens, Antwerp, september 11, 2006)
- Lecture at the Fourth Thematic day of the TST Central "STEVIN: de gebruiker centraal" (H. van den Heuvel, Rotterdam, November 30, 2006)
- Lecture at the Fifth Thematic day of the TST Central "STEVIN: de gebruiker centraal" (J.P. Martens, Antwerp, November 23, 2007)

Outreach activities

- "AUTONOMATA: naar een betere behandeling van eigennamen in spraaktechnologie" (contribution to special issue of DIXIT, a Dutch journal for language and speech professionals)
- "Albertlaan, niet Albertláán", Senternovem Monitor 4 (2006)

1.7.Exploitation of results

- The p2p technology and the spoken name corpus will be exploited in a follow-up research project, called Autonomata Too, which is being funded by the STEVIN programme under Call 3.
- The transcription tools have been used to generate the pronunciation lexicons of the speech recognizers that are developed in the context of NBEST and NEON (two other STEVIN projects).
- Partner ELIS has used the spoken name corpus in its research on cross-lingual aspects of automatic speech recognition, and this research has already resulted in several publications.

2. External validation

2.1. External validation of the software tools

Organization:

The transcription tools incorporating the Nuance g2p have been evaluated by Dutcheer BV

Delivery:

The validation resulted in a validation report that is transferred to the NTU as deliverable D8 (delivery date was T22).

Summary of conclusions:

The main conclusion was that an analysis of the difference in total weighted error points of Nuance and Autonomata shows that Autonomata scores 30% less weighted points than Nuance, when stress errors are excluded, which means an improvement for Autonomata. If stress would be included in the comparison, this score would reduce to 23% less weighted points.

There were also some recommendations for further improvements, some of which (m+b bug, consonant doubling and words without any stress) have in the meantime been fixed.

2.2. External validation of the spoken name corpus

Organization:

The spoken name corpus was evaluated by BAS (Munich)

Delivery:

The external validation resulted in three validation reports together constituting deliverable D9. Two reports on the validation of the example transcriptions and the transcription protocol (available at T12) and one report on the full corpus distribution (available at T24).

Summary of conclusions:

The first two reports did not really formulate conclusions but mainly confirmed that the lexicon and the transcription protocol were ok. The final corpus validation report mentions the following conclusions:

“The findings presented in the validation report indicate a good quality of the AUTONOMATA corpus. In a hypothetical ranking from 1 (not usable) to 20 (without flaws) we would assign the AUTONOMATA corpus the mark 17.

The English documentation could be improved by clearly separating the *specification* from the *documentation* of the corpus (see remarks in 1.3.15). A standard checklist for speech corpora documentation ([1]) revealed a moderate number of missing informations that can easily provided by the producer for the next corpus release.

The quality of the recordings is sufficient for technical applications and phonetic investigations. More emphasis might have been given to a proper placement of microphones to avoid blowing noise (approx. 11% of the recordings).

The manual validation of the transcription in two subsamples of 1000 each revealed a moderate deviation between transcriber's and validator's opinions. We deem that these deviations are still in an acceptable range for a speech corpus of this kind.”

Responses to the external validation:

Most of the recommendations made by the external validator were taken into account before transferring the corpus to the NTU. How exactly the consortium responded is documented in a report D9_antwoord_op_BAS which is also transferred to the NTU as part of deliverable D9.

Project name	Title: Extension of CGN with speech of children, non-natives, elderly and human-machine interaction (JASMIN-CGN) (Jongeren, Anderstaligen, Senioren en Machine Interactie voor het Nederlands)
Project number:	STE04017
Planned starting date project:	1 April 2005
Real starting date project:	1 April 2005
Planned end of project date:	30 September 2007
Real end of project date:	31 December 2007

Consortium partners

Name	Affiliation
Dr. C. Cucchiarini	CLST, Radboud University Nijmegen
Prof. dr. H. Van hamme	ESAT, Katholieke Universiteit Leuven
Dr. ir. F.M.A Smits	TalkingHome

Names of participating researchers per partner

- CLST
 - a. Catia Cucchiarini
 - b. Andrea Diersen, left unexpectedly on 01-12-2005
 - c. Olga van Herwijnen, on maternity leave from 01-07-2006 until 31-10-2006, left unexpectedly on 22-03-2007
 - d. Leontine Aul, from 01-03-2006, left unexpectedly on 31-08-2007
 - e. Eric Sanders, from 01-09-2007
- ESAT
 - a. Hugo Van hamme
 - b. Maarten Van Segbroeck, left on 01-01-2006 to pursue a "specialisatiebeurs IWT"
 - c. Alain Sips, left unexpectedly on 31-03-2006
 - d. August Oostens, from 1-04-2006
 - e. Joris Driesen, from 1-02-2007
- TalkingHome
 - a. Felix Smits
 - b. Barry van der Veen, left unexpectedly on 16-06-2006
 - c. Erik Stegeman, left unexpectedly in March 2007.
 - d. Chantal Mülders, left unexpectedly on 01-12-2006
 - e. Koen Snijders, from 01-05-2007 to 01-08-2007

1. Final report

1.1. Summary of the project

The aim of JASMIN-CGN project was to extend the Spoken Dutch Corpus (Corpus Gesproken Nederlands -CGN- a corpus of about 9 million words that constitutes a plausible sample of standard Dutch as spoken in the Netherlands and Flanders and contains various annotation layers) in three dimensions. First, by collecting a corpus of contemporary Dutch as spoken by children of different age groups, non-natives with different mother tongues and elderly people in the Netherlands and Flanders (JAS-CGN), an extension along the age and mother tongue dimensions was achieved. In addition, speech material was collected in a communication setting that was not envisaged in the CGN: human-machine interaction. These three dimensions are reflected in the corpus as five user groups: native primary school pupils, native secondary school students, non-native children, nonnative adults and senior citizens. For each group, both read and human-machine-interaction data had to be collected for a total of about 95 hours. One third of the data was to be collected in Flanders and two thirds in the Netherlands.

1.2. Overview deliverables (+time of delivery: planned and real)

Code	Workpackage	Parties	Deliverables	Delivery time
A1	Finalization of corpus design	CLST, ESAT, TH, user group	Specifications for corpus design	Planned: 30-06-2005 Real: 30-06-2005
A2	Recording platform and dialog development	TH, CSLT, ESAT	Written report Recording platform dialog specifications	Planned:31-07-2005 Real: 31-08-2005
A3	Speakers recruiting, Speech recordings, Metadata	CSLT, ESAT, TH	Specifications of speakers, recordings and metadata. Speech files Metadata	Planned: 31-03-2006 Real: 31-07-2007
B1	Orthographic annotation	CSLT, ESAT	Orthographic annotations	Planned: 31-01-2007 Real: 31-08-2007
B2	Automatic phonetic transcription	CSLT, ESAT	Automatic phonetic transcriptions pronunciation lexicon report	Planned: 31-03-2007 Real: 30-11-2007
B3	POS tagging and lemmatization	CSLT, ESAT	Protocol specifying the tagset and its application. Protocol. POS tagger + documentation	Planned: 31-03-2007 Real: 30-11-2007
B4	Quality checking	CLST, ESAT	Written report	
C1	IPR	HLT agency	IPR regulation	
C2	Validation and Evaluation	ELDA	Written Report	
C3	Dissemination of results	CLST, ESAT, TH	Website, conference papers	
D	Project Management	CLST, ESAT, TH	Progress reports and final report	

Corpus:

Note: Due to the 5% budget cut, the proposed corpus size was reduced from 100 to 95 hours divided as follows:

In The Netherlands:

- native children between 7 and 11 (12h 21m)
- native children between 12 and 16 (12h 21m)
- non-native adults (12h 21m)
- non-native children between 7 and 14 (12h 21m)
- native adults above 65 (9h 26m)

In Flanders:

- native children between 7 and 11 (6h 10m)
- native children between 12 and 16 (6h 10m)
- non-native adults (6h 10m)
- non-native children between 7 and 14 (6h 10m)
- native adults above 65 (5h 5m)

Where not specified, about 50% of the material should be read speech and 50% extemporaneous speech recorded in the human-machine interaction modality (HMI).

1.3. Changes in content of deliverables and motivation for those changes

An important aspect of the realisation of this corpus that remained underspecified in the original deliverable description is the annotation of the human-machine interaction phenomena that were elicited through the dialogues. For this part of the annotation a new protocol had to be drawn up for transcribing the HMI phenomena, which had not been made for CGN, since this type of annotation was not envisaged in CGN. Therefore, it seems appropriate to include this transcription protocol in the list of deliverables and, for that matter, also the one for the orthographic transcription, which is a slightly adapted version of the one used for CGN. On the other hand, for quality checking no separate report was made. Quality checking was carried out for the orthographic transcriptions, for the annotations of the HMI phenomena and for the automatically generated phonemic transcriptions. In all cases the orthographic transcriptions and the HMI annotations were made by one transcriber and checked by a second transcriber who listened to the sound files, checked whether the transcription was correct and, if necessary, improved it. The speech material recorded in the Netherlands was also transcribed in the Netherlands, whereas the speech material recorded in the Flanders was transcribed in Flanders. To prevent the inconsistencies in the transcription, cross checks were carried out.

The quality of the automatically generated phonemic transcriptions was verified for 3 randomly selected files per Region (FL/NL) and category (non-native child, non-native adult, native child and senior) (a total of 24 recordings) by inspection of the proposed transcription. Lexicon and crossword assimilation rules were adapted to minimize the number of errors. Most of the required corrections involved hard/soft pronunciation of the “g” and optional “n” in noun plurals and infinitive forms.

Code	Workpackage	Parties	Deliverables	Delivery time
A1	Finalization of corpus design	CLST, ESAT, TH, user group	Specifications for corpus design	Planned: 30-06-2005 Real: 30-06-2005
A2	Recording platform and dialog development	TH, CSLT, ESAT	Written report Recording platform dialog specifications	Planned: 31-07-2005 Real: 31-08-2005
A3	Speakers recruiting, Speech recordings, Metadata	CSLT, ESAT, TH	Specifications of speakers, recordings and metadata. Speech files Metadata	Planned: 31-03-2006 Real: 31-07-2007

B1	Orthographic annotation	CSLT, ESAT	Orthographic annotations ((in Dutch and English) HMI annotations + protocol (only in Dutch)	Planned: 31-01-2007 Real: 31-08-2007
B2	Automatic phonetic transcription	CSLT, ESAT	Automatic phonetic transcriptions pronunciation lexicon report	Planned: 31-03-2007 Real: 30-11-2007
B3	POS tagging and lemmatization	CSLT, ESAT	Protocol specifying the tagset and its application. Protocol. POS tagger + documentation	Planned: 31-03-2007 Real: 30-11-2007
B4	Quality checking	CLST, ESAT		
C1	IPR	HLT agency	IPR regulation	
C2	Validation and Evaluation	ELDA	Written Report	
C3	Dissemination of results	CLST, ESAT, TH	Website, conference papers	
D	Project Management	CLST, ESAT, TH	Progress reports and final report	

In total 111 h and 40 m of speech were collected divided as follows.

In The Netherlands:

- native children between 7 and 11 (15h 10m)
- native children between 12 and 16 (10h 59m)
- non-native adults (15h 01m)
- non-native children between 7 and 14 (12h 34m)
- native adults above 65 (16h 22m)

In Flanders:

- native children between 7 and 11 (7h 50m)
- native children between 12 and 16 (8h 01m)
- non-native adults (8h 02m)
- non-native children between 7 and 14 (9h 15m)
- native adults above 65 (8h 26m)

About 50% of the material is read speech and 50% extemporaneous speech recorded in the human-machine interaction modality (HMI). For some groups the distribution is slightly skewed.

1.4. Problems and solutions

Speaker recruiting

Since the JASMIN-CGN corpus was collected for the aim of facilitating the development of speech-based applications for children, non-natives and elderly people, special attention was paid to selecting and recruiting speakers belonging to the group of potential users of such applications. However, speaker recruiting turned out to be much more problematic than envisaged. This applied across the board for all speaker groups.

The most efficient way to make recordings of children was to approach them through schools. However, recruiting speakers in schools was difficult because schools are reluctant to participate in individual projects owing to a general lack of time. In fact this was anticipated and the original plan was to recruit children through pedagogical research institutes that have regular access to schools for various experiments. Unfortunately, this form of mediation turned out not to work because pedagogical institutes give priority to their own projects. So, eventually, we had to contact the schools ourselves. For these reasons, recruiting children turned out to be much more time-consuming than we had envisaged

In the case of non-native speakers the applications we had in mind were especially language learning applications because there is considerable demand for CALL (Computer Assisted Language Learning) products that can help making Dutch as a second language (L2) education more efficient. In selecting non-native speakers, mother tongue constituted an important variable because certain mother tongue groups are more represented than others in the Netherlands and Flanders. For instance, in Flanders the choice was made to record mainly Francophone speakers since they form a significant fraction of the population in Flemish schools, especially (but not exclusively) in major cities and are therefore a very relevant group from the point of view of Dutch as a second language. So the difficulties encountered in recruiting non-native speakers in this case were similar to those encountered in recruiting children in schools, except that the choice of schools was more restricted and that the "yield" of a school was more difficult to predict.

In the Netherlands, the situation was different. The original idea to select speakers with Turkish and Moroccan Arabic as their mother tongue, which seemed preferable because Turks and Moroccans constitute two of the four most substantial minority groups (Dagevos, Gijsberts and Van Praag, 2003). However, this turned out not to be feasible. As a matter of fact, it was too difficult and time-consuming to recruit exclusively Turkish and Moroccan speakers. In addition, a new immigration law that imposes new obligations with respect to learning Dutch for people from outside the EU, led to considerable changes which clearly had an impact on the whole Dutch L2 education landscape. Moreover, in this new context it was no longer so straightforward to imagine which mother tongue groups would be the most obvious candidates for using CALL and speech-based applications. After various consultations with experts in the field, we decided not to limit the selection of non-natives to Turkish and Moroccan speakers and opted for a miscellaneous group that more realistically reflects the situation in Dutch L2 classes.

Considerable problems were encountered by TalkingHome in recruiting elderly speakers in the Netherlands and Flanders. The logistic difficulties in Flanders became apparent in November 2006, so that it was decided that ESAT would organise recruiting and make those recordings. The difficulties with finding elderly speakers in the Netherlands persisted to the very end of the project. In fact, a substantial group of speakers was actually recorded in the prolongation period, in July 2007, with obvious consequences for the subsequent processing of the data.

Unexpected changes in the project team

A substantial part of the project team consisted of researchers that were hired on a temporary basis. Unfortunately, several of them (three at TalkingHome, three at RU and two at ESAT) left the project before their contracts expired because they could get other jobs elsewhere. Therefore we had to look for substituting personnel, which of course caused unforeseen delay.

1.5. Recommendations for future research

A first recommendation concerns the reduction by 5% in the project budget that was required by the Programme Bureau. This was implemented by applying a reduction of the amount of speech material by 5% (from a total of 100 hours of recorded speech to 95) across the board. In hindsight, it would have been better to make specific choices and eliminate one type of material, for instance the HMI dialogues or read speech, for one group of speakers because this would have entailed a real reduction in effort, because it would not have been necessary to design dialogues or select texts for this specific group.

With respect to other problems that we encountered we have clear ideas about the causes, but it is difficult to say how these could have been prevented.

The difficulties in contacting schools and having them participate in the project had somewhat been anticipated and we had thought of reasonable solutions, but unfortunately these turned out not to work. With respect to recruiting the non-native speakers, we can say that in the Netherlands we initially based the selection on criteria inspired by the user groups and we underestimated the logistic problems involved in recruiting speakers from such a specific group. Basically, if you decide to select only Turkish and Moroccan speakers it means that you have to contact schools and teachers and that you may end up finding only one or two suitable subjects in a group of twenty. This is of course less efficient than making recordings of all twenty subjects.

With respect to the unexpected departure of researchers from the team we do not have clear recommendations, as this seems to be a more general problem that particularly applies to data collection projects.

1.6. Dissemination of results

- o **Publications (specify the type - refereed journal, proceedings, workshop,**
 - o Cucchiarini, C, H. Van hamme, O. van Herwijnen, and F. Smits (2006) JASMINCGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality, Proceedings LREC2006, Genoa, Italy.
 - o JASMIN-CGN: een uitbreiding van het Corpus Gesproken Nederlands, Jaarboek Dixit.
 - o Cucchiarini, C, J. Driesen, H. Van hamme, and E. Sanders (2008) Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus, Proceedings LREC2008, Marrakesh, Morocco.

- o **Presentations (specify the type, ...)**
 - o Taal in Bedrijf op 22-11-2005
 - o LREC2006, 22-28 May 2006, Genoa, Italië.
 - o STEVIN-Programmaday, 11 September 2006, Antwerp.
 - o Themaday STEVIN: de gebruiker centraal, 30 November 2006.
 - o Dag van de Fonetiek, 21 December 2006, Utrecht.
 - o Symposium OSTT, 15 December 2006, St. Maartenskliniek, Nijmegen.

- o **Outreach activities**

- o **other**

1.7. Exploitation of results

- o **New project proposals**
 - o STEVIN DISCO project

- o **New partnerships**
 - o There is much interest from publishers of (Dutch L2) learning material in using HLT and the JASMIN data can be used to improve this technology

- o **Patent, spin-off, other agreement with industry**

- o

2. External validation

2.1. Name organisation that performed the external validation

BAS

2.2. Delivery date external validation report

Not yet known

2.3. Summary conclusions external validation report

Project name	Identification and lexical Representation of Multiword Expressions (IRME)
Project number	STE-04-019
Planned starting date project	1 May 2005
Real starting date project	1 June 2005
Planned end of project date	31 May 2007
Real end of project date	31 August 2007

Consortium partners

- UIL-OTS, Utrecht
- Alfa-Informatica, Groningen
- Van Dale, Utrecht

Names of participating researchers per partner

Unless indicated otherwise, participant were involved during the whole project.

- UIL-OTS
 - Jan Odijk, project leader
 - Nicole Grégoire, AiO,
 - Michiel Hildebrand, from August 15, 2005 till October 1, 2005, as a programmer, to work on reviving the Rosetta3 system
 - André Schenk, as a consultant (12 person days), from November 2005 till February 2006. André Schenk has been hired for a limited amount of time as an expert in the area of Multiword Expressions in general and their treatment in Rosetta in particular, with the purpose of introducing Nicole Grégoire into these matters at an accelerated pace.
 - GridLine, has been hired to make a study of the complexity and required effort and lead time to revive the Rosetta3 system.
- Alfa-Informatica
 - Gertjan van Noord, supervisor
 - Gosse Bouma, supervisor
 - Begoña Villada Moirón, post-doc
- Van Dale
 - Johan Zuidema (did not participate in the extension period of the project)

1. Final report

1.1. Summary of the project

The IRME project addressed the following central problems: (i) the lack of large and rich formalized lexicons for multi-word expressions for use in NLP; (ii) the lack of proper methods and tools to extend the lexicon of an NLP-system for multi-word expressions given a text corpus in a maximally automated manner. Therefore, the project aimed to develop innovative methods and tools for the automatic identification and lexical representation of multi-word expressions. Concomitantly, a 5000 entry corpus-based multi-word expression lexical database

for Dutch has been developed. The database has been externally validated, and its usability has been evaluated in full in one NLP-systems for Dutch, and in part in a second NLP system for Dutch.

The project has contributed to the development of electronic lexicons, in particular for Dutch. The MWE database developed fills a gap in existing lexical resources for Dutch. The project has carried out strategic research into generic methods and tools for MWE identification and lexical representation, focusing on Dutch, but these tools are largely language-independent and can also be used for other languages, new domains, and beyond this project. In this way the project has contributed directly to strengthening the digital infrastructure for Dutch.

1.2. Overview deliverables (+time of delivery: planned and real)

Del.	Target Date	Description	Responsible	Actual Date
D1.1	T3= 31-Aug-2005	Specification of MaxEntropy and LSA/SVM models	Villada Moirón	8-Dec--2005
D2.1	T3= 31-Aug-2005	Report on formalizing and elaborating Parameterized ECM for Dutch	Linguist Utrecht	15-Dec-2006
D3.1	T8= 31-Jan-2006	Running version of Rosetta3 system	Programmer Utrecht	See below
D3.2	Not originally planned	Report on the effort and lead time estimations for making a running version of Rosetta3	GridLine	31 May 2007
D4.1	T12 = 31-May-2006	ECM Incorporation Tools	Programmer Utrecht	See below
D1.2	T6 = 30-Nov-2005	Report on performance of MaxEntropy models on acquisition of support verb constructions	Villada Moirón	Merged with D1.4 16-Feb-2007
D1.3	T9= 28-Feb-2006	Report on performance of LSA/SVM models on acquisition of phrasal verbs	Villada Moirón	26-Oct-2006
D6.1	T15= 31-Aug-2006	Report on results of incorporating idiomatic expressions into Rosetta	Linguist Utrecht	29-Aug-2007
D1.4	T12 = 31-May-2006	Report on model comparison (MaxEntropy and LSA/SVM) and evaluation on other MWEs types	Villada Moirón	Merged with D1.2 16-Feb-2007
D2.2	T12 = 31-May-2006	Report on extending the ECM to semi-idioms	Linguist Utrecht	Merged with D2.3 9-Mar-2007
D1.5	T18 = 30-Nov-2006	Specifications of tools to acquire valence patterns \ morpho-syntactic restrictions	Villada Moirón	31-Jan-2007
D2.3	T18 = 30-Nov-2006	Report on extending the ECM to support verb constructions	Linguist Utrecht	Merged with D2.2. 9-Mar-2007
D1.6	T20 = 31-Jan-2007	Delivery of automatically acquired data	Villada Moirón	28-Feb-2007
D1.7	T21 = 28-Feb-2007	Report on integration of acquired lexical knowledge into ECM-based database	Linguist Utrecht (responsible changed)	9-Aug-2007

D2.4	T22 = 30-Mar-2007	Initial version of MWE Lexical database, with associated documentation	Linguist Utrecht	20-Jun-2007
D2.5	Not originally planned	"The PEC method applied to Rosetta"	André Schenk	3-Feb-2006
D5.1a	T24= 31-May-2007	Tools to integrate MWEs into Alpino	Linguist Utrecht	31-Aug-2007
D5.1b	T24= 31-May-2007	Report on Evaluation of integrating ECM Lexical database in Alpino system	Villada Moirón	30-Nov-2007
D6.2	T24= 31-May-2007	Report on Evaluation of integrating ECM lexical database in Rosetta	Linguist Utrecht	See below
D7.1	T24= 31-May-2007	Final version of MWE Lexical database, with associated documentation, validated	Linguist Utrecht	30-Oct-2007
D7.2	Not originally planned	DuELME GUI	Linguist Utrecht	31-Aug-2007
D8.1	T4 = 30-Sep-2005	Delivery of Van Dale MWE database	Van Dale	15-Jan-2006
D9.1	T24= 31-May-2007	Report on evaluation by Van Dale	Van Dale	See below

1.3. Changes in content of deliverables and motivation for those changes

An extension of the project with 3 months has been requested and approved, so that the new end date of the project is August 31, 2007 (instead of the original end date of May 31, 2007).

Deliverables 1.2 and 1.4 have been merged into a single deliverable because they had large parts in common. In deliverable 1.3 and 1.4, the machine learning technique investigated is "decision trees" instead of the LSA/SVM models originally planned. There were two reasons for this change. First, research published on application of LSA to identify MWEs on other languages have shown poor results. These result became available while this project was already running. Second, the interpretation of the results produced by the decision trees is easy and more intuitive, which benefits the empirical process of selecting learning features and testing and improving the identification model.

Deliverables 2.2 en 2.3 have been merged into one deliverable *Report on extending the ECM to subclasses of MWEs*. The reason is that the standard representation of the various MWE classes shows considerable overlap, which makes it better for presentation purposes to describe them in a single document.

Deliverable D3.1, has, despite various attempts, not been realised as originally planned. As was reported already in various half yearly reports, at first a delay in the realization of this deliverable arose because the employee appointed for it decided to work elsewhere already soon after he started. Even though his employee worked on this deliverable only briefly, it had also become clear that the problem was considerably more complex than originally expected. We did not succeed in finding a replacement in the short term. Rather late in the project, it has been decided to assign the company GridLine the task to do an exploration of the problem and its complexity and to make an estimate of the required effort and lead time to realize this deliverable. GridLine has carried out this task but has started working on it much later than was agreed originally. Gridline's conclusions have been written down in a report that was available only at the

end of the project. This report confirms that the complexity of the problem, and the required effort and lead time to realize the deliverable were seriously underestimated. Apart from the report, GridLine has produced a DVD with a partially reworked version of certain Rosetta modules, and a series of files that contain the results of a system analysis of the (Dutch part) of the Rosetta system.

The report delivered by Gridline has been delivered and is available on the WIKI as Deliverable D3.2..

As a consequence, also some other deliverables (D4.1 and D6.1) could be realised only partially, and D6.2 could not be realized at all.

The work for deliverable D4.1 has remained limited to the development of tools for incorporation of an MWE-database in Alpino. Two versions have been developed. The first version operates on an earlier version of the MWE-database design, and does not follow specifically the ECM method. The second version operates on the final MWE-database design and follows the ECM method strictly. The latter version has been delivered and has been combined with Deliverable D5.1a.

Deliverable 5.1 has been split up into two parts (5.1a and 5.1b), with different creators/authors. Part (a) evaluates the automatic conversion from the ECM format into the Alpino lexicon format. Part (b) assesses the effect of incorporating the ECM database into the Alpino parsing system. The data used to carry out the evaluation with Alpino have also been made available on the WIKI.

For deliverable D6.1, an extensive study has been made how Rosetta deals with MWEs, and what would be needed for incorporation of MWEs from the MWE database into Rosetta. This has been laid down in a report.

Though an informal positive evaluation of the results of the extraction of multiword expressions and their contexts and properties was received via e-mail by Van Dale, Deliverable D9.1 has not been realized in the form of a report.

A few deliverables have been made though they were not originally planned. This concerns D 3.2, discussed above, deliverable D2.5, a report titled "The PEC method applied to Rosetta" by André Schenk, which describes in detail how the Parameterized Equivalence Class method can be applied to Rosetta, and Deliverable 7.2, a Graphical User Interface to the MWE database.

All other deliverables are available and have been realised as planned.

Internal Evaluation

The accuracy in extracting the minimum required lexemes, valency pattern and morpho-syntactic constraints, has indeed been evaluated against a small existing internal database, while the accuracy and coverage of the models has been evaluated against data provided by van Dale, in both cases according to the original plan. In addition, the database *Referentie Bestand Nederlands* has also been used to evaluate the accuracy and coverage of the identification models.

The lexical representation of MWEs, in particular the ECM, has been evaluated by testing whether it can be successfully used for the purpose it was designed for: semi-automatic incorporation of lexical representations into NLP systems. It has been tested in theory and practice (as planned) by applying the incorporation method to the Alpino system, and in theory only by applying the method to the Rosetta machine translation system. A test in practice could not be carried out here for reasons indicated above.

Comparison with Success Criteria

The original application lists the following success criteria:

- new insights into the (semi-)automatic acquisition and lexical representation of MWEs and their properties for NLP-purposes have to been obtained and laid down in 4 publications
- the project delivers a high-quality 5000 entry MWE lexical database for Dutch, independently validated, which is as neutral as possible with regard to grammatical framework, theory or specific implementation
- this database has been successfully tested in at least one NLP-system for Dutch.

These success criteria have clearly been met.

Additional higher aims were listed as well:

- obtain not only insights but actually concrete methods and supporting tools to (semi-) automatically acquire and lexically represent MWEs.
 - These tools have been created and have been applied.
- reduce the manual part of these methods as much as possible.
 - This has been achieved (inter alia) by introducing formalized patterns in the based on CGN-like dependency structures and by maximizing parameterization
- test the database to yield successful results for both NLP-systems mentioned
 - This been tested only on paper but not in practice
- obtain a maximally neutral lexical representation of MWEs based on the ECM or a modification thereof.
 - This has indeed been achieved

So these higher aims have also been achieved with the exception of the third bullet.

1.4. Problems and solutions

For a variety of reasons (in particular later start of certain participants, technical problems, unnatural split up of deliverables) many deliverables were available later (sometimes significantly later) than originally planned. This has led to delays in other deliverables that were dependent on them and required an extension of the project duration. However, all relevant deliverables have been created and made available.

The problem of revitalizing the Rosetta system, which has not been solved, has been mentioned before.

1.5. Recommendations for future research

Be careful with making revitalization of old software an important ingredient of the research plan. The risk that the revitalization will not succeed is high.

1.6. Dissemination of results

International Publications

(a=journal; b= book chapter;
c=conference proceedings; w= workshop proceedings)

Grégoire, Nicole, 2006, 'Elaborating the Parameterized Equivalence Class method for Dutch', in Calzolari (et al.), *LREC2006: 5th International Conference on Language Resources and Evaluation: Proceedings*, pages 1894-1899, Genoa, Italy. [c]

Grégoire, N., 2007, 'Design and Implementation of a Lexicon of Dutch Multiword Expressions', in (Grégoire et al., 2007), pp. 17-24. [w]

Grégoire, N., S. Evert and S.N. Kim (eds.), 2007. 'Proceedings of the Workshop on A Broader Perspective on Multiword Expressions', ACL 2007, Prague, June 28, 2007. [w]

Van de Cruys, T. and B. Villada Moirón, 2007, 'Semantics-based Multiword Expression Extraction'. In (Grégoire et al., 2007), pp. 25-32. [w]

Van de Cruys, T. and B. Villada Moirón, 2007, 'Lexico-Semantic Multiword Expression Extraction'. In P. Dirix *et al.* (eds.), *Computational Linguistics in the Netherlands 2006*, pp. 175-190. [c]

Villada Moirón, Begoña, 2005. 'Linguistically enriched corpora for establishing variation in support verb constructions'. In *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)*. R. of Korea. [w]

Villada Moirón, Begoña and Jörg Tiedemann, 2006. Identifying idiomatic expressions using automatic word-alignment. Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context, p.33-40. Trento, Italy. [w]

Villada Moirón, B., A. Villavicencio, D. McCarthy, S. Evert and S. Stevenson (eds.) (2006). Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties. Sydney, Australia. [w]

National Publications

Odijk, J., 2006, 'IRME', DIXIT 4.2, p. 23, December 2006.

Workshops

Villada Moiron, B., A. Villavicencio, D. McCarthy, S. Evert and S. Stevenson, 'Multiword Expressions: Identifying and Exploiting Underlying Properties', Workshop held at COLING/ACL 2006 in Sydney, Australia.

Grégoire, N., S. Evert and S.N. Kim, 2007. 'A Broader Perspective on Multiword Expressions', workshop held at ACL 2007, Prague, June 28, 2007.

Grégoire, Nicole, Stefan Evert and Brigitte Krenn, 2007 'Towards a shared task for MWEs (MWE2008)' workshop proposal accepted for LREC 2008, Marrakech, Morocco.

Presentations

Grégoire, Nicole, 2005, 'Lexical representation of Multiword Expressions', Presentation CLIN, UvA, Amsterdam, 16 December 2005

Grégoire, Nicole, 2006, 'Lexical Representation of Multiword Expressions', CTL Colloquium, 20 January 2006. UiL-OTS, Utrecht.

Grégoire, Nicole, 2006, 'Elaborating the Parameterized Equivalence Class method for Dutch', poster presentation at LREC 2006, Genoa, Italy, May, 2006d.

Grégoire, N. 2007, 'Converting a Standard Representation of MWEs to the Alpino Lexicon' presentation held at CLIN 17, Leuven, 12 January 2007.

Grégoire, N., 2007, 'Design and Implementation of a Lexicon of Dutch Multiword Expressions', presentation held at the ACL 2007 workshop 'A Broader Perspective on Multiword Expressions', Prague, Czech Republic, 28 June 2007.

Grégoire, N., 2007, 'DuELME: Dutch Electronic Lexicon of Multiword Expressions', presentation held at the TST-Centrale workshop 'De Gebruiker Centraal', Antwerpen, 23 November 2007.

Grégoire, Nicole and Begoña Villada Moirón, 2005: 'Het STEVIN-IRME Project', Poster presentation at 'Taal in Bedrijf 2005', Eindhoven, 22 November 2005.

Grégoire, Nicole and Begoña Villada Moirón, 2006: 'IRME: Identification and Representation of Multiword Expressions', Presentation at the STEVIN programmadag, Antwerpen, 11 September 2006.

Odijk, J., 2007, 'The IRME Project: Lexical Representation of MWEs', poster held at the STEVIN-dag, Hoeven, 21 September 2007.

Van de Cruys, T. and B. Villada Moiron, 2007, 'Semantic extraction of multiword expressions'. presentation held at CLIN 17, Leuven (12 January 2007).

Van de Cruys, T. and B. Villada Moirón, 2007, 'Semantics-based Multiword Expression Extraction', presentation held at the ACL 2007 workshop 'A Broader Perspective on Multiword Expressions', Prague, Czech Republic, 28 June 2007.

Van den Heuvel, Theo, 2006, 'Talige ondersteuning: meer dan woorden (IRME)', TST-centrale workshop 'STEVIN: De Gebruiker Centraal', Rotterdam, 30 November 2006.

Villada Moirón, B. Establishing relevant features for identification of multiword expressions. CLS, Nijmegen, March 9, 2006.

Villada Moirón, B., 2006, 'Capturing idiosyncratic linguistic behavior for automatic Multiword Expression Identification', talk given at the Computational Linguistics Colloquium at COLI, Saarbruecken University, Germany. 23 November, 2006

Villada Moirón, B., 2007, 'Identification and Representation of Multiword expressions (IRME)', poster held at the STEVIN-dag, Hoeven, 21 September 2007.

Villada Moirón, Begoña and Nicole Grégoire. 'Quantifying and qualifying lexicalized and idiomatic expressions'. *Collocations and Idioms 1: The First Nordic Conference on Syntactic Freezes*. Joensuu, Finland, May 19-20, 2006

Villada Moirón, B. and J. Tiedemann. Identifying idiomatic expressions using automatic word-alignment. EACL workshop on 'MWEs in a multilingual context'. Trento, Italy. April 3, 2006.

Submissions

Grégoire, N., 2007, 'DuELME: A Dutch Electronic Lexicon of Multiword Expressions', submitted to the *Journal of Language Resources and Evaluation*, special issue on Multiword Expressions. [a]

Grégoire, N., (2007), 'The Selection and Representation of Support Verb Constructions', paper submitted to *LREC 2008*, Marrakech, Morocco. [c]

Other

Website From the beginning of the project a Website has been set up for the IRME project, with both a public section and an internal section for the project participants. URL: <http://www-uilots.let.uu.nl/irme/>

Villada Moirón, B. and N. Gregoire: Quantifying and qualifying lexicalized and idiomatic expressions. Utrecht/Groningen, 2006.

N. Grégoire, MWE lexicon for Dutch - Representation protocol, Utrecht, 2006. (revised version)

N. Grégoire, Installatiehandleiding Windows voor de MWE database. Utrecht, 2006

1.7. Exploitation of results

The results of the project will of course be made available via the TST-Centrale and thus become available to the whole language and speech technology community of the Netherlands and Flanders and beyond.

Van Dale has shown great enthusiasm concerning the results of the acquisition of detailed morpho-syntactic properties of MWEs from large contemporary text corpora, has received the results and will make use of these results.

The research by Nicole Grégoire is going to be continued beyond the IRME project in a PhD project at Utrecht University (UIL-OTS).

Interest from an unexpected corner, for the MWE database, viz. the speech technology community, has already been shown, so that the database may also prove its usefulness in this area.

Options to link the MWE database and the Cornetto database are being investigated by the TST-centrale..

Gertjan van Noord has announced that he will integrate the IRME database into the Alpino parser for Dutch.

2. External validation

The resulting MWE lexical database has been successfully validated by an external organisation, CST (Copenhagen). This validation has resulted in an updated MWE lexical database that is available and will soon be transferred to the TST-centrale.

2.1. Name organisation that performed the external validation

CST, Copenhagen

2.2. Delivery date external validation report

30 August 2007

2.3. Summary conclusions external validation report

Citations from the validation report are represented in italics and between double quotes.

" The overall impression of the linguistic annotations after the validation is that the IRME database of MWEs is a skilfully elaborated language resource of a high quality. We have some recommendations (see below) for improvement of the documentation and the resource which could

be taken into consideration for future work.

The content validation, which checked the correctness and consistency of the linguistic annotations, identified few errors considering the complexity of the database

<i>MWE patterns</i>	<i>21 errors</i>
<i>MWE descriptions</i>	<i>93 errors</i>

The formal validation, which checked the technical quality and integrity of the file archive and the organisation of the files, revealed that all the validation criteria were met."

The errors reported (the largest part of which were minor and had no consequences for the linguistic correctness, according to the validation report) have been corrected in an updated version that has been made available.

"The validation of the documentation, which checked three specific documentation files for availability of specific information, completeness of information, and adequacy for future users of the corpus, identified a number of areas where the criteria were not met or inadequately met."

These have also been corrected in the final version.

Proposals funded in the 2nd Call for Proposals for strategic research proposals and HLT resources (data & tools)

<i>acronym</i>	<i>coordinating institute and other academic partners</i>	<i>industrial partners</i>	<i>STEVIN priorities addressed</i> <i>(subject)</i>	<i>planned duration</i>	<i>funding</i>
DAESO STE05024	Tilburg University (Emiel Kraher) Antwerpen University Universiteit van Amsterdam	Textkernel	Language research Language resources (Semantic / discourse annotation)	36 mnths	€ 487.000
DPC STE05026	KU Leuven (Piet Desmet) Hogeschool Gent		Language resources (Multilingual corpora / translational equivalents)	34 mnths	€ 498.000
LASSY STE05020	Groningen University (Gertjan van Noord) KU Leuven		Language resources (Syntactic treebank)	36 mnths	€ 496.000
MIDAS STE05030	KU Leuven (Hugo Van hamme) Radboud Univ. Nijmegen	Nuance	Speech research (Robust ASR)	48 mnths	€ 499.000
NBest STE06012	TNO-TM (David van Leeuwen) KU Leuven, Twente University, Radboud Univ. Nijmegen, Ghent University, SPEX, TU Delft		Speech resources (ASR benchmarks for evaluation)	29 mnths	€ 470.000
STEVIN can PRAAT STE05035	Universiteit van Amsterdam (Paul Boersma) Leiden University SPEX	Speech-Minded	Speech resources (ASR, annotation tool)	24 mnths	€ 114.000

Project name DAESO (Detecting And Exploiting Semantic Overlap)

Project number STE05024

Reporting period October 1, 2006 - April 1, 2008 (months 1-18)

Participants

prof. dr. E. Krahmer (UvT)
prof. dr. W. Daelemans (UA)
prof. dr. M. de Rijke (UvA)
drs. J. Zavrel (Textkernel)

Empolyees

dr. E. Marsi, postdoc UvT
dr. I. Hendrickx, postdoc UA
NN, postdoc UvA
researcher, Textkernel
6 student-assistants UvT

1. Summary of the project

The well-known fact that similar information can be expressed in many different ways is one of the major challenges in building robust NLP applications. It is commonly assumed that such applications can be improved with knowledge of how natural language expressions relate to each other, for instance in terms of paraphrases (same semantic content, different wording) or entailments (one expression implied by the other). DAESO investigates the detection of semantic overlap between Dutch sentences and the exploitation of this knowledge in existing NLP applications. For this purpose, tools will be developed for the automatic alignment and classification of semantic relations (between words, phrases and sentences) for Dutch, as well as for a Dutch text-to-text generation application which fuses related Dutch sentences into a single grammatical sentence, which may be a generalization, a specification or a reformulation of the input sentences. To facilitate development and testing of these tools, an annotated monolingual Dutch parallel/comparable corpus of 1M words will be developed, consisting of pairs of texts that express comparable information. The utility of the resources and tools will be demonstrated in the context of three applications: (1) question-answering systems (improved recall, more complete answers), (2) information extraction (improved recall), and (3) summarization (beyond extraction: sentence compression, sentence fusion, anaphora resolution).

1.1. Overview deliverables (+time of deliverable) according to the proposal

In the first half of the project, we primarily concentrated on developing the DAESO corpus; a parallel monolingual treebank for Dutch required for detecting semantic overlap (WP1A and WP1B). In addition, we started working on automatic alignment and paraphrasing (WP2A and WP2B) and on applying the techniques in the context of multidocument summarization (WP3A and WP3C).

The excel file in the appendix gives a detailed overview of the relevant deliverables for the first 18 months of the project, with full IPR details. We are pleased to state that the IPR with our data providers is now fully settled.

1.2. Previously completed deliverables

See the excel sheet in the appendix for an overview of the deliverables of the first 18 months.

1.3. Changes requested (contents/timing of deliverables) and motivation

Most of the work is going as planned. However, soon after the project started it became clear that developing the corpus was going to take more time than planned originally, in part due to problems with the data collection and in part to an overestimation of availability of student-assistants (we describe these in more detail in section 2.2 below). This means that the final version of the annotation of the corpus (WP1B-1) will be finished later than anticipated. The prognosis is that the annotation will be finished by October 2008, and we are currently hiring two additional student-assistants to further speed up the annotation. Some other WPs cannot be finalized without the corpus data, most notably WP2B. However, preparatory work for this WP has been done, and when WP1B is finished, WP2B can be completed soon afterwards as well. Since we have also been working on WPs that were planned for later (including WP3A and WP3C) we do not expect that this will cause serious problems for the project as a whole.

1.4. Employee involvement in relation to the original plan

Erwin Marsi (Tilburg University) has been working on DAESO since the beginning. Iris Hendrickx (Antwerp University) joined the project on November 1, 2007 (somewhat earlier than originally planned).

1.5. Dissemination of the results

[Nota bene: All papers, presentations, etc. mentioned here can be downloaded from <http://daeso.uvt.nl/>]

1.5.1. Publications,

Emiel Kraemer, Erwin Marsi and Paul van Pelt (2008), Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion. Accepted for The *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, USA, June 15-20, 2008.

Erwin Marsi and Emiel Kraemer (2008), Detecting semantic overlap: Announcing a Parallel Monolingual Treebank for Dutch, submitted to the *Proceedings of the 18th meeting of Computational Linguistics in the Netherlands (CLIN)*.

Iris Hendrickx, Walter Daelemans, Kim Luyckx, Roser Morante and Vincent Van Asch (2008), CNTS: Memory-Based Learning of Generating Repeated References. Accepted for the *Referring Expression Generation Challenge 2008*, held in conjunction with the *5th International Natural Language Generation Conference (INLG 2008)*, Salt Fork, Ohio, USA, June 12-14, 2008.

Mariët Theune, Jette Viethen, Iris Hendrickx and Emiel Kraemer (2008), GRAPH: Realizing the Costs. Accepted for the *Referring Expression Generation Challenge 2008*, held in conjunction with the *5th International Natural Language Generation Conference (INLG 2008)*, Salt Fork, Ohio, USA, June 12-14, 2008.

Emiel Kraemer, Mariët Theune, Jette Viethen and Iris Hendrickx (2008), The Costs of Redundancy in Referring Expressions (GRAPH). Accepted for the *Referring Expression Generation Challenge 2008*, held in conjunction with the *5th International Natural Language Generation Conference (INLG 2008)*, Salt Fork, Ohio, USA, June 12-14, 2008.

Erwin Marsi and Emiel Kraemer (2007), Annotating a parallel monolingual treebank with semantic similarity relations. In: *The Sixth International Workshop on Treebanks and Linguistic Theories (TLT'07)*, Bergen, Norway, December 7-8, 2007.

Erwin Marsi, Emiel Kraemer and Wauter Bosma (2007). Dependency-based paraphrasing for recognizing textual entailment. In: *Proceedings of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007.

1.5.2. Presentations

Emiel Kraemer and Erwin Marsi, *Further Explorations in Sentence Fusion*, Talk at Symposium on Natural Language Processing and Multimodal Interaction, Twente University, March 27, 2008.

Erwin Marsi and Emiel Kraemer, *Detecting Semantic Overlap: Annotating a parallel monolingual treebank with semantic similarity relations*. Poster presented at the 2nd STEVIN Program Meeting, Hoeven, The Netherlands, September 21 2007.

Erwin Marsi, Emiel Kraemer. *Detecting semantic overlap: Announcing a Parallel Monolingual Treebank for Dutch*. Presented at the 18th Computational Linguistics in the Netherlands (CLIN2007), Nijmegen, The Netherlands, December 7, 2007.

Erwin Marsi, Emiel Kraemer. *Annotating a Parallel Monolingual Treebank with Semantic Similarity Relations*. Presented at The Sixth International Workshop on Treebanks and Linguistic Theories (TLT'07), Bergen, Norway, December 7-8, 2007

Erwin Marsi, Emiel Kraemer. *Shallow approaches to sentence alignment in comparable text*. Presented at IDI DIS Meeting, NTNU, Trondheim, Norway, November 22, 2007.

Emiel Krahmer, Erwin Marsi, Paul van Pelt. *Question-Driven Sentence Fusion is a Well-Defined Task. But the Real Issue is: Does it matter?*. Presented during Stevin site visit Meeting, UvT, Tilburg, November 8, 2007.

Emiel Krahmer, Erwin Marsi. *Detecting And Exploiting Semantic Overlap*. Presented during Stevin site visit Meeting, UvT, Tilburg, November 8, 2007.

Erwin Marsi, Emiel Krahmer. *DAESO: Detecting And Exploiting Semantic Overlap*. Presented at ILK Meeting, UvT, Tilburg, November 7, 2007.

Emiel Krahmer, Erwin Marsi. *Detecting And Exploiting Semantic Overlap (DAESO)*. Presented during Stevin day, Antwerp, 2006.

1.5.3. Outreach activities

Emiel Krahmer, Erwin Marsi en Lilian Beijer (2008), *Taaltechnologie voor mensen met communicatieve beperkingen: een optie?* In: *DIXIT, Tijdschrift over Toegepaste Taal- en Spraaktechnologie* (Dutch magazine on applied language and speech technology), to appear.

Emiel Krahmer, Erwin Marsi, Walter Daelemans, Maarten de Rijke, Jakub Zavrel (2006), *Hetzelfde, maar dan anders: Semantische overlap detecteren en gebruiken (DAESO)*. In: *DIXIT, Tijdschrift over Toegepaste Taal- en Spraaktechnologie* (Dutch magazine on applied language and speech technology), 4(2), December 2006, p. 18.

1.5.4. Other

The DAESO website contains a lot of information about the project, and is updated regularly. See <http://daeso.uvt.nl/>

We made a set of Dutch summaries available (source texts with sentence rankings and extraction), useful for evaluation purposes. See <http://ilk.uvt.nl/~marsj/data/de-Vries-summaries.html>

The first public release of the Hitaext tool for manual alignment of text. See <http://daeso.uvt.nl/hitaext>

The first public release of the Algraeph tool for manual alignment of linguistic graphs. See <http://daeso.uvt.nl/algraeph>

1.6 Exploitation of the results

1.6.1. (New) collaborations

Apart from the collaborations that were planned initially, various new collaborations have been initiated. Together with Polderland Language and Speech Technology and Twente University (in particular with Mariet Theune and Wauter Bosma) we have been exploring ways to collaborate on automatic text summarization. With Lilian Beijer (St Maartenskliniek, Nijmegen) we have discussed potential ways in which the DAESO tools could be applied for language therapy (see our Dixit paper on this. Finally, with Antal van den Bosch (Tilburg University) we will work on applying the DAESO tools and data for Memory-based paraphrasing (potentially useful for Machine Translation). The Parallel Dutch Corpus, another Stevin project, has expressed interest in using Hitaext, our tool for manual alignment of text, to edit their sentence alignments.

1.6.2. (Accepted) project proposals based on present project

Memphix: MEMory-based paraPHrasing with Implicit and eXplicit semantics, PhD project proposal, sponsored by Tilburg University (with Antal van den Bosch and Harry Bunt). Accepted, currently hiring.

BOF: An evaluation corpus for Dutch multi-document summarization. Sponsored by Antwerp University. Accepted, currently hiring student-assistents.

Q-SUM: Query-based summarization, Stevin proposal 2007 (not accepted; we consider resubmitting it in the STW Open Competition).

1.6.3. Other (patent, ..)

None.

2. Progress per work package and deliverable

2.1. Activities completed in the past 18 months

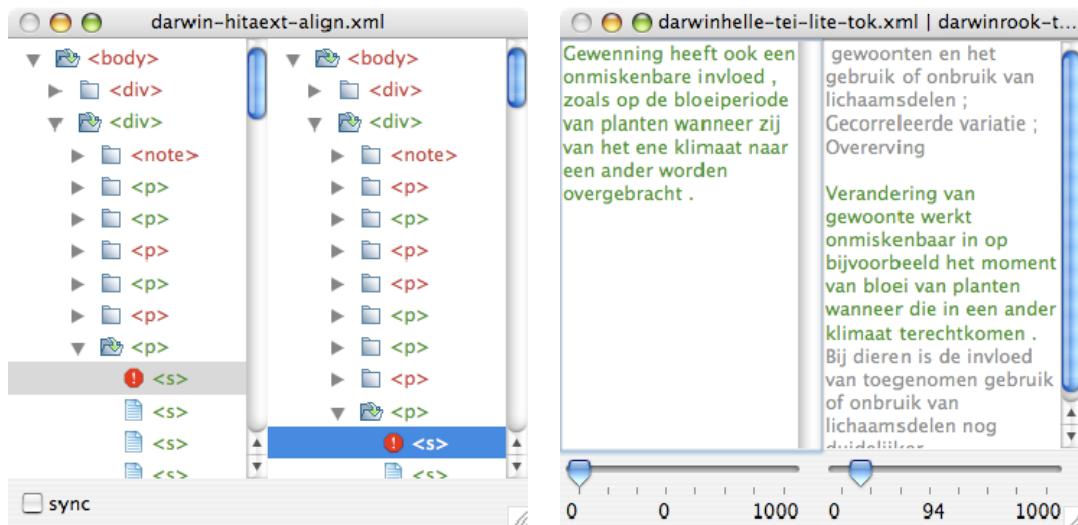
2.1.1. Work on the DAESO corpus

The entire first year of the DAESO project was devoted to building the DAESO corpus. The following table captures the corpus construction, how much of the data is manually processed (500k), and how much data is available in all (much more). The data collection forms deliverable WP1A-1

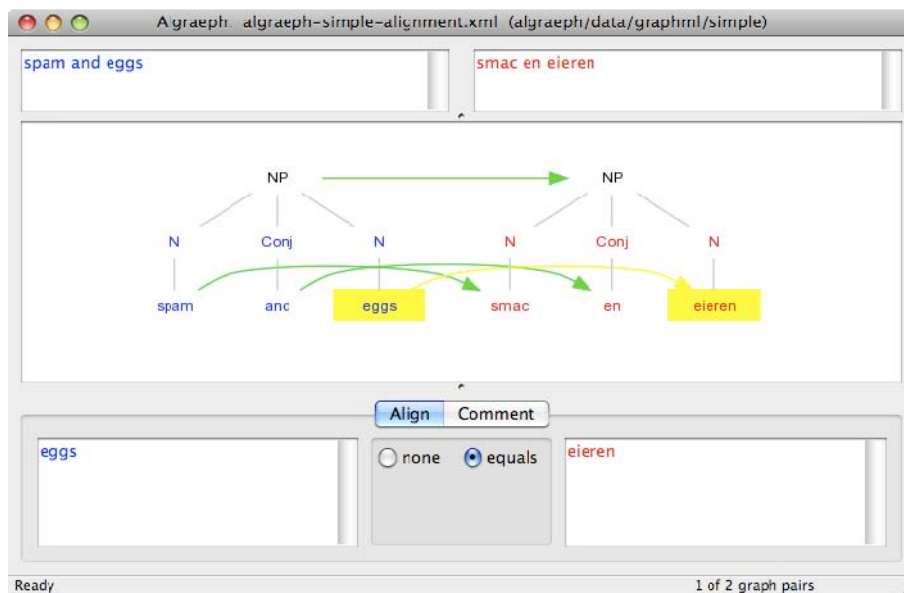
Corpus part	Source	Manually processed (words)	Available (words)
Autocue-subtitles	[TwNC]	125k	192k
Translations	Darwin 1	25k	154k
	Darwin 2	25k	191k
	Montaigne 1	25k	462k
	Montaigne 2	25k	~500k
	Saint-Exupery 1	15k	15k
	Saint-Exupery 2	15k	15k
News headlines	[web]	24k	>900k
Press releases	ANP	125k	197k
	NOVUM	125k	136k
QA system output	[imix]	1k	1k

We are pleased to announce that the IPR with all the relevant data providers (i.e., all those not between square brackets) has been formally regulated by now. With respect to the original plan, one change was made: it turned out that there was not enough suitable QA material within the IMIX corpus (different, comparable answers to the same question): 25k words were planned, but only 1k words was actually available. To compensate for this, we introduced an extra corpus part, namely comparable head lines, mined from Google news.

All the data has been put in XML format (XML TEI, Text Encoding Initiative for the translations, and custom XML formats for the other sources), and all sentences have been split and tokenized (using the D-COI tokenizer), and parsed with Alpino. In addition, all sentences have been aligned in pairs, and this alignment was manually checked. For this a specific tool (Hitaext) has been developed and released under the GPL license. This tool was not promised, and is an extra result (WP1A-3). Hitaext allows for n-to-n alignments at the sentence level in a generic way, and can be applied to both monolingual and multilingual sentence pairs.



In addition, we have worked on a tree-graph alignment tool (Algraeph, WP1A-2) and a set of annotation guidelines (WP1A-4). The Algraeph tool allows for within sentence alignment of words and phrases, and is likewise applicable to both monolingual and multilingual alignment and can be configured for use with different semantic relations (the picture below illustrates multilingual alignment, with only the “equals” relation). All of this is finished, and concludes WP1A. To round up the DAESO corpus, we still need to finish the manual alignment of relations between words and phrases in pairs of related sentences (WP1B-1). This work is currently ongoing (achieving high rates of inter-annotator agreement). The publication about the corpus has been realized (in fact, we have written one international and one national paper about it). All in all, the production of the DAESO corpus is going well, but taking more time than foreseen, and we return to this in section 2.2.



For details about the corpus collection and annotation, and the tools that were developed for this purpose, we refer to Marsi and Kraemer (2008), “Detecting semantic overlap: Announcing a Parallel Monolingual Treebank for Dutch” .

2.1.2. Work on Detecting Semantic Overlap

In the first half of the project, we have started working on a number of WPs related to detecting semantic overlap. Software was developed to automatically align parallel texts (e.g., books) and comparable texts (e.g., press releases). So far, this software has primarily been used to speed up the manual annotation. In addition, we have explored different features and distance metrics for aligning sentences from comparable texts. Up to now this was purely based on shallow features but in later stages we want to include more advanced linguistic features (annotated in the corpus) as well. This lays the foundation for the work in WP2A ("Alignment and labelling software"). Marsi presented the results so far in a talk "Shallow approaches to sentence alignment in comparable text" at NTNU, Norway. In this context, it is also worth mentioning our contribution to the third RTE ("Recognizing Textual Entailments) Challenge, which was not planned but which seemed highly relevant for our current purposes. For results, we refer to Marsi et al. (2007), "Dependency-based paraphrasing for recognizing textual entailment".

In addition, some preliminary work has been carried out to be able to extract paraphrases from the DAESO corpus (relevant for WP2B), but this work can only be finalized when the manual annotation of the corpus is done.

Finally, we performed a number of experiments on sentence fusion (relevant for WP2C), which resulted in an ACL 2008 paper ("Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion"). The data that were collected for this work (2200 human-produced sentence fusions) will be made available to the research community as well.

2.1.3. Work on Exploiting Semantic Overlap

Work in this part of the project started when Iris Hendrickx joined the project team on November 1, 2007 (note that this is somewhat earlier than initially planned, which became possible since her previous contract ended earlier than anticipated at the time of drafting the DAESO project proposal). Since the DAESO corpus was not completed when Iris joined the project, it was decided that she started working on WP3C, "Applications in Multidocument Summarization". As starting point, the open source software package MEAD (<http://www.summarization.com/mead/>) was taken. MEAD is a multi document summarizer for English and Chinese, and we adapted it to Dutch. This gave us a baseline summarizer in which we can test the usefulness of the DAESO tools. The baseline summarizer is functional and will provide the backbone for the online demo of a Dutch multi-document summarizer.

Evaluation of summaries is a difficult problem, which is further complicated by the lack of data for Dutch. To remedy this problem, a small, new project proposal was written (and subsequently accepted by the University of Antwerp), which allows us to develop an extra data-set of source documents and summaries, which will be made available for research purposes as well. The first step of this project is to make a list of 25 topics or questions and for each topic we collect a set of 5-10 relevant newspaper articles (based on the DAESO and D-COI corpora). Next, 5 student assistants are paid to write summaries for these document sets and to rank the importance of the sentences in the texts. The final evaluation corpus (which is planned to be finished by August 1, 2008) will be used to evaluate the automatic multi-document summarizer in the DAESO project.

In addition, some preliminary work has been done on WP3A ("Coreference resolution"). In particular, Hendrickx and colleagues (2008) participated in the co-reference task of the REG Challenge, with a machine learning algorithm which makes predictions about how co-referential NPs

should be realized in text (as a pronoun, a full description, etc.), which is directly relevant for WP3A. Notice that Hendrickx also contributed to two other Tasks in the REG Challenge (described in Theune et al., 2008 and Krahmer et al., 2008).

2.2. Problems and solutions

There are no serious problems, but the preparation of the corpus took more time than foreseen.

The reasons for this have been described in more detail in the report of the first six months. Here we briefly summarize them. To begin with, it turned out that some of the books to be included in the corpus were not available in electronic form. These books have been OCRed, and the scanning was manually corrected. In addition, the sentence splitting and tokenization software from the D-COI project proved to be more error prone than expected, so we decided to manually correct this as well (for the translations). To facilitate the manual annotation of related sentences we developed an extra annotation tool (Hitaext). And finally, since we could not get sufficient useful QA material, an additional source of data needed to be collected: the news headlines from Google News Dutch. These factors together implied that the data collection (WP1A) took longer than anticipated.

As a result of this, the manual annotation (WP1B) started later than originally planned. Unfortunately we overestimated the availability of student-assistants, which implied that this work package also took more time than originally anticipated. In addition, we experienced some startup problems, mainly with debugging the annotation software on all platforms (Linux, Windows and Mac OS), and fleshing out the details of the annotation guidelines. Our original planning may have been a little too optimistic. However, we feel that we are on the right track. We are currently hiring two additional student-assistants, to make sure that the final annotation will be finished before the next half-year report.

Finally, we would like to reiterate that while the delays in WP1B will lead to some delays for other WPs as well, they will not cause serious overall problems for the project, since we are already working on various work packages that are scheduled for later on. We also believe that the extra work we put in maintaining the quality of the corpus (additional data, additional manual checks of OCR and splitting/tokenization) will pay off, since they will result in a better corpus.

2.3. Proposed schedule for the upcoming period

We aim to finish WP1B-1 by October 2008. Furthermore, we request extension of the related WP2B to December 2008. Other than that, we will proceed as planned.

Appendix A: EXCEL sheet with deliverables for first 18 months

Deliverable #	Omschrijving	Datum (gepland)	Datum (uitstel)	Datum (gereed)	Type	Achtergrond (kennis)	Voorgrond(kennis)	Beschikbaarheid	IPR / licenties / prijzen / bijzonderheden
WP1A-1	XML corpus (1M woorden) + preprocessing	31-Mar-07	30-Jun-07	31-Oct-07	Data	None	A parallel monolingual treebank	Non-commercial and commercial	IPR completely settled
WP1A-2	Gadget annotatie-tool	31-Mar-07	30-Jun-07	31-Oct-07	Software	None	A graphical (tree-) graph alignment tool	Open source	GPL license
WP1A-3 [extra]	Sentence Alignment tool [Hitaext]	31-Mar-07	30-Jun-07	31-Oct-07	Software	None	A graphical sentence alignment tool	Open source	GPL license
WP1A-4 [extra]	Annotation guidelines [1st version]	31-Mar-07	30-Jun-07	31-Oct-07	Documentation	None	A first version of the annotation guidelines	Available upon request (for everyone)	
WP1B-1	Manual annotation of half the corpus	30-Sep-07	31-Mar-08		Data	None	A detailed annotation of a parallel monolingual treebank	Non-commercial and commercial	IPR completely settled
WP1B-2	Conference paper	30-Sep-07		31-Oct-07	Documentation	None	Two papers describing the DAESO corpus (one national and one international)	Free for everyone	
WP2B	Paraphrase extraction tool	31-Dec-07	30-Jun-08		Software	None	A tool that extracts paraphrases from a parallel monolingual treebank	Open source	GPL license

Project title: DPC – Dutch Parallel Corpus,
A multifunctional and Multilingual corpus
(Dutch – English, Dutch – French)

Project number: STE-05-26

Reporting period 01/08/07 – 31/01/08 (S3)

Participants – co-ordinating research institutes

K.U.Leuven Campus Kortrijk
Etienne Sabbelaan 53
B-8500 Kortrijk

Hogeschool Gent
Jozef Kluyskensstraat 2
B-9000 Gent

Research partners – Core research team

Prof. Dr. Piet Desmet (promotor)
Prof. Dr. Willy Vandeweghe (co-promotor)

Senior staff members

Dr. Hans Paulussen (start: 1-Dec-2007)
Dra. Lieve Macken

Project collaborators

Dr. Julia Trushkina (end: 19-Nov-2007)
Lic. Antoine Besnehard (end: 31-Aug-2007)
Lic. Maribel Montero Perez (start: 17-Dec-2007)
Lic. Lidia Rura

Contents

Part 1: Global Planning of the DPC project.....	3
Part 2: Report on Semester 3 (S3)	11
Part 3: Planning of Semester 4 (S4).....	16
Part 4: DPC presentations in Semester 3 (S3):	18
Appendix 1: DPC Text Providers	19
Appendix 2: Deliverables of S3	20

Introduction

This document provides a report on the third semester of the DPC project. The report consists of four parts:

Part 1 lays out a global planning of the project. It is based on the planning introduced in the STEVIN project proposal and represents a more detailed version of it. The planning is introduced in a form of a table, where the first column contains information on project work packages (WP1 to WP6) with details on project tasks and deliverables. Columns 2 to 6 stand for the five semesters of the project (S1 to S5). Crosses in table cells provide information on scheduling of specific tasks and on the deadlines for the project deliverables.

Part 2 of this document presents a detailed report for the third semester (S3) of the project. The report is based on the global planning.

Planning of Semester 4 is introduced in Part 3.

Part 4 contains a list of presentations and publications on the DPC project during Semester 3.

Part 1: Global Planning of the DPC project

GLOBAL PLANNING	S1 <i>1/5/06-31/1/07</i>	S2 <i>1/2/07-31/7/07</i>	S3 <i>1/8/07-31/1/08</i>	S4 <i>1/2/08-31/7/08</i>	S5 <i>1/8/08-1/3/09</i>
WP1: Corpus Design and Data Collection					
<i>Tasks</i>					
T1.1: Design of the Corpus					
T1.1.1: Predefine a composition of the corpus (which text types and in which proportion)	X				
T1.1.2: Prepare a protocol for corpus design	X				
T1.1.3: Prepare an initial list of potential data providers	X				
T1.2: Contacts with data providers					
T1.2.1: Establish contacts with data providers	X	X	X	X	
T1.2.2: Arrange IPR agreements	X	X	X	X	
T1.2.3: Collect data from providers	X	X	X	X	
T1.3: Initial processing of the data					
T1.3.1: Prepare protocol for data collection	X				
T1.3.2: Check the quality of the data	X	X	X	X	
T1.3.3: Store the data on the server	X	X	X	X	

T1.3.4: Document the incoming data	X	X	X	X	
T1.3.5: Compile a list of metadata for incoming data	X	X	X	X	
T1.3.6: Establish correspondences between language versions	X	X	X	X	
T1.3.7: Prepare an overview of collected data	X	X	X	X	
T1.4: Feedback from the user group					
T1.4.1: Prepare a questionnaire for the user group	X				
T1.4.2: Prompt users to fill in the questionnaire	X				
T1.4.3: Analyze the feedback	X				
T1.5 User Documentation					
T1.5.1: Write user documentation (corpus design and data collection)				X	
<i>Deliverables</i>					
D1.1: Specifications of the corpus design	X				
D1.2: IPR agreements				X	
D1.3: Protocol for data collection	X				
D1.4: User documentation - part 1: Corpus Design and Data Collection				X	
WP2: Text Normalization					
<i>Tasks</i>					
T2.1: Cleaning the incoming data					

T2.1.1: Prepare protocol for data cleaning	X	X			
T2.1.2: Convert the data to txt format	X	X	X	X	X
T2.1.3: Normalize character encoding	X	X	X	X	X
T2.1.4: Introduce presentational mark-up in the data	X	X	X	X	X
T2.2: Sentence splitting					
T2.2.1: Find/implement sentences splitters for NL, FR, EN	X				
T2.2.2 Prepare protocol for sentence splitting	X	X			
T2.2.3: Split sentences	X	X	X	X	X
T2.2.4: Manual checking of sentence splitting	X	X	X	X	X
T2.3: Tokenization					
T2.3.1: Find/implement tokenizers for NL, FR, EN		X			
T2.3.2 Prepare protocol for tokenization		X			
T2.3.3: Tokenize the data		X	X	X	X
T2.3.4: Manual checking of tokenization		X	X	X	X
T2.4: Encode monolingual data in XML					
T2.4.1: Prepare protocol for XML-coding		X			
T2.4.2: Encode the data in XML format		X	X	X	X
T2.5: User documentation					
T2.5.1: Write user documentation (text normalization)				X	
<i>Deliverables</i>					

D2.1 Protocols text normalization		X			
D2.2 XML-files in TEI/XCES-format (monolingual)					X
D2.3 User documentation – part 2: Text normalization					X
WP3: Alignment					
<i>Tasks</i>					
T3.1: Paragraph alignment					
T3.1.1: Prepare protocol for paragraph alignment	X	X			
T3.1.2: Align data on paragraph level	X	X	X	X	X
T3.2: Sentence alignment					
T3.2.1: Explore potential aligners to be used	X	X			
T3.2.2: Prepare protocol for sentence alignment	X	X			
T3.2.3: Align the data on sentence level with different tools		X	X	X	X
T3.2.4: Combine alignments of different tools		X	X	X	X
T3.2.5: Manually check the alignment quality		X	X	X	X
T3.2.6: Prepare a final version of alignment to be included in the DPC corpus		X	X	X	X
T3.3: Sub-sentence alignment					
T3.3.1: Select portions of data for sub-sentence alignment			X	X	X
T3.3.2: Prepare protocol for sub-sentential alignment			X		
T3.3.3: Manually align selected portions of data on sub-sentence level			X	X	X

T3.4: User documentation					
T3.4.1: Write user documentation (alignment process)					X
<i>Deliverables</i>					
D3.1 Protocol for sentence alignment		X			
D3.2 Release of sentence alignments					X
D3.3 Protocol for sub-sentence alignment			X		
D3.4 Release of sub-sentence alignments					X
D3.5 User documentation - part 3: Alignment					X
WP4: Linguistic Annotation					
<i>Tasks</i>					
T4.1: Lemmatization					
T4.1.1: Find/implement lemmatizers for NL, FR, EN		X			
T4.1.2 Prepare protocols for lemmatization			X		
T4.1.3: Lemmatize the data			X	X	X
T4.1.4: Manual checking of lemmatization			X	X	X
T4.2: Part of speech tagging					
T4.2.1: Find/implement PoS taggers for NL, FR, EN		X			
T4.2.2 Prepare protocols for part of speech tagging			X		
T4.2.3: Annotate the data			X	X	X
T4.2.4: Manual checking of PoS annotation			X	X	X

T4.3: Syntactic annotation					
T4.3.1: Select portions of data for syntactic annotation			X	X	X
T4.3.2: Find parsers for NL, FR, EN		X			
T4.3.3: Prepare protocols for syntactic annotation			X		
T4.3.4: Annotate the data			X	X	X
T4.3.5: Manual checking of syntactic annotation			X	X	X
T4.4: User documentation					
T4.4.1: Write user documentation (linguistic annotation)					X
<i>Deliverables</i>					
D4.1 Protocols for lemmatization and part-of-speech tagging			X		
D4.2 Release of the data files enriched with lemmata and PoS tags					X
D4.3 Protocols for syntactic annotation			X		
D4.4 Release of the data files enriched with syntactic annotations					X
D4.5 User documentation - part 4: Linguistic annotation					X
WP5: Exploitation and Corpus management					
<i>Tasks</i>					
T5.1: DPC server					
T5.1.1: Set up a server for storing DPC data	X				
T5.1.2: Design and document server structure	X	X			

T5.2: DPC website					
T5.2.1: Design and document website structure	X	X			
T5.3: Web interface to query the corpus					
T5.3.1: Prepare specifications of web interface		X			
T5.3.2: Implement corpus query tool					X
T5.4: Write documentation					
T5.4.1: Write user documentation for exploitation of the corpus					X
<i>Deliverables</i>					
D5.1 Specifications web interface		X			
D5.2 Corpus exploitation tools: web interface to query data					X
D5.3 Full text TEI-format XML-files + alignment links					X
D5.4 User documentation - part 5: exploitation - final version					X
WP6: Validation					
<i>Tasks</i>					
T6.1: Internal validation by a user group					
T6.1.1: Set up a user group	X				
T6.1.2: Prepare data for the internal validation			X	X	X
T6.1.3: Get feedback from the user group				X	X
T6.2: External validation					

T6.2.1 Set up external validation group					
T6.2.2: Prepare validation plan for Xplanation		X			
T6.2.3: Prepare data for the external validation by Xplanation			X	X	X
T6.2.4: External validation by Xplanation			X	X	X
T6.2.5 Prepare validation plan for CST		X			
T6.2.6: Prepare data for the external validation by CST			X	X	X
T6.2.7: External validation by CST			X	X	X
T6.3: Write documentation					
T6.3.1: Prepare validation reports for each WP				X	X
T6.4: Organize a workshop					
T6.4.1: Design a workshop (location, dates, participants, program)					X
T6.4.2: Prepare presentations for the workshop					X
T6.4.3: Host the workshop					X
T6.4.4: Prepare a report on the workshop					X
<i>Deliverables</i>					
D6.1 External validation report for each WP					X
D6.2 Workshop					X
D6.3 Internal validation			X		

Part 2: Report on Semester 3 (S3)

Summary

The main focus of S3 (1 August 2007 - 31 January 2008) was on the optimization of all topics involving data acquisition. In the first two semesters, contacts with text providers were not always successful, since the IPR agreement models were not transparent enough and required lots of modifications. The acquisition procedure has improved considerably since the TST-Centrale has validated two basic IPR agreement models: one for commercial use and one for publishers. Moreover, the validation of a short version of the standard IPR agreement for commercial use is in the pipeline, that will be a very useful model for providers of text materials publicly available on the web.

The finalisation of the IPR agreement models has resulted in better communication with text providers, so that the data acquisition procedure is eventually getting under control. During S3, this has resulted in an increase of text data, which is documented in the data matrix provided in the annex.

During this reporting period, a number of protocols have been added concerning the annotation of sub-sentence alignment, the lemmatization and PoS tagging procedure for the three languages. The syntactic annotation will be carried out for a subselection of the corpus, for which the co-operation of the LASSY team is being organised.

There has been some change of personnel in the DPC team. At the end of her contract, Yulia Trushkina (computational linguist) has decided to take up another job. Antoine Besnehard (linguist French) has also opted for another position. Both positions have been taken over by Hans Paulussen (computational linguist) and Maribel Montero Perez (linguist French). This change of personnel has had no influence in the working conditions of the DPC team. First of all, Yulia Trushkina has carefully prepared the transfer of knowledge to Hans Paulussen, who is well acquainted with the DPC project. Also in the case of Antoine Besnehard the linguistic know-how mainly regarding the acquisition of French text data could easily be transferred.

The next section provides details on realization of each task scheduled for Semester 3.

WP1: Corpus Design and Data Collection*Tasks***T1.2: Contacts with data providers**

Task 1.2 started in Semester 1 and will continue throughout the whole project period. Details on the results achieved in S3 are presented below.

T1.2.1: Establish contacts with data providers

Numerous attempts have been made to contact other data providers than listed in the previous report. Some of these attempts were successful and led to collaboration agreements. Other data providers were deleted from the previous list as they refused cooperation. See Appendix 1.

T1.2.2: Arrange IPR agreements

During S3 considerable progress has been made with respect to IPR agreements: four contracts have been finalised (RIZIV, FOD Justitie, FOD Sociale Zekerheid, BMM), i.e. signed by all three sides (the data provider, K.U.Leuven R&D and the INL).

Four contracts have been signed by the data providers (Ons Erfdeel, DNS, communication agency Forte, GazeTTe) and forwarded to K.U.Leuven R&D.

For a number of contracts negotiations are currently conducted on modifications, suggested by the data provider, and the counter proposals, made by the HLT Agency (TST-centrale). Such is the case of the NMBS Group, IBM and Sociale Verzekeringsbank. The NMBS Group has accepted the changes and is now ready to sign the contract. IBM and Sociale Verzekeringsbank still have to agree to the last modifications of the HLT Agency.

C2 contracts for commercial use were sent to Westtoer, Bosch, premier.be, Quarterly, Eli Lilly, Eisai, Cis Bio, Campuskrant and KBC and their reply is awaited.

A cooperation agreement has been achieved with the Flemish government.

Some data providers, mainly pharmaceutical companies, refused to sign the C2 agreement for commercial use because of its length and the complexity. It has therefore been decided to use a short (one-page) version of the C2 agreement for commercial use.

During this semester, significant progress has been achieved in the contacts with publishers for whom a special IPR agreement was needed and drawn up during this semester. It has been validated by the HLT Agency and sent to Transmed, Roularta, Lannoo and Ons Erfdeel. The last one has already signed the contract.

A cooperation agreement has been achieved with De Standaard.

IPR issues are currently being negotiated with some other publishers: Actes Sud and The Guardian/ The Observer.

Deliverable 1.2 gives a detailed overview of the current state of affairs of the IPR agreements that have been sent to the data providers

T1.2.3: Collect data from providers

The corpus size increased to **5.4 million tokens**, covering data from 22 providers. Appendix 1 contains an actualised list of the DPC text providers.

T1.3: Initial processing of the data

Task 1.3 started in Semester 1 and, with the exception of subtask 1.3.1 finished in S1, will continue through the whole project period. Details on new results achieved in S3 are presented below.

T1.3.5: Compile a list of metadata for incoming data

The metadata are initially stored in ODS files (ODS = Open Document Spreadsheet), as part of OpenOffice. The ODS files will be later on automatically transformed to XML-documents.

T1.3.7: Prepare an overview of collected data

An overview of all DPC data is recorded in a form of a matrix and kept updated on the DPC plone-site ([date]_Overzicht_jul.doc) by the computational linguist. The DPC matrix is also presented and described in the deliverable Data Collection Protocol (D1.3).

Deliverables

D1.2: IPR agreements:

Deliverable 1.2 gives a detailed overview of the current state of affairs of the IPR agreements that have been sent to the data providers.

D1.3: Data Collection Protocol

Since S2, a large number of documents have been added to DPC. This is reflected in the DPC matrix added and documented in the Data Collection Protocol.

WP2: Text Normalization

T2.1: Cleaning the incoming data

T2.1.1: Prepare protocol for data cleaning

T2.1.2: Convert the data to txt format

T2.1.1 and T2.1.2 were developed in S2

T2.1.3: Normalize character encoding

Most of the original documents are encoded in one of the following code pages: single byte encoding CP1252 or ISO-8859-1, or unicode encoding UTF8. However, some characters in CP1252 are not represented in ISO-8859-1, and some tools are limited to particular single byte encodings.

Therefore, a mapping is required. The mapping tool `neutralize_CP1252.pl` has been adapted to reduce incompatible

S3: THIRD PROGRESS REPORT 31/01/08

characters. This will be further tested in S4, where the required character mapping conversion per tool will be listed.

T2.1.4: Introduce presentational mark-up in the data

Developed in S2.

T2.2: Sentence splitting:

This part was developed in S2.

T2.3: Tokenization:

For the three languages, tokenization is performed as part of the PoS tools selected for each language. In some cases (e.g. apostrophes), some preprocessing or postprocessing of the tokenization is required, depending on the language and the PoS tools used. Further information is found in WP4 Linguistic Annotation.

T2.4: The corpus data will be released in XML format. Some of the tools output the data in XCES format. This format will later on be converted to TEI. The mapping from XCES to TEI is foreseen for S4. The metadata are at the moment stored in ODS-format (ODS = Open Document Spreadsheet), which is easier to edit than XML-files. A first version of a mapping tool `odt2xml.pl` is implemented which converts the ODS files into XML data. Meta data will be stored in separate XML documents. A mapping tool from XML to ODS is also being implemented, which will allow to create metadata files automatically for those text files which already contain metadata. Fusion of XML metadata files and the related DPC texts will also be envisaged.

WP3: Alignment*Tasks*

In semester S2, the focus regarding the alignment procedure was on Paragraph alignment (T3.1) and Sentence alignment (T3.2). This semester, we had a closer look at sub-sentential alignment (T3.3).

T3.3: Sub-sentence alignment

T3.3.1: Select portions of data for sub-sentence alignment

The texts that will be aligned under sentence level are subset of the data that will be annotated on a syntactic level (see T4.3.1).

T3.3.2: Prepare protocol for sub-sentential alignment

The procedure for sub-sentential alignment is entirely based on the procedure and tools developed (or adapted) by Lieve Macken in the framework of her PhD thesis. For the language pair Dutch-English, her annotation guidelines will be used.

For the language pair Dutch-French, a first version of the annotation guidelines were developed based on the Dutch-English guidelines.

T3.3.3: Manually align selected portions of data on sub-sentence alignment

No action yet.

*Deliverables***D3.3: Protocol for sub-sentence alignment**

Subsentence Alignment Protocol (D3.3) is attached to the report. The protocol introduces the problem of sub-sentential alignment and contains the annotation guidelines for the language pairs Dutch-English and Dutch-French. It provides details on the annotation tool and conversion tool (to convert the sub-sentential correspondences to a human-readable format).

WP4: Linguistic Annotation*Tasks*

In the previous semester a number of annotation tools (lemmatization, PoS tagging) have been compared for the three languages. This semester, the available tools have been tested and initial versions of protocols for linguistic annotation have been prepared, regarding lemmatization (T4.1.2) and PoS tagging (T4.2.2).

T4.1.2: Prepare protocols for lemmatization**T4.2.2: Prepare protocols for PoS tagging**

Next to aligning DPC at sentence level, the corpus will also be annotated linguistically. Since the parallel corpus consists of three languages, which have different annotation standards, different encoding systems are being used. Since DPC is first of all a Dutch corpus (with alignments in English and French), the linguistic annotation of D-COI was taken as encoding system. In the case of English, the Penn Treebank conventions are used. For French, the Cordial and TreeTagger scheme are used, depending on the type of text used.

*Deliverables***D4.1: Protocols for lemmatization and PoS tagging**

The enclosed protocol on lemmatization and PoS tagging discusses the different approaches used for linguistic annotation specific for the three languages. In some cases, the workflow involves manipulations over different platforms (Windows and Linux). Some of the tasks are carried out by partners with experience of a number of annotation tools for Dutch.

D4.3: Protocols for syntactic annotation

Part of DPC will be syntactically annotated. In the first version of the protocol, the annotation workflow is explained for Dutch, a task which will be carried out by the LASSY team.

WP5: Exploitation and Corpus management

The first part of this WP consisted in setting up the DPC server for storing the DPC data (T5.1) and setting up a website for communication on the DPC project (T5.2).

T5.3: Web interface to query the corpus**T5.3.1: Preparation of the web interface specifications**

The deliverable 5.1 has been updated during the third semester. The web interface specifications have been made more accessible and readable for the user group. The document contains concrete examples of search queries, described in detail.

A survey of the user group is scheduled for S4.

WP6: Validation**T6.1: Internal validation**

Thanks to publications and presentations, DPC is being talked about, even before it is finished, which shows that there is a clear need for a Dutch parallel corpus. After having consulted the representatives of the STEVIN committee, present during the DPC meeting of 17 September 2007, we have decided to issue pre-releases of selections of the DPC data. The pre-releases will only be available for researchers belonging to K.U.Leuven and HoGent (in fact the institutions of the core research team), under the condition that the researchers involved will sign an agreement to not disseminate the data, and under the condition that they will return an evaluation report.

T6.2: External validation**T6.2.2: Prepare validation plan for Xplanation**

A first draft of the external validation plan for Xplanation had been sent to Xplanation.

The validation plan consists of two parts. In the first part, the existing terminology extraction system of Xplanation will be tested in two different settings (with and without DPC as reference corpus). In the second part, the added value of DPC to enhance the current terminology extraction system will be described with the focus on the sentence alignment and extra linguistic annotations.

T6.2.5: Prepare validation plan for CST

The validation plan for CST would consist of an informal validation plan (during production of the DPC data) and a formal validation plan at the final phase of the project. In the coming weeks a first sample set of aligned data will be sent to CST, so that the actual validation plan can be finalised.

Part 3: Planning of Semester 4 (S4)

During semester 4, acquisition of data will continue. Thanks to the standardisation of the model contracts, it will become easier to contact the remaining text providers. Incoming data will be cleaned and aligned as usual. From now on, the data will also be linguistically annotated so that pre-releases can be prepared for internal validation by researchers from K.U.Leuven & HoGent.

Further work will be carried out to improve the workflow of alignment and linguistic annotation, and output formats in XML will be developed, so that the main focus can go to the documentation tasks (D1.1, D1.4). Next to aligning and annotating the data, a procedure will be developed to check linguistic annotation and alignment output.

Part 4: DPC presentations in Semester 3 (S3):

The following presentations have been made during Semester 3:

- Presentation EUROCALL conference (7 September 2007, Coleraine)
- Presentation ACCENTA 2007 (21 September 2007, Gent)

The following articles have been published during Semester 3:

- Dutch Parallel Corpus : A multifunctional and multilingual corpus. Hans Paulussen, Lieve Macken, Julia Trushkina, Piet Desmet, Willy Vandeweghe. Cahiers de l'Institut de Linguistique de Louvain (CILL) 32. 1-4 (2006), 269-285. Louvain-La-Neuve.
- Dutch Parallel Corpus: MT corpus and translator's aid. Lieve Macken, Julia Trushkina, Lidia Rura. The 11th Machine Translation Summit, Copenhagen, Denmark, September 10-14 2007.
- Dutch Parallel Corpus: A Multilingual Annotated Corpus. Lieve Macken, Julia Trushkina, Hans Paulussen, Lidia Rura, Piet Desmet, Willy Vandeweghe. The fourth Corpus Linguistics conference, Birmingham, 27-30 July 2007

Appendix 1: DPC Text Providers

- Barco NV
- BMM
- Bosch
- Campuskrant
- Cis Bio
- De Post
- De Standaard
- DNS
- Editions Actes Sud
- Eisai
- Electrolux
- Eli Lilly
- EU (Europese Commissie en Europees Parlement)
- FOD Justitie
- FOD Sociale Zekerheid
- Ford Dodge Animal Health
- GazeTTe
- IBM
- Infrabel
- KBC
- Ministerie van de Vlaamse gemeenschap
- NMBS Holding
- Orphan
- Provinciale diensten voor toerisme (Vlaanderen): Westtoer
- Quarterly
- RIZIV
- Site Guy Verhofstadt
- Sociale verzekeringsbank (NL)
- Stichting 'Ons erfdeel'
- The Guardian/ The Observer
- Transmed, Site met 7 medische tijdschriften
- Uitgeverij LANNOO
- Uitgeverij ROULARTA
- Verenigde Naties
- Vlaamse overheid

Appendix 2: Deliverables of S3

The following deliverables have been scheduled for Semester 3:

The following four deliverables were scheduled as planned:

- D1.2 IPR agreements
 - D1-2 IPR Agreements.pdf
- D3.3 Protocol for sub-sentence alignment
 - D3-3a Subsentential Protocol.pdf
 - D3-3b Subsentential NL-EN.pdf
 - D3-3c Subsentential NL-FR.pdf
 - D3-3d Subsentential Handalign.pdf
- D4.1 Protocol for lemmatization and part-of-speech tagging
 - D4-1 Protocol for lemmatization and tagging.pdf
- D4.3 Protocol for syntactic annotation
 - D4-3 Protocol for syntactic annotation.pdf

The following two deliverables required some update:

- D1.3 Protocol for data collection
 - D1-3a Data Collection.pdf
 - D1-3b Data Acquisition.pdf
 - D1-3c Data Matrix.pdf
 - D1-3d Data Collection.pdf
- D5.1 Specifications web interface
 - D5-1 Specifications web interface.pdf

One new deliverable has been introduced as part of the internal validation task (T6.1):

- D6.3 User agreement and validation report model
 - D6-3 Internal Validation.pdf
 - D6-3-1 Beoordelingsovereenkomst.pdf
 - D6-3-2 Validatierapport.pdf

All submitted deliverables can be found on the STEVIN-WIKI.

Progress Report STEVIN Projects

Project Name Large Scale Syntactic Annotation of Written Dutch
Project Number STE05020
Reporting Period september 2007 - march 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

Not applicable.

1.3 Changes requested

The timing of the project, and the completion of its deliverables, faced four problems. The first problem concerned the difficulty to hire a post-doc in Groningen. We are happy to be able to state that as of February 1, 2008, we were able to find a suitable post-doc, so this problem has now been solved.

The second problem concerned the lack of availability of the final D-Coi corpus. Although we have (as members of the D-Coi consortium) a fairly clear understanding about the contents of the D-Coi corpus, we need the precise final version in order to ensure that the details between D-Coi and Lassy line up. The delay of D-Coi causes corresponding delays for our deliverables 1.1 and 1.2.

A related problem concerns our desire to cooperate with the STEVIN Dutch Parallel Corpus (DPC) project. Since DPC has an interest to have part of their corpora syntactically annotated, and Lassy has an interest to include some more Flemish material in its corpus selection, we are currently negotiating with DPC what corpus material can be selected. This is a further cause for delay of deliverable 1.1 and 1.2.

The final problem concerns the delayed start of the D-Coi successor project, now known as SoNaR. It is our desire to base Lassy Large on the corpus selection for SoNaR. This final problem is another cause for the delay of deliverable 1.2.

We request the following changes with respect to the time line of Lassy. First, we propose to set the end date of the project to May 1, 2010 (rather than Nov 1, 2009), and to add six months to each of the deliverables. As for the deliverables 1.1 and 1.2, we would propose to move the dates for these deliverables to September 2008.

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.

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/>

During ACL 2007, Lassy sponsored the ACL Workshop entitled *Deep Linguistic Processing*. The Lassy sponsoring enabled an invited keynote lecture by Annete Frank, entitled *Across*

Languages and Grammar Paradigms - New Perspectives on Resource Acquisition, Grammar Engineering and Application.

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 Computational Linguistics in the Netherlands.

1.5.1 Publications

- Nelleke Oostdijk, Martin Reynaert, Paola Monachesi, Gertjan van Noord, Roland Ordeman, Ineke Schuurman, Vincent Vandeghinste. From D-Coi to SoNaR: A reference corpus for Dutch. Accepted for LREC 2008.
- Roelien Bastiaanse and Gosse Bouma. Linguistic Complexity and Frequency in Agrammatic Speech Production. 2008. Submitted.
- Gosse Bouma, Geert Kloosterman, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jrg Tiedemann. Question Answering with Joost at CLEF 2007, CLEF 2007 Working Notes.

1.5.2 Presentations

- Gertjan van Noord, Self-trained Bilexical Preferences for Improved Syntactic Disambiguation. CLIN, December 7, 2007, Radboud University, Nijmegen.

1.6 Exploitation of the results

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

2 Progress per deliverable

2.1 Deliverable 1.1

As described in our previous report, the corpus selection for Lassy has been done on the basis of input from the Lassy user group.

The selection of Lassy Small is in place, but there are two factors which cause a delay before the deliverable is final. First of all, because we still have no access to the final D-Coi corpus, the precise word counts of the various D-Coi parts are still unreliable. Secondly, we have an agreement with the DPC project to supply us with corpus material that we will annotate syntactically, but at present we have not yet received information concerning the precise specification of this material.

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

Note that most of the material is material originally collected and pre-processed in D-Coi, except for the lines marked *new* in the table. We decided to switch from the Wikipedia material

material	POS-tagged and lemmatized	syntactically annotated
wikipedia (new)	100,000	100,000
DPC (new)	100,000	100,000
e-magazines	7,000	7,000
wikipedia	200,000	200,000
brochures	60,000	60,000
autocues	200,000	300,000
total in Lassy	667,000	767,000
D-Coi	500,000	200,000
Lassy Small	1,167,000	967,000

Table 1: Corpus selection Lassy Small

material	POS-tagged, lemmatized, syntactically annotated
D-Coi minus Wikipedia, Europarl	22M
CLEF Wikipedia	58M
Europarl version 3	38M
selection from TwNC/Mediargus	382M

Table 2: Corpus selection Lassy Large

collected in D-Coi to a more recent version of Wikipedia which was made available to CLEF participants. The reason is, that the CLEF-version of Wikipedia (provided by the University of Amsterdam) is much better represented in XML, making the corpus clean-up and tokenization much more successful, while at the same time keeping track of the meta-information from Wikipedia. The resulting (annotated) material should therefore be much more useful. If time allows, we are also considering re-annotating the old Wikipedia material with the corresponding new material.

Furthermore note that we plan to put more effort in POS-tagging and lemmatization in comparison to the project proposal. This is motivated by our desire to work with Flemish DPC material.

The deliverable 1.1 can be submitted as soon as the final release of D-Coi is accessible to us, and we have a final agreement with the colleagues from the DPC project.

2.2 Deliverable 1.2

This deliverable is somewhat behind schedule, due to the fact, once again, that we have no access to the final D-Coi release. Also, 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 a fall-back option if material from SoNaR is not available in time.

layer	annotated	target
lemmatization	560	667
POS-tagging	560	667
Syntactic	614	800

Table 3: Progress of Annotation Efforts. All numbers are Kilo-words.

2.3 Deliverables 2 and 3

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

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/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. Alpino is available from <http://www.let.rug.nl/~vannoord/alp/Alpino/>.

2.4 Deliverables 4

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 CLEF-version of) Wikipedia (58 million words), Europarl version 3 (38 million words), TwNC (Dutch newspapers; almost 400 million words) and part of Mediargus (Flemish newspapers; about 100 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 the 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 further improvements to Alpino available later.

2.5 Deliverables 5

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

2.6 Deliverables 6

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.

2.6.1 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.

2.6.2 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 as 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.

2.6.3 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.

2.6.4 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. This 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. Once again, Joost obtained the best result for Dutch QA at CLEF in 2007.

Project name	MIDAS (Missing Data Solutions)
Project number	STE05030
Reporting period	September 30, 2007 - April 1, 2008

Participants

Nuance: Rudi Vuerinckx
K.U.Leuven: Hugo Van hamme
R.U.Nijmegen: Jort Gemmeke, Bert Cranen

1. Summary of the project

The general status of the project is as follows:

- despite the late start due to recruiting problems, the scientific progress is encouraging (see submitted and accepted publications). New missing data masks were designed but need more testing (WP3.1)
- the baseline MDT system is up and running (WP2) and a significant speed-up was already obtained (WP4.1)
- the benchmark of the baseline MDT system is not fully complete due lack of a grammar compiler from the SPRAAK project

In the near future, the consortium will work towards joint publications.

1.1. Overview deliverables (+time of deliverable) according to the proposal

Name	Month (proposal)	Month (agreed)	Description	Contributors
D1	3	8	Training and test material defined	SSFT, KUL, RUN
D2	6	12	Benchmark report of state-of-the-art system	SSFT
D3	12	16	Benchmark report of baseline MDT system	KUL, RUN
D4	24		Benchmark report of version 2	KUL, RUN
D5	30		Benchmark report of version 3	KUL, RUN
D6	42		Final MDT system integrated in SPRAAK	KUL, RUN
D7	44		Benchmark report of final version	KUL, RUN

1.2. Previously completed deliverables

D1 and D2 (on Wiki)

1.3. Changes requested (contents/timing of deliverables) and motivation

During the site visit, the problem of part-time employment (80%) of Jort Gemmeke at R.U.Nijmegen was discussed. The consensus was to extend the participation of RUN from month 42 to month 48 (also end of the project). It is therefore reasonable to achieve the following milestones on the following dates (as discussed at the site visit):

WP3:

- Improved binary masks (WP3.1) - M20 RU - medium risk
- Soft masks (WP3.2) - M27 RU - medium risk
- Sound class dependent masks (WP3.3) - M38 RU - medium risk
- Confidence measures for MDT decoders (WP3.4) - M42 RU & KUL - medium to high risk

WP2: final MDT system - M48 KUL&RUM - medium risk

Since both RUN and KUL can now contribute till the end of the project, it would make more sense to reschedule the final version of the MIDAS recognizer at month 48 instead of 42, and to organise an annual benchmark. This leads to the following deliverable table:

Name	Month (agreed)	Month (new)	Description	Contributors
D1	8	8	Training and test material defined	SSFT, KUL, RUN
D2	12	12	Benchmark report of state-of-the-art system	SSFT
D3	16	16	Benchmark report of baseline MDT system	KUL, RUN
D4	24	24	Benchmark report of version 2	KUL, RUN
D5	30	36	Benchmark report of version 3	KUL, RUN
D6	42	48	Final MDT system integrated in SPRAAK	KUL, RUN
D7	44	48	Benchmark report of final version	KUL, RUN

1.4. Employee involvement in relation to the original plan

Cumulative effort table in person-months

	anticipated	Yujun Wang KUL	Jort Gemmeke RUN	Rudi Vuerinckx Nuance	Bert Cranen RUN	Kris Demuynck KUL	total
D1	3			1.0			1.0
D2	1.6			1.1			1.1
D3	14	5.2				1.7	6.9
D4	35	5.8	12.8		2.4		21.0
D5	24						
D6	22						
D7	2						
total	101.6	11.0	12.8	2.1	2.4	1.7	30.0

Like for D1 and D2 (see report second period), we have underspent on D3 by lack of staff. Yet, by pulling in effort from permanent staff and Maarten Van Segbroeck (IWT funding), this had little impact on the content of the deliverables.

1.5. Dissemination of the results

Presentations or posters:

- TST-dag September 2006 (Antwerpen)
- TST-dag September 2007 (Hoeven)

Conference publications:

[1] Noise reduction through Compressed Sensing - Jort Gemmeke and Bert Cranen (*submitted to Interspeech 2008*)

[2] State dependent oracle masks for improved dynamical features - Jort Gemmeke and Bert Cranen (*submitted to Interspeech 2008*)

[3] Using sparse representations for missing data imputation in noise robust speech recognition - Jort Gemmeke and Bert Cranen (*submitted to EUSIPCO 2008*)

[4] Gaussian Selection for Fast Decoding with PROSPECT features in Missing Data Techniques - Yujun Wang and Hugo Van hamme (*submitted to Interspeech 2008*)

[5] Classification on incomplete data: imputation is optional - Jort Gemmeke (*Benelearn 2008*)

[6] On the relation between statistical properties of spectrographic masks and recognition accuracy - Jort Gemmeke, Bert Cranen, Louis ten Bosch (*IASTED SPPRA 2008*)

Workshop publications:

[7] Noise robust digit recognition using sparse representations - Jort Gemmeke and Bert Cranen (*ISCA 2008 ITRW "Speech Analysis and Processing for knowledge discovery"*)

1.6 Exploitation of the results

- o (New) collaborations
- o (Accepted) project proposals based on present project
R.U.Nijmegen (CSLT) and K.U.Leuven (ESAT/PSI) have submitted a project proposal "BATS" of type IM-PACT to the IBBT/ICT-Regie which was accepted. The topic is indexation and search in audio (and multimedia containing audio) archives. One of the work packages deals with robustness to background noise and music in particular, which will be approached with missing data techniques as well
- o Other (patent, ..)

2. Progress for deliverable "Benchmark report of baseline MDT system" (M16)

2.1. Activities completed in the past 6 months

- *Building an acoustic model for Dutch that is compatible with the PROSPECT speech representation.*
As mentioned in the previous report, the acoustic models built for Dutch (Netherlands - shorthand "DUN") before October 2007 did not seem to meet the quality requirements. They were rebuilt and appear to yield acceptable accuracy now (see benchmark report).
- *Baseline MDT system based on ESAT's recognizer*
The following activities were completed:
 - Acoustic model for DUN (see previous bullet)
 - Codebook design for VQ-based masks for DUN. These replace the WP3.1 masks of the DBN approach - see previous report.
 - Solving software problem mentioned in previous report related to speaker restart
 - Building the grammar for digit string and isolated word recognition. This should have been a "push-the-button task" for the grammar compiler of the SPRAAK project. However, due to the delays in this project, we hand-crafted the grammars for these two simple tasks. For the same reason, we were unable to run the more complex tasks (a.o. the "when" grammar) - see benchmark report D3 of the baseline MDT system
 - Selection of a subset of the test database for the isolated word tasks. As mentioned in the report on the test set design, the number of utterances for the isolated word task is significantly higher than for the other tasks. Processing them all takes too much CPU time in our current implementation (to improve with the work in WP4). Hence, a subselection of the test set was made.

2.2. Problems and solutions

To finish the "complex" grammars, we await the delivery of the grammar compiler for SPRAAK. Once delivered, there should be no major problems (except hard work of course) to complete the benchmark.

However, from these two recognition tasks, we can obviously conclude that the performance of the MDT system is inferior to the Nuance Vocon 3200 (state-of-the-art recognizer) - see D3.

2.3. Proposed schedule for the upcoming period

The SPRAAK project is to make a software delivery on May 31. The benchmarks for the other grammars will be run than as well.

3. Progress for deliverable “MDT-based recognizer version 2” (M24)

3.1. Activities completed in the past 6 months

- *WP3.1: Improving binary masks - M20 RU*
 A Support Vector Machine mask estimator was constructed and experiments on the AURORA-2 database were run for evaluation. The estimator is described in [1] (submitted to Interspeech 2008).

WP3.2: Soft masks - M27 RU
 The Support Vector Machine framework supports probabilistic output. No evaluation has been performed yet.
- *WP3.3: Decoding with sound class dependent masks - M33 RU*
 After having established that statistical properties of a mask can be related to speech recognition accuracy [6], a framework for class dependent mask estimation and MDD was constructed. Experiments on the AURORA-2 database to test the feasibility of a statedependent masking approach (using oracle masks) show promising results. These are described in [2] (submitted to Interspeech 2008).
- *WP4.1: Faster algorithms for NNLSQ problem - M18 KUL*
 Additional ideas in Gaussian selection were explored and additional experiments on the AURORA-4 database were run for evaluation: (1) subspace clustering (2) neighbourhood method and (3) the method presented last time.
 The technical report is summarized in [4] (Interspeech 2008 submission). We achieved a factor 3 speedup in the imputation and selected the best of the three designs. The time that has become available due to the rearrangements in WP4.3 will be spent in speeding up the evaluation of individual Gaussians.
- *WP4.2 “Dynamic features” alias “Missing data imputation using wider time-context” - M30 KUL*
 The original goal of this work package was to do joint imputation of static and dynamic features (instead of solving them as independent streams). Implicitly, this means that spectra are modeled/imputed over a wider time window (context). Hence, a more generic approach is to set up an imputation framework which imputes missing data using a worldwide context rather than frame-by-frame. This is deemed to be important especially at SNR's below 0 dB. Promising results were described in the Interspeech [1,2], and EUSIPCO [3] submissions and the ISCA ITWR [7] and Benelearn [5] publications.
- *WP4.3: Soft masks used in back-end - M26 KUL*
 This topic was addressed by Maarten Van Segbroeck (IWT – specialisatiebeurs) and published in the proceedings of the ICASSP 2008 conference. The code is integrated in the MIDAS research platform and will become available in the final MIDAS recognizer. Additional work will be spent in WP4.1, which takes more time than anticipated.

3.2. Problems and solutions

- WP3.3: Decoding with sound class dependent masks – M33 RU*
 A first implementation of the likelihood normalizations proposed by Van hamme (2003) did not yield the desired results. This is possibly due to the masks associated with silence states. Research is currently underway to further analyse the results obtained on AURORA-2.

3.3. Proposed schedule for the upcoming period

- *WP3.1: Improving binary masks*
Training the mask estimator on the provided training data from Nuance.
Evaluation of the mask estimator on the test data from Nuance.
- *WP3.2: Soft Masks*
Evaluation of the mask estimator using the soft-mask back-end.
- *WP3.3: Decoding with sound class dependent masks - M33 RU*
Further investigation of likelihood normalization schemes.
Using mixtures of masks weighted by a-priori state likelihood.
- *WP4.1: Faster algorithms for NNLSQ problem*
Where the focus was on avoiding the computation of Gaussians (with imputation), we will now try to speed up the imputation itself.
- *WP4.2: Missing data imputation using wider time-context*
Consider implementation of the wide time-context imputation method in the missing data system. Needs some care, since imputation is not a non-negative least squares problem any more.

Project name **N-Best**

Project number

Reporting period (from beginning of project until 31/3/2008) May 2006—April 2008

Participants

TNO Human Factors, Soesterberg
CLST+SPEX, Radboud Universiteit Nijmegen
ESAT, KU Leuven
ELIS, Universiteit Gent
HMI, Universiteit Twente
EWI, TU Delft

Participants not part of the project, but who have voluntarily subscribed to participate in the evaluation:
Brno University of Technology, Tsjechie
LIMSI/CNRS en Vecsys research, Parijs
Telecats, Twente.

1. Summary of the project

The STEVIN project N-Best (North- and South Dutch Benchmark Evaluation for Speech recognition Technology) aims at establishing the state-of-the-art performance of large vocabulary, speaker independent, continuous speech recognition (LVCSR) systems for the Dutch language. Because it is the first such evaluation held for Dutch, it also functions as a stimulator for research institutes in the Dutch speaking area and beyond to develop LVCSR systems for Dutch, and to focus on areas that are specific to the Dutch language, such as word compounding and verb constructions spanning the whole sentence.

The project is set up along the lines of earlier held international benchmark evaluations held by NIST in the US for many years in the context of various DARPA projects, and more recently in France by DGA in the ESTER project in the context of the Technolanguage program, funded by the French Ministry of Research. The basic idea is that common Dutch linguistic training resources are defined, collected and distributed amongst participants in the evaluation, who can then build, train and optimize their systems. New speech data is then recorded and annotated, and in a fixed evaluation period the participants can run their systems on the new evaluation data, and submit their system hypothesis results to the evaluator. This party then scores the system hypotheses, and distributes the scores among the participants. Finally, a workshop is organised in which participants present and discuss their approaches, such that researchers can exchange experiences and ideas and can learn from each other in order to improve their systems in the future.

An important result of this evaluation is that a standard evaluation set is defined for LVCSR in Dutch, so that researchers in the future can build new systems and evaluate and optimize with respect to this data set. Also, because it is a formal evaluation, performance results in N-Best will act as a benchmark that can be referred to in future research on Dutch speech recognition.

The domain covered in N-Best is both Broadcast News (BN, comparable to "hu b-4" NIST evaluations) and Conversational Telephone Speech (CTS, comparable to "hu b-5" in NIST evaluations), and separate evaluation sets for North and Southern Dutch accents (as spoken in The Netherlands and Flanders, respectively) have been defined. In each of the four conditions the evaluation material consists of about 2 hours of speech.

The project consortium consists of TNO, acting as coordinator and evaluator, SPEX, collecting and annotating the evaluation data, and research groups of five universities in The Netherlands and Flanders, namely ELIS (Gent), ESAT (Leuven), RU (Nijmegen), EWI (Delft) and HMI (Twente).

The project has almost completed, as the evaluation period (April 2008) has ended, and system results have been submitted and a first round of scoring has been performed. Six participants have submitted in total 54 recognition results for the various conditions and system configurations.

1.1. Overview deliverables (+time of deliverable) according to the proposal

The project started in May, 2006 (m0).

D1.1 Draft evaluation protocol (m6)

D1.2 Final evaluation protocol (m12)

D1.3 Evaluation participation form (m12)

D2.1 List of participants (m22)

D3.1 Specification of data and annotation (m10)

D3.2 Evaluation data plus reference transcriptions (m18)

D4.1 Evaluation data set and support (m22)

D5.1 Proposed phone set, vocabulary, dictionary (site-internal, m12)

D5.2 Training phone-alignment, acoustic and word language models (site-internal, m15)

D5.3 ASR output of development test data, in evaluation protocol format (m18)

D5.4 Systems capable of running evaluation index files (m22)

D6 ASR output of evaluation data (m23)

- D7.1 Workshop on evaluation results (m26)
- D7.2 Draft journal manuscript describing evaluation and results (m27)

Completed deliverables

- D1.1 Draft evaluation protocol
- D1.2 Final evaluation protocol
- D1.3 Evaluation participation form

- D2.1 List of participants

- D3.1 Specification of data and annotation
- D3.2 Evaluation data plus reference transcriptions

- D4.1 Evaluation data set and support

- D5.1 Proposed phone set, vocabulary, dictionary
- D5.2 Training phone-alignment, acoustic and word language models
- D5.3 ASR output of development test data, in evaluation protocol format 4/5 as one partner did not submit dry-run results.
- D5.4 Systems capable of running evaluation index files

D6 ASR output of evaluation data. 4/5---as one partner did not run the evaluation yet.

1.2. Changes requested (contents/timing of deliverables) and motivation

No change in the deliverables is required, other than the timing. The Workshop (D7.1) has been planned to be on 10 September 2008, as a satellite to MLMI'08. This is three months after the original planning (m26), which would have ended up around the summer holidays, given the later than anticipated start of the project. Similarly, this will have repercussions to the draft manuscript (D7.2, m27) which will need input from the workshop.

1.3. Employee involvement in relation to the original plan

note: figures are estimates.

Partner	Planned	October/06	April/07	October/07	April/08	July/08	totaal
TNO	14	2.1	5.2	9.6	14.2		
SPEX	9	0.4	1.3	5.8	8.4		
CLST	7	0.1	0.6	4.5	6.5		
HMI	9	0.1	1.2	5.8	8.4		
ESAT	6	0.1	0.2	2.0	3.0		
ELIS	10	0.1	1.0	6.4	9.3		
EWI	9	0.1	1.0	5.8	8.4		
Total	64	3.0	10.5	39.8	58.1		

1.4. Dissemination of the results

- o **Publications (specify the type – refereed journal, proceedings, workshop,).. ,**
 - o Judith Kessens and David van Leeuwen. N-Best: The Northern and Southern Dutch Evaluation of Speech Recognition Technology, Proc. Interspeech, Antwerp, 2007, 1354–1357. This paper receiver the COCOSDA prize for best paper related to standards in speech databases and assessment techniques.
 - o Marijn Huijbregts, Roeland Ordelman and Franciska de Jong
Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition. Proceedings of the Second International Conference on Semantic

and Digital Media Technologies, SAMT 2007, Genoa, Italy, Lecture Notes in Computer Science, volume 4816, Springer Verlag, Berlin, ISBN 3-540-77033-X, pp. 78--90, December 2007.

- o David van Leeuwen, Contribution to DIXIT special issue on STEVIN projects, December 2007.
- o **Presentations (specify the type,)..**
 - o STEVIN day in Antwerpen, 11 september 2006. Presentaion of the N-Best project, by David van Leeuwen (TNO), project leader N-Best.
 - o Poster related to the Interspeech article above, 29 augustus 2007, Interspeech, Antwerp, Judith Kessens and David van Leeuwen, "N-Best: The Northern and Southern Dutch Evaluation of Speech Recognition Technology".
- o **Outreach activities**
 - o Several Calls for Participation have been sent out on the various research email lists.
- o **Other**

1.6 Exploitation of the results

- o **(New) collaborations**
Outside the STEVIN project N-Best, three parties have subscribed to participate in the evaluation. These are the Brno University of Technology (Czech Republic), LIMSI + Vecsys Research (France) and TeleCats (Twente, The Netherlands).
- o **(Accepted) project proposals based on present project**
N-Best is related to the STEVIN project PRAAT, where a LVCSR toolkit for the the Dutch language community is developed. Evaluation is "outsourced"t o N-Best for this project.
- o **Other (patent, ..)**

2. Progress per deliverable

2.1. Activities completed

One of the first things that have been carried out was the specification of the evaluation protocol. In here, the precise definition of the task, the conditions, the evaluation metric, method and rules are specified. Based on existing evaluation protocols by NIST and within ESTER, this document was drafted by the evaluator but in discussion with data collector and ASR sites, but also in consultation with Jon Fiscus from NIST.

The evaluator has made a specification of the test data, in terms of the variability in content, speakers, subjects, regions, channels, etc., in discussion with the data collector. This has functioned as a guide for the data collector to find sources to match the specifications, and ways to provide the data with annotation (transcription of the speech and meta-information like speaker identity, channel quality).

The coordinator has negotiated the use of textual data from Flemish and Dutch newspapers for N-Best training with data suppliers Mediargus (in Belgium) and PCM (Netherlands, though HMI Twente as mediator). The same negotiations have been made with the TST-Centrale for use of the Spoken Dutch Corpus (CGN) for acoustic training within CGN. Similarly, The data collector has negotiated the use of broadcast news material for N-Best, and for later distribution through the *Nederlandse Taalunie*, from Dutch and Belgium data providers. These legal issues have taken a fair bit of effort, which was not anticipated at the time of writing of the proposal.

The coordinator has designated part of the acoustic training material of CGN for *development testing*. The aim was to set up a “dry-run” of the N-Best evaluation, using this development data. This meant that the data selection and formats were to be made similar to the new evaluation material. With the given CGN data, this proved to be hard in some cases. For instance, the Dutch BN material did not consist of full news broadcasts, but of only selections of the broadcast. Therefore the CGBN material was augmented with some news broadcast recordings and transcriptions made by TNO. The CTS part of the dry-run data proved to be difficult in the sense that many of the CGN conversations showed very high levels of cross-talk. A lot of effort was spent in defining segments which did not have speakers speaking simultaneously, to avoid cross-talk.

The five ASR partners have obtained all parts of training data through signing the appropriate licenses. This has allowed them to develop Dutch ASR systems. These system were then applied to the dry-run data during the dry-run evaluation period. The results were scored by the evaluator, distributed among the ASR partners, and in December 2007 a small workshop was held in Roosendaal to discuss progress and results, and exchange ideas and experiences related to the dry run. Not every partner managed to complete the dry-run in time, but all partners were represented at the workshop.

The data collector has obtained BN data from the owners of the data. For CGN, a recording platform similar to that used in collecting CGN was set up, and subjects according to the test data specification were recruited. After recordings were made, a first transcription pass was made. This data was sent to the evaluator, who selected material based on content and meta-data. This selection was sent back to the data collector who verified the transcriptions in a second annotation run. The results thereof were sent back to the evaluator who combined the various selection into sets of a single audio file, reference transcription file and segment information file. These have then been made anonymous--stripping off information about the source-- and have been prepared for distribution. A final consistency check was made of all the data.

The coordinator has further advertised N-Best amongst colleagues and mailing lists. Most effective in obtaining external participants has turned out direct contact during (other evaluation) workshops. In total participants external to the project partners subscribed to the evaluation. One of these, TeleCats, had to withdraw, unfortunately.

The data has been distributed to all sites by download over the internet. Sites were given about a month to process all data and submit results, which could be done by e-mail. In total six participants have delivered their system results for all required four conditions in time, and in total 54 submissions were made among the sites. Some sites had multiple systems, some had multiple run-time conditions for the same system.

The deadline of the evaluation (2 May 2008) is right at the end of this reporting period, but we can report that basic scoring has been performed without too many glitches, and a period of adjudication is on-going.

2.2. Problems and solutions

One site (ESAT) had indicated that during the evaluation period not enough person power was available to run the evaluation. They will run the evaluation after the official evaluation period, but before the workshop. TeleCats had to withdraw. They were going to participate with third-party ASR software, but the evaluation data turned out to be too challenging and the third party systems not flexible enough to utilize the (additional) N-Best training. This is a bit of a pity, as the evaluation plan had specifically been altered to allow participation with third-party systems.

2.3. Proposed schedule for the upcoming period

After the adjudication period, the evaluator will revisit scoring and transcription once more, and results will be finalized. Then ASR participants, data collector and evaluator will prepare for a final workshop to be held 10 September at TNO in Soesterberg, the Netherlands. As a result the coordinator will prepare a manuscript about the N-Best evaluation and its results, and partners are expected to present their work in related papers.

The evaluation data and all transcriptions will have to be handed to the sponsor, the TaalUnie, who can then maintain the data set through the *TST-Centrale*.

Finally, before 31 Dec 2008, ASR sites must remove training data from their systems if they obtained the data through a license for evaluation within N-Best.

Projectnaam:

STEVIN can PRAAT

Projectnummer

STE-05-35

Rapporteringsperiode

jan 2008 - mei 2008

Deelnemers

prof. dr. P. Boersma
prof. dr. F. Hilgers
prof. dr. V. van Heuven
dr. H. van den Heuvel
dr. D.J.M. Weenink

Uitvoerders van het onderzoek

Paul Boersma (projectleider)
David Weenink (uitvoerder)

1. Samenvatting

We propose to develop a number of improvements and added functionality to the PRAAT program that will then additionally and freely become available for speech scientists. This is the final report.

1.1. Deliverables (+tijdstip) volgens het projectvoorstel

The seven deliverables are:

- D1. Klatt Synthesizer
- D2. Sound-Follows-Mouse
- D3. Graphical Manipulation of Formant Frequencies and Bandwidths
- D4. Software Bandfilter Analysis
- D5. Availability of Elementary Scientific Functions
- D6. Improve Formant Frequency Measurements
- D7. Search and Replace with Regular Expressions

No ordering of the deliverables nor any time schedule were specified in the project.

1.2. Reeds eerder afgewerkte deliverables

1.3. Gewenste wijzingen (inhoud/tijdstip van deliverables)

All deliverables are finished except for 1. In conformance with the visting committee, the Klatt synthesizer will be delivered in october 2008.

1.4. De personeelsinzet in relatie tot het oorspronkelijke plan

The number of hours has been in concordance with the description of the project.

Boersma has worked in conformance with the proposal.

	voorzien	per 1	per 2	per 3	per 4	per 5	per 6	per 7	per 4	totaal
D1&D3										200
D2										340
D4										160
D5										120
D6										120
D7										120
totaal										1060

1.5. Disseminatie van resultaten (publicaties, lezingen, ...)

2. Vorderingen per deliverable

- D1. Klatt Synthesizer (25% complete)
- D2. Sound-Follows-Mouse (100% complete)
- D3. Graphical Manipulation of Formant Frequencies and Bandwidths (100%)
- D4. Software Bandfilter Analysis (100% complete)
- D5. Availability of Elementary Scientific Functions (100% complete)
- D6. Improve Formant Frequency Measurements (100% complete)
- D7. Search and Replace with Regular Expressions (100% complete)

2.1. Gerealiseerde werkzaamheden in afgelopen periode

D1 and D3. Klatt Synthesizer & Graphical manipulation of formants and bandwidths:
The main data structure for the synthesizer and the manipulation will be the FormantGrid.
This will be integrated into new datatypes Klatt and KlattEditor in honour of Dennis Klatt.
To be delivered in october 2008.

D2 Sound-follows-mouse:

This has been implementated as the VowelEditor. The VowelEditor is 100% functional.

D5. Availability of Elementary Scientific Functions:

We have included the Gnu Scientific Library, gsl, version 1.10 and included several test scripts for checking.

D6. Improve Formant Frequency Measurements:

Implemented as the robust LPC algorithm. If model order conforms to the signal, this algorithm performs significantly better on artificial signals. On real speech where model and real speech signal often do not agree the algorith performs at least as well as the standard algorithms.

D7. Search and Replace with Regular Expressions:

The search and replace functions for the scripting language and for TextGrids have been implemented and testscripts have been made.

Part of the deliverables, general infrastructure changes and other new functionality have been implemented, as can be traced on <http://www.praat.org>. Praat versions changed from version 4.4.20 (May 3, 2006) to 5.0.24 (May 22, 2008).

2.2. Knelpunten en oplossingen

No real problems occurred. Because of teaching obligations of Weenink, delivery of D1 was postponed to October 2008. At that time we will probably have Praat version 5.2.

2.3. Voorgestelde planning voor de komende periode (May – October 2008)

After May, 1

Finish remaining part (KlattEditor).

Handleiding.

Het voortgangsverslag moet de Programmacommissie en het Bestuur in staat stellen om het project op te volgen in het licht van de beloofde resultaten (deliverables). Het is dus geen wetenschappelijk rapport over de verschillende onderdelen van het project, maar een managementrapport. Het bestaat uit een samenvatting en uit een sectie per deliverable die nog niet klaar was aan het einde van de vorige verslagperiode.

Samenvatting

- 1) Gevraagde wijzigingen van het project zijn wijzigingen met betrekking tot de inhoud of het oplevertijdstip van een deliverable. De motivatie voor een vraag tot wijziging dient te worden neergeschreven in het corresponderende stuk "vorderingen per deliverable".
- 2) De personeelsinzet kan worden beschreven aan de hand van een tabel met daarin, per partner, de totale voorziene mensmaanden en de gerealiseerde mensmaanden in de opeenvolgende periodes van 6 maanden.
- 3) Van elke publicatie dient men aan te geven of ze op de publieke respectievelijk de interne website van STEVIN geplaatst werd (of zal worden).

Vorderingen per deliverable

- 1) De vorderingen per deliverable zijn bedoeld om kort aan te geven (1) wat er tijdens de afgelopen periode gerealiseerd werd, (2) op welke problemen men eventueel gestoten is en welke oplossingen men voorstelt, en (3) welke planning men voor ogen heeft voor de komende 6 maanden.
- 2) Indien er knelpunten zijn dan dient men in de sectie "knelpunten en oplossingen" duidelijk te beschrijven wat die zijn, hoe ze zullen aangepakt worden en in welke mate de oorspronkelijke doelstellingen in het gedrang komen. Indien men bijstellingen van de oorspronkelijke doelstellingen voorstelt dan is dit de plaats om ze te motiveren en te preciseren. Dit gedeelte kan uiteraard wel technische informatie bevatten, voor zover die nodig is om de problematiek goed te kunnen begrijpen.

Proposals funded in three Calls for tender for specific HLT resources

<i>acronym</i>	<i>coordinating institute and other academic partners</i>	<i>industrial partners</i>	<i>STEVIN priorities addressed</i> <i>(subject)</i>	<i>planned duration</i>	<i>funding</i>
SPRAAK STE05038	KU Leuven (Patrick Wambacq) Radboud Universiteit Nijmegen - CLST TNO Human Factors Universiteit Twente - HMI		Speech resources (ASR)	26 mnths	€ 400.000
CORNETTO STE05039	Free University Amsterdam (Piek Vossen) Universiteit van Amsterdam KU Leuven	Irion Technologies bv	Language resources (Semantic lexicon)	24 mnths	€ 399.000
SoNaR	Radboud Universiteit Nijmegen - CLST (Nelleke Oostdijk) Antwerpen University Hogeschool Gent Leuven University Instituut voor Nederlandse Lexicografie Groningen University Tilburg University Twente University Utrecht University Universiteit van Amsterdam SPEX, Nijmegen	Polderland Logica-CMG Dutchear Nuance IRION Van Dale Lexicografie Dutch HLT Agency	Language resources speech resources (Annotated written Dutch corpus)	36 mnths	€ 836.000 (final commitment pending; preparatory 12 month start-up phase has been funded)

Project name: SPRAAK: Speech Processing, Recognition & Automatic Annotation Kit

Project number: STE05038

Reporting period: this is a translated version of the Dutch report covering the period of 01/02/2007 until 30/09/2007, extended with information from 30/09/2007 until 31/03/2008

**Participants: K.U.Leuven-ESAT,
R.U.Nijmegen-CLST,
TNO,
U.Twente-HMI**

1. Summary of the project

The availability of a speech recognition system for Dutch is mentioned as one of the essential requirements for the language and speech technology (LST) community. Indeed, researchers now are faced with the problem that no good speech recognition tool is available for their purposes or existing tools lack functionality or flexibility.

This project has two primary goals that will be accomplished within a single software framework. The first goal is to develop a highly modular toolkit for research into speech recognition algorithms. It allows researchers to focus on one particular aspect of speech recognition technology without needing to worry about the details of the other components. The second goal is to provide a state-of-the-art recogniser for Dutch with a simple interface, so that it can be used by non-specialists with a minimum of programming requirements. Next to speech recognition, the resulting software will enable applications in related fields as well. Examples are linguistic and phonetic research where the software can be used to segment large speech databases or to provide high quality automatic transcriptions.

We choose the existing ESAT recogniser, augmented with knowledge and code from the other partners in this project, as a starting point. This code base will be transformed to meet the specified requirements. The transformation is accomplished by improving the software interfaces to make the software package more user friendly and adapted for usage in a large user community, and by providing adequate user and developer documentation written in English, so as to make it easily accessible to the international LST community as well.

Next to providing a reference speech recognition platform for the Dutch speaking community, this project also encompasses knowledge transfer between the different partners, hence strengthening the ties between the Netherlands and Flanders, and between research institutions and application developers.

Summary of the work accomplished in the reporting period:

In the reporting period, work was mainly devoted to code conversion, development of the demo recognizers, writing of documentation, and building tools (BNF compiler and text normalization tools). A first version of the recognizer was completed and distributed between the project partners.

1.1. Overview deliverables (+time of deliverable) according to the proposal

The table below shows a list of deliverables with their dates. The numbering corresponds to the table in appendix 6 of the consortium agreement. The table below reflects the situation on 31/03/2007.

Deliverable	original date of delivery	status
WP0.1 project management and reporting	continuous	running
WP1.1 specifications of the code conversion	30/09/06	finished*
WP1.2 User level API specifications	30/09/06	finished*
WP1.3 API for the configuration of the recognizer's components	30/09/06	finished*
WP1.4 list of low-level routines, macros, types and global variables useable by higher level modules	30/09/06	running**
WP2.1 regression tests	31/07/06	running**
WP2.2 development of test suites	31/01/08	running**
WP3.1 BNF compiler	31/01/08	running
WP3.2 tools for text normalization	31/01/08	running
WP3.3 interface to the Python scripting language	31/07/07	finished*
WP4.1 users manual, version 1	31/07/07	running**

WP4.1 users manual, final version	31/01/08	running**
WP4.2 developers manual, version 1	31/07/07	running**
WP4.2 developers manual, final version	31/01/08	running**
WP5.1 demo recognizer for read speech	31/12/07	running
WP5.2 demo recognizer for telephone speech	31/12/07	running
WP6.1 licensing conditions and contracts	31/08/06	running
WP6.2 dissemination	continuous	running

* the work is finished in principle, but in the course of the project, modifications or additions can happen.

** the shift of the delivery date is caused by a different view on the creation of the specified deliverable than originally planned in the project proposal. It was found much more efficient to generate the deliverable along with the code conversion process (see also below in section 2).

1.2. Completed deliverables

See the table above.

1.3. Changes requested (contents/timing of deliverables) and motivation

A more detailed explanation of the changes listed below, can be found further down in this report, in section 2.

WP1.4 list of low-level routines, macros, types and global variables useable by higher level modules: this list is continuously growing, following the amount of converted code, and it will only be completed when the code conversion is finished.

WP2.1 regression tests: the final delivery date of this deliverable is shifted to the point when all code of the new recognizer is available, since the goal of the regression tests is to test the code.

WP4.1 and WP4.2 manuals: since the documentation is extracted from the code, documentation is generated incrementally along with the code conversion progress; a final version will become available with the final version of the recognizer (May 2008).

WP6.1 licensing conditions and contracts: there is agreement on the open source model for research purposes and on the procedure to follow for commercial contracts. The texts that describe this are ready, but need some small final modifications.

1.4. Employee involvement in relation to the original plan (man months)

partner	budget	period 1 01/02/06- 31/07/06	period 2 01/08/06- 31/01/07	period 3 01/02/07- 30/09/07	period 4 01/10/07- 31/05/08	total as of 30/09/07
ESAT	38	9,05	9	15,75		33,80
RU	8	0,5	0,95	1,80		3,25
TNO	6	0,8	0,05	0,45		1,30
UT	3	0,2	0,04	1,01		1,25
total	55	10,55	10,04	19,01		39,60

Remark: the figures above show the total amount of manmonths, including those that were not paid by the project (such as performed by staff members or financed by other sources, as mentioned in the project proposal).

1.5. Dissemination of the results

- Publications

- local article in DIXIT (Dec 2006), see http://www.notas.nl/dixit/dixit_2006_december.pdf
- article submitted to Interspeech2008: *SPRAAK: an open source “SPeech Recognition and Automatic Annotation Kit”*, by Kris Demuynck, Jan Roelens, Dirk Van Compernelle, Patrick Wambacq
- **Presentations**
 - 09/06/2006: presentation for the FWO scientific community AVS (Audio Visual Systems, see <http://www.av.s.vub.ac.be/>)
 - 15/06/2006: first workshop of the SPRAAK user community in Leuven . A second workshop is planned at the end of the project.
 - 11/09/2006: Stevin day in Antwerp
 - 21/09/2007: Stevin day in Bovendonk
- **Website:** the domain [spraaak.org](http://www.spraak.org) has been acquired and the website <http://www.spraak.org/> is under construction
- **through followup projects:** new Stevin demonstration projects (NEON, HATCI, AAP, DISCO) and an IBBT-ICTRegie project (BATS) have been approved; demonstration projects are seen as an excellent dissemination channel.

1.6 Exploitation of the results

- **(Accepted) project proposals based on the present project:** the following Stevin projects will use the speech recognizer: N-Best, NEON, HATCI, AAP, DISCO; the project BATS (funded by IBBT-ICTRegie) will also use the recognizer.

2. Progress per deliverable

2.1. Activities completed in the reporting period

WP1.1 specifications of the code conversion

As mentioned in the previous report, all conventions have been established. These conventions are for the moment only available in a temporary format, i.e. links to existing documentation (eg. doxygen), or as a template (eg. C code, object interface). The majority of the conventions have been documented, only a few details need to be added.

Todo: adding the missing details

WP1.2 User level API specifications

Finished: the required functionality of user level API has been fixed (see previous report). The high level code that implements this API will be finished in the coming period.

Todo: code and documentation; since the documentation resides in the code, the formal description will be generated in synchronism with the actual implementation.

WP1.3 API for the configuration of the recognizer's components

Finished: the required functionality of the implementation level API has been fixed (see previous report). Here also, the code that implements the API, will be finished in the coming period.

Todo: code and documentation; since the documentation resides in the code, the formal description will be generated in synchronism with the actual implementation.

WP1.4 list of low-level routines, macros, types and global variables useable by higher level modules

This list is growing steadily, following the code conversion process.

WP2.1 regression tests

High level regression tests for the existing speech recognizer are available as a starting point. These tests are adapted to the new recognizer, along with the conversion process (eg. new scripts that test the training are finished only after the training itself can be done with the new system). This means that the delivery date of this deliverable is shifted to the point where the converted code is available.

WP2.2 test suites

Test suites for the existing recognition system are available. These can be reused without changes. Nevertheless, faster versions of the tests have been developed for the larger part of the training process.

WP3.1 BNF compiler

The required functionality has been established and specs were detailed. The compiler is built as a combination of three steps:

1. Perl script reads BNF files and outputs an FST that can be processed by the AT&T tools
2. determination of the FST text file in AT&T format

3. the resulting FST is mapped to an (AT&T) WFST file
Version 4 (the prefinal version) is ready. Some small changes will be incorporated in the final version, based on feedback.

WP3.2 tools for text normalization

All tools are ready. Documentation is available but is still scattered and needs to be consolidated.

WP4.1, WP4.2 users and developers manuals

In SPRAAK the documentation is extracted automatically from the code. Conventions for this process have been established. The user interface description has already been made according to these conventions; this is an essential part of the users and developers documentation. A first version of a general description of the system is ready.

WP5.1 demo recognizer for read speech

All data for the Northern Dutch version is ready and training can be started. For the Southern Dutch version, some additional data preparation work is needed.

WP5.2 demo recognizer for telephone speech

All data for both versions (Northern and Southern Dutch) is ready and training can be started.

WP6.1 licensing conditions and contracts

The consortium has reached agreement on the following model:

- SPRAAK is available freely for research purposes (at distribution cost), according to the open source model. No support is given unless SPRAAK users involve the SPRAAK developers in their projects and part of the budget is reserved for support.
- new functionality can be added, also on users' request, if financing is available, eg. through participation in projects.
- licences are available for commercial contracts. Royalties are put in a fund to be used for support and extension of the software, to the benefit of the whole SPRAAK community. This means that no complicated income distribution rules need to be agreed upon. The fund can be regarded as a remuneration of the extensive background knowledge that was put in the project.
- the fund is managed by the members of the SPRAAK consortium (initially consisting of the project partners).
- support and development on a commercial basis is possible by SPRAAK partners at commercial prices. Incomes generated by these contracts go to the involved SPRAAK partners.
- adapted or newly developed functionality that is generated by the SPRAAK community MUST end up in SPRAAK in open source.

These principles were written down in legal texts and communicated to the TST-centrale. Some final changes need to be implemented.

WP6.2 dissemination

see 1.5 above.

2.2. Problems and solutions

The only problem is an overall delay in the development. There is no particular bottleneck. An extension of two months (until 31/05/2008) has been requested (and granted). Contacts with partnerships of other projects that will use the SPRAAK software, have shown that this poses no problems for these projects.

2.3. Proposed schedule for the upcoming period

In the upcoming period, the project will be finalized: the code conversion (as described in WP2.d of the project proposal), the demo recognizers, the documentation, the BNF compiler and the text normalization tools will be finished. The work should be completed by 31/05/2008. A workshop for all users and interested third parties will be organized.

Project name: Cornetto: Combinatorial and Relational Network as Toolkit for Dutch Language Technology

Project number: STE05039

Reporting period: 1 October 2007 – 31 March 2008

Participants: Vrije Universiteit, Amsterdam, Faculteit der Letteren (VU-AMS)
Universiteit van Amsterdam, Instituut voor Informatica (UVA-AMS)
K.U. Leuven- ICRI (KU-LEU)
Irion Technologies B.V. (IRION)

Persons involved:

Piek Vossen (VU-AMS/IRION)
Willy Martin (VU-AMS)
Hennie van der Vliet (VU-AMS)
Isa Maks (VU-AMS)
Roxane Segers (VU-AMS)
Maarten de Rijke (UVA-AMS)
Erik Tjong Kim Sang (UVA-AMS)
Katja Hofmann (UVA-AMS)
Marie-Francine Moens (KU-LEU)
Erik Boiy (KU-LEU)
Hetty van Zutphen (IRION)
Agata Cybulska (IRION)

1 Summary

1.1 Deliverables

The next table lists the deliverables with related work packages, the date of delivery and the status. The numbering is in accordance with appendix-4 of the consortium agreement.

Table 1: Deliverables

Id	Wp	Title	delivery date	status	Mo.
	0	project management en rapportering	continuing	in progress	
D01	1	Cornetto Database	30/09/06	done	6
D02	2	Aligned RBN, DWN and Wordnet2.x	30/09/06	done	6
D03	3	Top-level ontology, relation constraints and assignments	31/12/06	done	9
D04	4	Core Cornetto	30/07/07	done	15
D05	4	Extended Cornetto	31/01/08	in progress	22
D06	5	Acquired concepts	30/03/07	done	12
D07	5	Acquired relations	30/07/07	done	15
D08	5	Cornetto Acquisition toolkit	30/07/07	done	15
D09	6	Ontology for the legal and finance domain	30/03/07	in progress	12
D10	6	Sub-database tuned, trimmed and extended to the legal and finance domain	31/01/08	in progress	22
D11	6	Domain tuning toolkit	31/01/08	in progress	22
D12	7	Evaluation report of the ontology advisory-board	30/03/07	WORKSHOP	12
D13	7	Task-based evaluation of Core Cornetto	30/09/07	in progress	17
D14	7	Task-based evaluation of Extended Cornetto	31/01/08	in progress	22
D15	7	Evaluation report of the user-group on the legal and finance sub-database	31/01/08	WORKSHOP	22
D16	8	Final Cornetto database and software	30/03/08	in progress	24

1.2 Finished deliverables

DEL	WP	Title
D01	1	Cornetto Database
D02	2	Aligned RBN, DWN and Wordnet2.x
D03	3	Top-level ontology, relation constraints and assignments
D04	4	Core Cornetto
D06	5	Acquired concepts
D07	5	Acquired relations
D08	5	Cornetto Acquisition toolkit

1.3 Changes

The editing work will continue until June 2008. The task-based evaluation will take place in June and as well as the completion of documentation and packaging of the results. External evaluation was not budgeted in the project but we will see if we can ask Paola Monachesi to carry out the evaluation for a small fee. Her students already worked with Cornetto.

Instead of an evaluation report by the ontology advisory board (D12) and the user-group (D15), we decided to organize another workshop in September to which we will invite both groups and other interested parties. The previous workshop was very successful and it is a better way to present the results and discuss it then a technical report based on intermediate results.

1.4 Spent effort compared to planned effort

partner	Total person months whole project	Nr. person months spent for period 3	Nr. person months spent in total
VU-AMS	15 (18)*	3 (5.4)	15(21.4)+
UVA-AMS	15	0.8(0)	15(17.9)+
KU-LEU	14	5 (5.2)	10.53 (10.53)
IRION	13 (10)*	3,5 (2.8)	13(16.8)+
Totaal	57		

Comment: the straight number indicate the paid effort in person months, the numbers between brackets give the actual person months including the work exceeding the paid effort.

+ Overspending is based on the overhead of the organization and voluntary contributions.

* Due to the budget shift from VUA to Irion Technologies, the totals have been adjusted.

1.5 Dissemination of the results Cornetto

Papers

Period 4

A. Horák, I. Maks, A. Rambousek, R. Segers, H. van der Vliet, P. Vossen (fc) "[Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology](#)", in: Proceedings of the 18th International Congress of Linguists ([CIL18](#)), Seoul, Republic of Korea, July 21-26, 2008.

A. Pease, C. Fellbaum, P. Vossen, (fc) "[Building the Global WordNet Grid](#)", in: Proceedings of the 18th International Congress of Linguists ([CIL18](#)), Seoul, Republic of Korea, July 21-26, 2008.

Vossen P., Fellbaum C. (fc) "Universals and Idiosyncracies in Multilingual WordNets", in: Handbook Multilingual Lexicography, [Oxford University Press](#), 2008

Vossen P. (fc) "WordNet: principles, developments and applications", in: Dictionaries. An International Encyclopedia of Lexicography. Volume: Recent developments with special focus on computational lexicography, Walter/Mouton de Gruyter, [Handbooks of Linguistics and Communication Science](#) (HSK), Berlin, 2008

Vossen P., Maks I., Segers R., VanderVliet H. (fc) "[Integrating lexical units, synsets and ontology in the Cornetto Database](#)", in: Proceedings of [LREC 2008](#), Marrakech, Morocco, May 28-30 May 2008.

Maks I., P. Vossen, Segers R., VanderVliet H., van Zutphen H. (fc) "Encoding adjectives in the Dutch semantic lexical database Cornetto", in: Proceedings of [LREC 2008](#), Marrakech, Morocco, May 28-30 May 2008.

Erik Tjong Kim Sang and Katja Hofmann, Automatic Extraction of Dutch, Hypernym-Hyponym Pairs. In Proceedings of CLIN-2006, Leuven, Belgium, 2007.

Horak A., P. Vossen, A. Rambousek "[A Distributed Database System for Developing Ontological and Lexical Resources in Harmony](#)", in: Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics: [CICLing 2008](#), February 17-23, 2008, Haifa, Israel. Also to be published in the [Lecture Notes on Computational Linguistics and Intelligent Text Processing in Lectures Notes in Computer Science](#), Volume 4919/2008, ISBN 978-3-540-78134-9, 1-15, Springer-Verlag, Berlin, 2008.

Horak A., Vossen P., Rambousek A. (2008) The Development of a Complex-Structured Lexicon based on WordNet, in: Proceedings of the Fourth International GlobalWordNet Conference - [GWC 2008](#), Szeged, Hungary, January 22-25, 2008

Vossen P., Maks I., Segers R., VanderVliet H., van Zutphen H. (2008) "The Cornetto Database: the architecture and alignment issues", in: Proceedings of the Fourth International GlobalWordNet Conference - [GWC 2008](#), Szeged, Hungary, January 22-25, 2008

Period 3

Vossen P. and C. Fellbaum, 2007 (fc) "Universals and Idiosyncracies in Multilingual WordNets", In: Handbook Multilingual Lexicography, [Oxford University Press](#), 2007

Period 2

Fellbaum C. and P. Vossen, 2007 "Connecting the Universal to the Specific: Towards the Global Grid", In: Proceedings of [The First International Workshop on Intercultural Collaboration](#) (IWIC 2007), Kyoto, Japan, January 25-26, 2007 Publically available on external Website

Vossen, P. (2006). *Cornetto: Een lexicaal-semantische database voor taaltechnologie*, Dixit Special Issue, Stevin. ([PDF](#)) Publically available on external Website

Vossen, P. (2006). *Een communicatief lexicon door het verankeren van woordbetekenissen*, Oratie. ([PDF](#)) Publically available on external Website

Posters

Period 3

Vossen, P., Hofmann, K., de Rijke, M., Tjong Kim Sang, E., and Deschacht, K. (2007). The Cornetto Database: Architecture and User-Scenarios. In DIR 2007, pp. 89-96. ([PDF](#)) Publically available on external Website

Tjong Kim Sang, E. (2007) Extracting Hypernym Pairs from the Web. In Proceedings of ACL 2007 (poster), Prague, Czech Republic.

Period 2

Hofmann, K. and Tjong Kim Sang, E. (2007). Automatic Extension of Non-English WordNets. SigIR 2007, accepted. Publically available on external Website

Talks

Period 4

Guest Lecture on [Corpusgebaseerd tekstonderzoek](#) voor het [ICT](#) Onderwijscentrum van de Vrije Universiteit Amsterdam, December 13, 2007.

Guest Lecture on Corpus-based Methods: "[Automatic term extraction from domain corpora](#)", Universiteit Nijmegen, November 26, 2007

Invited speaker on [Taal, Intelligentie en Betekenis](#)" on the Lustrum symposium [Kunstmatige Intelligentie, Ontwikkelingen en Toepassingen in Muziek, Taal en Geneeskunde](#)" van het [Genootschap Physica](#), Alkmaar, October 27, 2007

Invited Speaker on [Global Wordnet Grid](#) on the 6th International [PLAIN](#) Language Conference, Amsterdam, Netherlands, October 11-14, 2007,

Invited speaker on [Global Wordnet Grid](#) at the [Politechnika Koszalin](#), Koszalin, Poland, October 8, 2007.

Keynote Speaker on [Global WordNet](#) at the [3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics](#), Poznan, Poland, October 5-7, 2007,

Period 3

Tjong Kim Sang, E.: Cornetto presentation on the [Stevin](#) Programmadag, Nederlandse TaalUnie, Hoeven, September 21, Publically available on external Website

Vossen, P, Guest lecture: "Cornetto: Een Combinatorische Semantische Database voor het Nederlands" for the [Capita Selecta](#) of the Instituut voor Nederlandse Lexicologie ([INL](#)), Leiden, May 22, 2007. Publically available on external Website

Vossen, P, Guest lecture: Global WordNet and Global WordNet Grid at the [University of Potchefstroom](#), South Africa, March 19, 2007. Publically available on external Website

Vossen, P, Guest lecture: Cornetto and Wordnets at the [University of Potchefstroom](#), South Africa, March 20, 2007. Publically available on external Website

Vossen, P, Guest lecture: "African Languages WordNet Workshop/Training" of the [Department African Languages](#) of the University of South Africa, Unisa), Pretoria, South Africa, March 21-23, 2007.

Period 2

Tjong Kim Sang, E.: "Extracting Dutch Hypernymy Pairs from the Web". Presented at CLIN-17, 12 January 2007 in Leuven, Belgium. Publically available on external Website

Vossen, P., Maks, I., Martin, W., van der Vliet, H., Hofmann, K., van Zutphen, H.: *The Cornetto Database*. Presented at CLIN-17, 12 January 2007 in Leuven. ([Abstract](#), [PDF](#), [PowerPoint](#)) Publically available on external Website

Vossen, P.: *Een communicatief lexicon door het verankeren van woordbetekenissen*, Oratie. Presented December 22 2006 at Vrije Universiteit, Amsterdam. ([PDF](#), [PowerPoint](#)) Publically available on external Website

Vossen, P.: *Er kan meer dan men doet met de Cornetto database*. Presented November 30 2006 at the TST Themadag: de gebruiker centraal in Rotterdam. ([PDF](#), [PowerPoint](#)) Publically available on external Website

Period 1

Vossen, P.: *Cornetto*. Presented September 11 2006 at the Stevin project day in Antwerp. ([PDF](#), [PowerPoint](#)) Publically available on external Website

2 Progress per deliverable

2.1 Realized work in the reporting period

WP0: Management

In the 4th period, we had 3 project meetings :

- November, 19th, 2007, Leuven, Consortium and Stevin PC members
- January, 22th, 2008, Szeged, VUA with Masaryk University
- February, 18th, 2008, Amsterdam, Consortium

In addition, we had technical meetings between the individual partners of the project.

For the final period of the project, which lasts 3 months no new project meetings are scheduled.

WP1: Cornetto database

The development of the Cornetto database software was completed during the previous reporting period. In this period, the software was deployed on one additional server, located at VU, for editing work. The server at UvA is used for guest access and is regularly synchronized with VU. Further, minor bug fixes were carried out and user support was provided.

WP2: Aligning semantic resources

Completed

WP3: Ontologize

The ontology assignment was imported from the English wordnet and is further revised and improved for important concepts. This work is parallel to the work done for the editing (see below). Mappings imported from English are either confirmed, corrected or replaced by more

complex mappings. In the case of adjectives, we also started to develop an extension of SUMO.

WP4: Cornetto editing

The editing of the aligned RBN-DWN is carried out for different groups of problematic (e.g. polysemous, frequent) words:

Editing and alignment RBN-DWN-PWN-SUMO-WDOMAIN:

WORDS	period 3	period 4
Adjectives (frequent and / or polysemous):	50	130
Adjectives with 2 senses :	110	110
Frequent verbs:	130	425
Polysemous nouns (4 or more meanings) and nouns with meaning shift labels in RBN:	630	1000
Total	920	1665

The total of 1665 words corresponds with about **9500 lexical units and synsets**. These make up the most frequent, the most complex and most ambiguous words of Dutch. The work carried out consists of mapping RBN to DWN, creating new synsets and lexical units if necessary, adding any other data or relations, mapping the synsets to PWN, assigned a proper mapping to SUMO and a wordnet Domain label.

WP5: Acquisition toolkit

In this period we have adapted one of the online extraction toolkits to the feedback of the lexicographers. We have packaged the required project deliveries for this work package. We have also been working on a conference paper which compares two preprocessing strategies used in our acquisition work.

WP6: Domain acquisition

We have implemented association techniques from the field of data mining among which are a chi-square and a likelihood ratio for a binomial distribution. These tools allow us to detect a variety of events that are correlated such as: terms (including compounds) associated with a domain, collocations associated with a domain, idiomatic expressions, meanings of new words, domain-specific meanings of old words, and the acquisition of horizontal frame patterns.

WP7: Evaluation

Irión implemented 3 methods to implement semantic relations in the classification software:

1. Adding any content word combination (bi-grams) to the index and the classification text;
2. Only create combinations based on the Cornetto database in a window of 15 words;

3. Combined method 1 and 2;

Irion has a Dutch classification system that assigns thesaurus labels from the IPTC thesaurus (<http://www.iptc.org/pages/index.php>) to Dutch news text. We will investigate whether the above methods improve the current system on a test set that is defined for Irion's customers.

UvA will evaluate the Cornetto database in one of the following end-to-end tasks

- a) Automatic Question Answering (QA) - UvA has developed a QA system with which we are participating in international evaluation efforts. The system currently uses Dutch WordNet for question analysis. The cornetto database can be evaluated by replacing DWN in this step.
- b) Electoral Search - UvA has developed a widely-used electoral search engine verkiezingskijker.nl (<http://verkiezingskijker.nl>). The search engine is currently using manually assigned expansion terms for search categories. Cornetto can be evaluated in the context of automatically selecting such query expansion terms.

Evaluation will be co-ordinated with members of other project groups.

WP8: Distribution

Several users have shown interest in the intermediate results. One user from Princeton University has signed a license for using the intermediate result. Various students have been working with samples from the database and the database is being used by PhD students at the VUA for research in metaphors and opinions in a VICI program. There have also been requests from other project proposals (Stevin and European Union) that intend to use the Cornetto database. The database will also be used in the 7FP KYOTO that started in March 2008 at VUA.

2.2 Problems and solutions

No further problems have been reported

2.3 Suggested planning for the next period

WP3: Ontologize

The final phase of ontologization will focus on

- synsets with many equivalent mappings to English Wordnet
- concepts with high position in the hierarchy
- concepts with conflicting ontology assignments

When the manual work is done for the frequent & polysemous words and for the top concepts, we will extract all inherited SUMO labels for all synsets and check these for conflicts. The conflicts will be traced to their cause and corrected.

WP4: Cornetto editing

Continue editing of problematic cases:

Editing and alignment RBN-DWN-SUMO:

WORDS	Planning period 5	<i>period 4</i>
Adjectives (frequent and / or polysemous)	250	130
Adjectives with 2 senses :	110	110
Frequent verbs:	500	425
Polysemous nouns:	1200	1000
Total	2060	1665

We further focus the editing in the last phase on:

- concepts with many descendants (hyponymy and meronymy);
- concepts with many unmatched LUs and synsets;
- concepts with many equivalence mappings to English wordnet;
- concepts with conflicting SUMO labels and Domain labels;


Finally, we will make sure that all manually edited LUs and synsets will be part of the complete hierarchy in the sense that their hyperonyms are also manually verified. In this way we guarantee a semantic closure of part of the database.

- **Linking DWN-SUMO**

Linking of all aligned synsets (approx. **6000**) to SUMO ontology

- **Compiling guidelines** for linking adjectives, nouns, verbs to the SUMO ontology

MultiWordUnits

- Compiling guidelines for selection, editing, and alignment of MultiWordUnits
-  Aligning to DWN, linking to SUMO of a set of 100 MultiWordUnits

WP5: Acquisition toolkit

Work on the acquisition toolkit has been completed. During the remainder of the project period we will publish evaluation / results of this work package.

WP6: Domain acquisition

UvA provides command line version of toolkit with preprocessed data to Leuven and offers support in their initial experiments with the toolkit.

The next step is to perform exhaustive and systematic testing (of which the scenarios have been drafted) of the domain acquisition toolkit. We aim to publish the findings and evaluation in a computational linguistics journal.

WP7: Evaluation

When the final database is delivered in June, Irion and UvA will carry out extrinsic evaluation of the Cornetto database.

WP8: Distribution

We are currently already writing the documentation. We will also carry out evaluation of the quality of the data for different samples. The samples will be drawn from different sets of data based on part of speech, polysemy, mapping-status and revision status. On the basis of these samples, we can assign a quality label to the data records in the total database.