CLST Annual Report 2014

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, directly or indirectly, to the development of language and speech technology. CLST is active in fundamental and applied 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 from commercial parties are also acquired.

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 an 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 databases
- A firm position in a network of national and international contacts with additional expertise and knowledge

In 2015 CLST will review and update its research domains by formulating a new policy plan for the years 2016-2020.

2014 was a very successful year for CLST, highlights being the acquisition of four projects in NWO’s Creative Industry track, and two ICT-projects in Horizon 2020.
Board

Board members in 2014 were A. van den Bosch (chair), R. van Hout, D. van Leeuwen, P. Fikkert and N. Oostdijk, J. Oomen (Netherlands Institute for Sound and Vision) R. Smeulders (BrilliantBrains). N. Schröder, senior controller of the Faculty, is advisor of the Board. Final responsibility for CLST 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 2014, sorted per research domain. See for detailed information on all projects: http://www.ru.nl/CLST.

Data mining & knowledge discovery:

- Speech as metadata II: Speaker diarization for the Netherlands Institute of Sound and Vision (completed in 2014)
- Automatic detection of threatening tweets (completed in 2014)

Various new projects have recently been acquired for this domain. See next section.

Language learning and language teaching:

- ACMoLA (A Computational Model of Language Acquisition) (completed in 2014)
- DigLin (webtools for feedback on words and short sentences read by nonnatives who start reading)
- FASOP (Corrective feedback and acquisition of syntax in oral proficiency) (completed in 2014)
- GOBL (Games Online for Basic Language learning)
- IntlUni (The Challenges of the Multilingual and Multicultural Learning Space in the International University)
- LangTERRA (EC network project)
- LIN (A validated reading level tool for Dutch)

Communication in Health Care:

- CHASING (CHAllenging Speech training In Neurological patients by interactive Gaming)
- ComPoli (LST to improve patients’ communication possibilities with a eHealth webportal) (completed in 2014)
- INSPIRE (Speech processing in realistic environments, e.g. in hearing aids and cochlear implants)

Forensics:

- BBfor2 (Bayesian Biometrics for forensics) (completed in 2014)
- COST Action IC1106: Integrating Biometrics and Forensics for the Digital Age

Resources, tools and research infrastructure:

- CLAM-Support (Supporting CLAM as interface for tools in a web application) (completed in 2014)
- Nederlab (Repository of digitised Dutch texts and search interfaces)
- RemBench (demonstrator project interlinking databases centered around the life of the painter Rembrandt van Rijn) (completed in 2014)
- VALID (Curation of 5 databases with language pathology data collected in the Netherlands) (completed in 2014)
- App development in the Gravity project Language in Interaction

In this overview we find projects from various funding sources:
- CLARIN-NL projects: CLAM-S, RemBench, VALID
- EU projects: BBfor2, COST IC1106, DigLin, GOBL, INSPIRE, IntlUni, LangTERRA
- NWO, STW & ZonMW projects: ACMOLA, CHASING, ComPoli, FASOP, LIN, Nederlab
- Other: Dutch Government, Netherlands Institute for Sound and Vision. Projects: Elftal, HLT4LL, Speech as metadata, Threatening tweets

CLST is the coordinating partner of two Marie Curie ITN projects, viz. BBfor2 and INSPIRE, and of two LLP projects, viz. DigLin and GOBL.

Last but not least, CLST was engaged in contract research with companies. We have set up collaborations in language technology with Davinci Products and Revisely. For CMG and ABN-AMRO we have carried out several batches of speech transcriptions. We sold materials for grapheme-to-phoneme conversion to MV-Nordic.

The overview above shows that relatively many projects were completed in 2014. They have been succeeded by a range of new projects which were acquired in 2014 and started in 2015.

**Acquisition**

In the new ambitious NWO program Creative Industry, David van Leeuwen and Henk van den Heuvel acquired a project named FAME! (Frisian AudioMining Enterprise). In this project we will disclose 2600 hours of radio broadcasts from the Omrop Fryslân (Frisian Broadcast). The radio broadcasts contain spoken Frisian and Dutch covering the period 1950–2000. We will use speech technology for spoken document retrieval (speech to text conversion) and for speaker tracking (speaker diarization & recognition). The project duration is from 2015 to 2018 and involves a postdoc for CLST. Project partners are the Fryske Akademy, Områp Fryslan, Tresoar, and GridLine BV.

Another project was acquired in the same NWO program by Antal van den Bosch and Iris Hendrickx. The project is called DISCOSUMO and aims to develop a new discussion thread summarizer, which will allow mobile users to quickly digest the key points of a discussion thread as it has developed so far, thereby making it easier for users to contribute. As a result, the threshold for mobile users to participate in online discussions will be significantly lowered, and the general quality and understanding of these discussions may be expected to increase. The project duration is from 2015 to 2018 and involves a postdoc for CLST. Project partners are Tilburg University and Sanoma.

CLST is coordinator of both projects.

Two short term seed projects were acquired in NWO’s Creative Industry’s program. These are so called KIEM projects. The first is named SMiLe (acquired by Helmer Strik) and deals with Serious Gaming for Minority Language Elicitation with Frisian as test case. Project partner are the Fryske Akademy, Gridline, Grendel Games, and NHL. The second KIEM project was acquired by Nicoline van der Sijis. In this project a transcription app will be developed which converts names in Cyrillic automatically and consistently to equivalents in Dutch spelling. Project partner is the Meertens Institute. Both KIEM projects have a duration of 6 months.

---

In NWO’s Free Competition a project named ISLA (Idiomatic Second Language Acquisition) was acquired by Helmer Strik together with Ton Dijkstra. In this project we investigate how L2 learners acquire and use formulaic expressions. Subjects receive focused teaching to learn FEs and are tested experimentally to measure FE acquisition, comprehension, and production using the automatic speech recognition facilities developed in platforms of earlier similar projects. The project duration is from 1 Jan. 2015 to 1 Jan 2019, and involves a postdoc and a PhD student for CLST.

In the LEIT ICT-calls of the Horizon2020 framework the project TraMOOC (Translation of Massive Open Online Courses) was acquired by Antal van den Bosch. The project aims at developing high-quality translation of all types of text genre included in MOOCs (e.g. assignments, tests, presentations, lecture subtitles, blog text) from English into eleven European and BRIC languages. There are nine European project partners, the University of Berlin being coordinator. Project duration is from 1 Feb. 2015 to 1 Feb. 2018. The project involves a postdoc for CLST.

The CLARIAH project obtained funding in 2015. Antal van den Bosch will be leader of WP3. In CLARIAH CLST will have over 65 PMs in a period of four years starting mid 2015 to continue our work on the Infrastructure for the Arts and Humanities.

Finally, we have started a promising collaboration with the Faculty of Philosophy, Theology and Religious Studies in the ERC project of Corien Bary where CLST takes care of the language technology needed for the project.

**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. In addition, he is member of the ELRA Board. ELRA’s missions are to promote language resources for the Human Language Technology (HLT) sector, and to evaluate language engineering technologies. 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. In CLARIN-NL Antal van den Bosch and Nelleke Oostdijk are members of the national advisory panel. They are also member of the technical advise committee of Taalportaal and of the editorial board of CLIN. In CLARIAH Antal van den Bosch is technical coordinator of WP3 and in Nederlab, he is member of the Steering Committee as chair of the technical track. CLST is represented in the International Speech Communication Association Special Interest Group (ISCA SIG) on Speech and Language Technology in Education (SLaTE) by H. Strik (vice-president) and C. Cucchiarini. Together with Radboud into Languages CLST has started the spin-off company Novolanguage. Helmer Strik holds a part time position as Chief Scientific Officer at Novolanguage with the objective of joint research and project acquisition.

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:

- CLARIAH, which aims to design, construct, and exploit a facility for eHumanities research.
- CLARIN-ERIC\textsuperscript{10}
- 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.\textsuperscript{11}
- 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.\textsuperscript{12}
- COST IC-1106\textsuperscript{13}

CLST is member of the LDC and of ELRA, and as such an important entry for the acquisition of language resources by CLS researchers.

\textit{ICT infrastructure}

Together with the Department of Linguistics and the Department of Business Communication (Funds Antal van den Bosch), CLST set-up a powerful computing and storage infrastructure at the Faculty of Science. The equipment was expanded in 2013 and 2014 and now consists of nine 16-core computing units and over 175 TB disk storage. CLST values this infrastructure as vital substitute of our computing and storage capacities. Co-financed by CLS, we have employed an ICT-specialist and programmer for system maintenance and user support. In 2014 we saw an increase of staff members with programming skills supporting our research.

\textit{Personnel}

Four temporary staff members left CLST in 2014\textsuperscript{14} and found employment elsewhere. Four new employees were added to the staff, two junior researchers\textsuperscript{15}, one senior researcher\textsuperscript{16} and one ICT-specialist\textsuperscript{17}. During 2014 the staff employed by CLST consisted of five tenure staff members, eight temporary staff members (of which five postdocs), six PhD students, and two tenure staff members with a temporary part-time posting from the Department of Linguistics, summing up to 10.7 fte (excluding 0.3 ICT support). 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 contracts and financing. Student assistants are excluded from this overview.

<table>
<thead>
<tr>
<th></th>
<th>NWO-like projects, incl. CLARIN-NL</th>
<th>Other external funding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure staff</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>PhD students</td>
<td>1.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Other temporary staff:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Postdoc</td>
<td>1.0</td>
<td>1.3</td>
</tr>
<tr>
<td>Secondments Dept. Linguistics</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>Sum</td>
<td>4.9</td>
<td>5.8</td>
</tr>
</tbody>
</table>

\textsuperscript{10}http://www.clarin.eu/
\textsuperscript{11}http://www.flarenet.eu
\textsuperscript{12}http://www.meta-net.eu/
\textsuperscript{13}http://www.cost.eu/domains_actions/ict/Actions/IC1106
\textsuperscript{14}Steven Bodnar, Marijn Huijbregts, Bart Penning de Vries, Cezara Pastrav
\textsuperscript{15}Sara Ahmadi, Mario Ganzeboom
\textsuperscript{16}Iris Hendrickx,
\textsuperscript{17}Peter Berck
Compared to 2013 the number of persons employed by CLST remained constant.

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.

**Internships**

This year have started extra efforts in training and educating students in LST-related research and development. Apart from attending fairs with internships offerings for students, we maintain a webportal where proposals for internships are displayed\(^\text{18}\), both at Bachelor and at Master level.

**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, 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). Business customers such as MV-Nordic and Revisely articulate that they value our work and would like to extend and intensify collaboration.

**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. Where applicable the affiliation to CLST and the Radboud University is mentioned.

Further PR activities in 2014 were:

- **Website demo and video updates**\(^\text{19}\) now including:
  - LST Webservice Portal: Portal to a variety of LST demos
  - Event predictor from tweets: LaMa Events
  - Add annotations to spoken interviews: Oral History Annotation Tool (INTER-VEWs project)
  - Data search in video recordings of popular TV series: BATS
  - Online Games for Basic Language Learning: GOBL
  - Digital Literacy Instructor: DigLin
  - Spelling correction for Dutch: Valkuil
  - Spelling correction for English: Fowlt
  - ML Translation system for Frisian <-> Dutch: Oersetter
  - Learn about the spelling of Limburgian dialects and practice
  - Rembrandt Documents project - RemDoc
  - FoLiA Linguistic Annotation Tool: FLAT

- **Folders and brochures (Dutch, English) for e.g.**
  - LESSLA Conference\(^\text{20}\), Aug. 2014

- **Publicity in newspapers and radio/TV broadcasts addressing:**
  - Language technology and making daily work easier (Suzan Verberne and Iris Hendrickx, Dutch TV, 6 June)
  - Speaker identification after beheading James Fowley by IS (David van Leeuwen, Dutch TV, 24 August)
  - Completion RemBench project\(^\text{21}\) (Suzan Verberne, 6 November)

---


\(^{19}\) [http://www.ru.nl/clst/demos-clst/](http://www.ru.nl/clst/demos-clst/)

Periodicals for a general or professional audience such as DIXIT

CLST introduced an annual demo award among its employees. The idea is to develop a portfolio of demos that show the application of our technology. In 2013 the award was given to Maarten van Gompel. He proposed a demo in which an English learner of Dutch can type a text in Dutch but switch to English when s/he does not know a part of a sentence in Dutch. The demo uses autocompletion and automatic translation to help the writer to complete the Dutch text. The demo was implemented by Iris Monster. In 2014 the award was given to Wessel Stoop for his proposal for a web page combining all kinds of automatic analyses on tweets of an arbitrary twitter account (gender, age, sentiment, aggression). This demo will be implemented in 2015.

PhD Defenses & Publications

Three CLST projects resulted in an academic PhD defense in 2014:


Bergmann, C.,(2014, July 7). Computational models of early language acquisition and the role of different voices. Radboud Universiteit Nijmegen. Prom.: Prof. dr J.P.M. Fikkert, Prof. dr L.W.J. Boves & Dr L.T.B. ten Bosch. Project: ACMOLA.

In CLST projects with PhD students, supervision of these students requires a substantial time investment by faculty staff members. By way of compensation CLST can provide extra funding for PhD students to complete their theses when faculty funds are exhausted. This is decided on a case by case basis by the CLST-director. In 2012 such a contribution resulted in a successful PhD defense by Folkert de Vriend. For 2014 such a provision was made for Steve Bodnar.

Publications by CLST employees or others involved in CLST projects are listed in the Appendix.

Henk van den Heuvel, Director CLST
H.vandenHeuvel@let.ru.nl
http://www.ru.nl/CLST

---


22 See [http://cls.ru.nl/swan/](http://cls.ru.nl/swan/)
Publications 2014

Journal contributions:

- [http://asmp.eurasipjournals.com/content/2014/1/36](http://asmp.eurasipjournals.com/content/2014/1/36) (Metis)


- [http://dx.doi.org/10.1016/j.langsci.2013.08.015](http://dx.doi.org/10.1016/j.langsci.2013.08.015) (Metis)


- [http://dx.doi.org/10.1007/s10791-014-9239-6](http://dx.doi.org/10.1007/s10791-014-9239-6) (Metis)


- [http://repository.ubn.ru.nl/handle/2066/132195](http://repository.ubn.ru.nl/handle/2066/132195) (IR)
- [http://dx.doi.org/10.1075/dujal.3.2.02heu](http://dx.doi.org/10.1075/dujal.3.2.02heu) (Metis)


- [http://dx.doi.org/10.1075/dujal.3.2.03sto](http://dx.doi.org/10.1075/dujal.3.2.03sto) (Metis)


- [http://dx.doi.org/10.1177/1745691614552498](http://dx.doi.org/10.1177/1745691614552498) (Metis)

Dissertation:

- http://repository.ubn.ru.nl/bitstream/handle/2066/127847/127847pre.pdf?sequence=1


- http://repository.ubn.ru.nl/bitstream/handle/2066/127124/127124.pdf?sequence=1

Book chapters:


- http://dx.doi.org/10.1093/oxfordhb/9780199571932.013.001 (Metis)

Proceeding papers:

- http://repository.ubn.ru.nl/handle/2066/135150 (IR)


- [http://repository.ubn.ru.nl/handle/2066/131547](http://repository.ubn.ru.nl/handle/2066/131547) (IR)
- [http://anthology.aclweb.org/C/C14/C14-2016.pdf](http://anthology.aclweb.org/C/C14/C14-2016.pdf) (Metis)

- [http://repository.ubn.ru.nl/handle/2066/135115](http://repository.ubn.ru.nl/handle/2066/135115) (IR)

- [http://repository.ubn.ru.nl/handle/2066/135159](http://repository.ubn.ru.nl/handle/2066/135159) (IR)


**Popular publications:**


**Datasets:**


Software:

Gompel, M. van 2014. Colibri Core Nijmegen: LST Software.
- https://github.com/proycon/colibri-core (Metis)

- https://github.com/proycon/colibrita (Metis)

Gompel, M. van 2014. FLAT: FoLiA Linguistic Annotation Tool Nijmegen: LST Software.
- http://github.com/proycon/flat (Metis)

- http://proycon.github.io/folia (Metis)