CLST Annual Report 2011

CLST is the “Centre for Language and Speech Technology”. CLST operates as a separate unit within the Faculty of Arts of the Radboud University Nijmegen.

CLST was founded in January 2003. Its objective is to contribute to the development of language and speech technology. CLST is active in research, application development and consultancy. Ideal projects for CLST are research projects with a link to applications. Whenever possible, CLST publishes about these projects in research journals and conference proceedings. Projects are funded and carried out within national and international research programmes, while contracts are also accepted from commercial parties.

In 2011, a new policy plan was drawn up. CLST focuses on research in and application of Language and Speech Technology (LST) in the following domains:

- Data mining & knowledge discovery
- Language learning and language teaching
- Communication in Health Care
- Forensics

Our cross-over activities are bundled in the domain:

- Resources and research infrastructure

Particular strengths of CLST are the following:

- Knowledge and expertise in speech technology, particularly in fundamental ASR technology and speech data
- Knowledge and expertise in language technology, particularly in the integration of rule-based and statistical processing of written language
- Combination of knowledge and expertise in the field of language and speech technology
- Combination of fundamental and applied research
- Strongly embedded in a eHumanities context
- Knowledge and expertise in the design, collection and validation of spoken and written language resources, and in the creation and maintenance of linguistic and lexical databases
- A firm position in a network of national and international contacts with additional expertise and knowledge
Board

Board members in 2011 were L. Boves (chair, emeritus since April) and R. van Hout (chair since April), A. van Kemenade, D. Iskra, D. Kenyon-Jackson and N. Oostdijk. N. Schröder is advisor of the Board. Final responsibility lies with the Dean of the Faculty. Executive director is H. van den Heuvel.

Projects

Below we present an overview of the research projects in which CLST was involved in 2011, sorted per research domain.

Data mining & knowledge discovery

- BATS (Speaker tracking and topic detection in audio archives)
- Google Award Project: Extracting factoids from Dutch texts
- TM4IP (Text Mining for Intellectual Property)
- MIDAS (Noise robust speech recognition)
- SCALE (Speech Communication with Adaptive Learning)

Language learning and language teaching

- DISCO (ASR-based CALL for training oral proficiency In Dutch)
- My Pronunciation Coach (Development of a pronunciation trainer for English as second language)
- TST-AAP (a demo system for pronunciation training of Dutch)
- FASOP (Corrective feedback and acquisition of syntax in oral proficiency)
- ACMoLA (A Computational Model of Language Acquisition)

Communication in Health Care

- ComPoli (LST to improve patients’ communication possibilities with a eHealth webportal)
- FeetBack (combining body sensors and persuasive messages to incite patients to healthy behavior)
- OptiFox (Automatic optimization of the settings of a cochlear implant using automatic pronunciation assessment techniques)
- PEDDS (automatic feedback on pronunciation of people with dysarthria)

Forensics

- BBfor2 (Bayesian Biometrics for forensics)

Resources, tools and research infrastructure

- AAM-LR (Demonstrator web application for annotation of audio recordings)
- Adelheid (Demonstrator web application for tagging historical Dutch texts)
- INTER-VIEWs (ASR-based access to 250 interviews with veterans & development of a web-based annotation tool)
- CLAM-Support (Supporting CLAM as interface for tools in a web application)
- DCS (Data Curation Service for CLARIN-NL)

1 See for detailed information on all projects: http://www.ru.nl/CLST
- Limburgian Spelling (a website with author environment to practice the spelling of various dialects of the Dutch province Limburg)
- SOMS2 (Second Opinion Music Recognition System – as distinguished from speech)
- SoNaR (A 500 M words text corpus of written Dutch)
- TQE (Demonstrator application for automatic quality evaluation of phonetic transcriptions in audio)
- TTNWW (Tools for Dutch LST as web services in a workflow)
- Projects for the Dutch Language Union (Nederlandse Taalunie, NTU): Masterplan LST & Overview LST resources for language education

In this overview we find projects from various funding sources:
- CLARIN-NL projects: AAM-LR, Adelheid, CLAM-S, DCS, INTER-VIEWs, TQE, TTNWW
- EU projects: BBfor2, OptiFox, SCALE
- NWO & ZonMW projects: ACMOLA, BATS, ComPoli, FASOP, PEDDS
- STEVIN projects: AAP, DISCO, MIDAS, SoNaR

CLST is the coordinating partner of two Marie Curie ITN projects, viz. BBfor2 and INSPIRE, and participates in a third one, viz. SCALE (coordinated by the University of Saarbrücken).

A number of projects were completed in 2011: INTER-VIEWs, MIDAS, Masterplan LST (NTU), LST resources for Dutch language education (NTU), SOMS2, and TQE.

**Acquisition**

Bert Cranen acquired a Marie Curie ITN project, named INSPIRE, investigating speech processes in realistic environments. INSPIRE aims to develop computational models that can accurately predict microscopic intelligibility improvements in devices for personal use. These include all kinds of telecommunication devices, hearing aids and cochlear implants, but also radios and TVs, that can be adapted to individual requirements. The consortium has 8 academic partners and one industrial partner (Philips). The project started 1 Jan. 2012 and lasts for four years. CLST is the coordinating partner.

In the Lifelong Learning Programme, Helmer Strik and Catia Cucchiarini acquired a project named GOBL – Games Online for Basic Language Learning. GOBL aims to provide youths and adults with means to improve their basic language skills with free access to materials for the development of communicative proficiency in French, Dutch and English through web-based minigames that support spoken interaction. The project has five partners, among which a company (Televic). The project started 1 Jan. 2012 and lasts for three years. CLST is the coordinating partner.

David Leeuwen takes part in a new COST action (IC1106): Integrating Biometrics and Forensics for the Digital Age. In this action parties are brought together in order to foster relationships between forensic and biometric research. The consortium has participants from 17 EU countries. The project started 1 Jan. 2012 and lasts for four years.

In the forensics domain Nelleke Oostdijk and Hans van Halteren acquired a project funded by the Dutch Ministry of Safety and Justice. In this (pilot) project, automatic procedures to assess threatening tweets are investigated and evaluated. It has a clear forensic impact.
Participation in new initiatives and awareness

At the national level CLST keeps track of new trends and developments. To this end, CLST’s director participates at Board level (Treasurer) in NOTaS (Nederlandse Organisatie voor Taal-en Spraaktechnologie), an organization in which national companies and knowledge centers cooperate to stimulate the development and application of language and speech technology in the Dutch language. Further, CLST is involved in CLARIN-NL, which is a large-scale national collaborative effort to create, coordinate and make language resources and technology available and readily usable in the Humanities and Social Sciences.\(^2\)

CLST members are also involved in European initiatives pertaining to the language resources infrastructure. Involvement is considered of some importance as the infrastructure impacts on the possibilities for research in the area of language and speech technology:

- FLaReNet, which aims at developing a common vision of the area of language resources and language technologies, and fostering a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide.\(^3\)
- Meta-Net which is a Network of Excellence dedicated to fostering the technological foundations of a multilingual European information society. META-NET is building the Multilingual Europe Technology Alliance (META). Bringing together researchers, commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders.\(^4\)

Personnel

Three temporary staff members left CLST in 2011 and found employment elsewhere. Four new employees were added to the staff, three junior researchers and one post-doc. Louis ten Bosch obtained a tenure position at the Dept. of Linguistics (0.5 fte) and at CLST (0.5 fte). At the end of 2011 the staff employed by CLST consisted of four tenure staff members, seven temporary staff members, nine PhD students, and two members with a temporary part-time posting from the Department of Linguistics, summing up to 17.0 fte. The institute’s daily management is in the hands of the director with the support of a secretary (together 0.5 fte). The Table below shows the distribution of the CLST staff (in fte) over various types of contract and financing. Student assistants are excluded from this overview.

<table>
<thead>
<tr>
<th></th>
<th>NWO-like projects</th>
<th>Other external funding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure staff</td>
<td>1.8</td>
<td>1.7</td>
</tr>
<tr>
<td>PhD students</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>Other temporary staff:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior</td>
<td>0.6</td>
<td>1.2</td>
</tr>
<tr>
<td>Postdoc</td>
<td>1.4</td>
<td>2.0</td>
</tr>
<tr>
<td>Secondeiments Dept.</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Linguistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>7.8</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Annual performance interviews were held with all staff members. These interviews serve to achieve a match between personal ambitions, job contents and career opportunities of individual employees.

CLST reserves part of its budget for training and education of its personnel. Most staff members followed one or more courses, most of which were offered by the Radboud University as part of its career policy.

\(^2\) [http://www.clarin.nl](http://www.clarin.nl), Due to the CLARIN-EU project similar initiatives have started in more European countries such as Germany and Norway. As of Feb. 29th, 2012, CLARIN is also an official ERIC.

\(^3\) [http://www.flarenet.eu](http://www.flarenet.eu)

\(^4\) [http://www.meta-net.eu](http://www.meta-net.eu)
The Faculty of Arts has appointed two professors in the field of LST. Since 1 September 2011, Antal van den Bosch obtained a full professorship in “Example based language modeling”. Together with David van Leeuwen, who was appointed as extraordinary professor in the field of “Forensic applications of Language and Speech Technology”, he fits very well in the research domains of CLST. Prof. Lou Boves formally retired in April 2011 but is still active in supervising PhDs. He also advises in proposal writing.

PhD Defenses & Publications

Two CLST projects resulted in an academic PhD defense in 2011. The PhD degree was part of the project obligations to which CLST had committed itself.


Publications by CLST employees are listed in the Appendix.

Customer Satisfaction

For most of its projects CLST is obliged to write progress reports. These progress reports follow the templates provided by the funding bodies (NWO, STEVIN, EU, CLARIN-NL). Customer satisfaction in other projects (typically with companies) is monitored on a regular basis (by personally contacting the customer e-mail or telephone). Customer responses are almost without exception very positive.

Public Relations

Every CLST research employee visited at least one international conference or workshop. Attendance of workshops is, as a rule, restricted to conferences to which the employee contributes an accepted paper. In all contributions the affiliation to CLST and the Radboud University is mentioned.

Further PR activities in 2011 were:

- Website updates now including
  - Demo of the Oral History Annotation tool from the INTER-VIEWs project (see also [http://wwwlands2.let.kun.nl/spex/annotationtooldemo/#search-by-words](http://wwwlands2.let.kun.nl/spex/annotationtooldemo/#search-by-words))
  - Videos of applications showing some of the results that were obtained in the DISCO project
- Folders and brochures (Dutch, English) for e.g. Biomedica in Eindhoven and Taal in Bedrijf in Rotterdam
- Publicity in newspapers and TV broadcasts
- Periodicals for a general or professional audience such as DIXIT

Substantial media attention was initiated around text acquisition for tweets, chats and sms in the SoNaR project. This generated news items on TV (EditieNL, TV Nijmegen) and in newspapers (Trouw, NRC-Next).

Henk van den Heuvel, Director CLST
H.vandenHeuvel@let.ru.nl
[http://www.ru.nl/CLST](http://www.ru.nl/CLST)
**Publications 2011**


Heuvel, H. van den (2011). Autonomata (Too). In J. Odijk (Ed.), *Stevin Programme, project results* (pp. 6-7).


