CLST Annual Report 2015

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.

The Centre for Language and Speech Technology (CLST) was founded in January 2003. In terms of research activities CLST operates under the CLS programme. Its objective is to contribute to the development of language and speech technology (LST). CLST aims at fundamental and applied research in the field of LST. The focus is on the use and development of LST in the following research domains:

1. Data mining & knowledge discovery (in written text and speech)
2. Language learning and language teaching
3. Health communication
4. Language resources and infrastructure

For this type of research CLST engages in collaborations with other academic institutions, commercial and non-profit organizations. CLST publishes its research findings in scientific and popularising articles. CLST also offers consultancy in each of the aforementioned domains.

Particular strengths of CLST are the following:

- Knowledge and expertise in speech technology, particularly in fundamental automatics speech recognition (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 reviewed and updated its research domains by formulating a new policy plan for the years 2016-2020. This plan is available via CLST’s website at http://www.ru.nl/CLST.

A number of new projects was acquired partly from new research funding sources. Our infrastructure has been improved in terms of ICT-facilities and the introduction of a Research Support Office and the initial steps to set up a Technical Support Group for the Humanities for the Faculty of Arts at large.

In 2015 we invested in Deep Neural Learning for ASR by extending our hardware and employing additional personnel. These investments have led into significant performance improvements for the ASR of a less resourced language like Frisian and for speech of pathological speakers.
**Board**

Board members in 2015 were Antal van den Bosch (chair), P. Fikkert, R. van Hout, D. van Leeuwen (until April), and N. Oostdijk. External members are: J. Oomen (Ned. Instituut voor Beeld en Geluid) en R. Smeulders (BrilliantBrains).

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 2015, sorted per research domain. See for detailed information on all projects: [http://www.ru.nl/CLST](http://www.ru.nl/CLST).

Data mining & knowledge discovery:

- DISCOSUMO: Summarizing threads in online discussion fora
- FAME!: Frisian AudioMining Enterprise, disclosing radio broadcasts in Frisian
- TraMOOC: Translation of Massive Open Online Courses

Language learning and language teaching:

- DigLin (webtools for feedback on words and short sentences read by nonnatives who start reading) (completed in 2015)
- IntlUni (The Challenges of the Multilingual and Multicultural Learning Space in the International University) (completed in 2015)
- LangTERRA (EC network project) (completed in 2015)
- ISLA (Idiomatic Second Language Acquisition)
- SMiLE (Serious Gaming for Minority Language Elicitation)

Communication in Health Care:

- CHASING (CHAllenging Speech training In Neurological patients by interactive Gaming)
- INSPIRE (Speech processing in realistic environments, e.g. in hearing aids and cochlear implants)
- EST2 (e-Learning based Speech Training, part 2)

Resources, tools and research infrastructure:

- Nederlab (Repository of digitised Dutch texts and search interfaces)
- App development in the Gravity project Language in Interaction¹
- CARE (Curation of Regional Dialect Dictionaries)
- CLARIAH (Common Lab Research Infrastructure for the Arts and Humanities): WP2 and WP3
- FutureTDM (Shaping the future of text and data mining in Europe)
- OpenCGN (a search interface for CGN, i.e. corpus of spoken Dutch, and SoNaR, i.e. corpus of written Dutch)
- Cyrillic Transcription App
- Tools for the ERC project ‘Unraveling the Language of Perspective’

¹ [https://www.languageininteraction.nl/apps.html](https://www.languageininteraction.nl/apps.html)
In this overview we find projects from various funding sources:

- CLARIN-NL projects: CARE, OpenSoNaR-CGN
- EU projects: DigLin, FutureTDM, INSPIRE, IntlUni, LangTERRA, TraMOOC, Unraveling the Language of Perspective
- NWO & ZonMW projects: CLARIAH, EST2, ISLA, Nederlab
- NWO Creative Industry: CHASING, DISCOSUMO, FAME, SMiLE, Transcription App for Cyrillic

CLST is the coordinating partner of the Marie Curie ITN project INSPIRE, and of the LLP project DigLin at European level, and of CHASING, DISCOSUMO, FAME and ISLA at the national level.

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 ABN-AMRO we have carried out several batches of speech transcriptions. With pride we mention the official launch of NovoLanguage in April 2015, a spin-off company from CLST. NovoLanguage makes products for language learners by using innovative language and voice technology. The company's strength is the attention given to spoken fluency in their new products.

**Consultancy**

Iris Hendrickx was active as consultant and tool developer for Centro de Linguística da Universidade de Lisboa. She will continue her consultancy in 2016.

Henk van den Heuvel was consulted by the Fryske Akademy with respect to corpora and ICT policy (May 2015). Furthermore he was invited as a consultant to the Disordered SpeechBank initiative organised by Martin Ball at Linköping’s University about the collection and curation of language and speech resources in the area of speech disorders (October 2015). This consultancy will continue in 2016.

**Acquisition**

In 2015 CLST acquired two projects in the final year of CLARIN-NL. The first project was acquired by Nicoline van der Sijs and Henk van den Heuvel. Its objective was the curation of three regional dictionaries for Brabant, Limburg and Gelderland. Part of these dictionaries was only available in semi-structured PDF format. The aim was to convert these resources into structured CLARIN-compatible formats with metadata. The second project was openCGN; its objective was to provide a similar web-based search interface for the CGN corpus as had been previously built for SoNaR and to integrate both interfaces into one. The project was acquired by Nelleke Oostdijk and Henk van den Heuvel.

Antal van den Bosch acquired an EU’s Horizon 2020 project called ‘FutureTDM: Scientific Information in the Digital Age’. The grant will be used to make an inventory of the state of affairs in the field of Text and Data Mining (TDM) and answer the question as to how best to increase the use of TDM in scientific texts. TDM enables researchers from different disciplines to analyse, extract insights and knowledge, and exploit diverse, complex datasets from various digital media. The aim of the Future TDM project is to improve the uptake of TDM in research environments in the EU.

Helmer Strik obtained a grant from NRO (Nationaal Regie-orgaan Onderwijsonderzoek) for a project entitled ‘The effectiveness of ICT-based initial reading education’. The project addresses the effectiveness of a common, software-based method in primary school reading education.
Martin Reynaert and Antal van den Bosch acquired a project with the NIOD in the Archief2020 Framework. This project has the name 'Fully Automatic Archive Disclosure' and explores ways to maximize OCR and Named Entity Recognition in large archives with typically hand-written documents.

Martin Reynaert obtained a grant from the Roosevelt Study Center to set up a project plan, which included the design, development and implementation of a digital preservation strategy to ensure the production, preservation and accessibility of the Marshall Plan Digital Photograph collection in Europe.

**Memberships Participation in new initiatives and awareness**

CLST keeps track of new trends and developments both at the national and the international level. This is reflected in memberships in boards, committees and other coordinating bodies. 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 technologies². 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.³ 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 Taalportaal⁴ and of the editorial board of CLIN⁵. 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 (SLaTE⁶) by Helmer Strik (chair since 2015) and Catia Cucchiarini. Together with Radboud into Languages CLST has started the spin-off company NovoLanguage⁷. Helmer Strik holds a part time position as Chief Scientific Officer at NovoLanguage with the objective of joint research and project acquisition. Furthermore, Catia Cucchiarini is member of FWO’s (Fonds Wetenschappelijk Onderzoek) (Vlaanderen)-Expertpanel Cult1: Languages. She is member of the evaluation panel of CHIST-ERA (European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net, member of the CLARIN-EU General Assembly and member of the Corpus Advisory Board Nederlab. In 2015 she joined the Advisory Board of Deditcon.

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

**ICT & infrastructure**

Together with the Department of Linguistics and the Department of Business Communication, CLST (Funds Antal van den Bosch), CLST set-up a powerful computing and storage infrastructure at the Faculty of Science. The equipment was expanded in 2015 and now consists of two webservers, eleven 16-core computing units and over 250 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. Due to our active participation in CLARIAH our programming capacity was increased as well.

---

² [http://www.elra.info](http://www.elra.info)
³ [http://www.clarin.nl](http://www.clarin.nl). As of Feb. 29th, 2012, CLARIN is also an official ERIC
⁴ [http://www.taalportaal.org](http://www.taalportaal.org)
⁶ [http://www.sigslate.org](http://www.sigslate.org)
⁷ [http://NovoLanguage.com/](http://NovoLanguage.com/)
In December of 2015 the Faculty of Arts has created a Technical Support Group for the Humanities Research in the Faculty. Backed by a substantial structural financial support from the University’s Board of Governors this group will further endorse our ICT infrastructure.

In 2015 the Faculty has also established a Research Support Office which brings together the existing expertise in the exploration of external funding opportunities, proposal writing and project management, including support with matters concerning financing and personnel.

**Personnel**

Two temporary staff members left CLST in 2015 and found employment elsewhere. Five new employees were added to the staff, one junior researcher, three senior researchers and one ICT-specialist. Three employees had a temporary employment at CLST in 2015. In the course of 2015 the staff employed by CLST consisted of five tenure staff members, sixteen temporary staff members (of which nine postdocs), four PhD students, and four tenure staff members with a temporary part-time posting from the Department of Linguistics, summing up to 12.3 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 contracts and financing in 2015. Student assistants are excluded from this overview.

<table>
<thead>
<tr>
<th>Scientific personnel in fte</th>
<th>NWO-like projects, incl. CLARIN-NL</th>
<th>Other external funding</th>
<th>Sum fte (+nr of employees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure staff</td>
<td>1.8</td>
<td>0.8</td>
<td>2.6 (5)</td>
</tr>
<tr>
<td>PhD students</td>
<td>1.8</td>
<td>1.6</td>
<td>3.4 (4)</td>
</tr>
<tr>
<td><strong>Other temporary staff:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior</td>
<td>1.5</td>
<td>0.5</td>
<td>2.0 (7)</td>
</tr>
<tr>
<td>Postdoc</td>
<td>1.8</td>
<td>1.7</td>
<td>3.5 (9)</td>
</tr>
<tr>
<td><strong>Secondments Dept. Linguistics</strong></td>
<td>0.0</td>
<td>0.8</td>
<td>0.8 (4)</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>6.7</td>
<td>5.6</td>
<td>12.3 (29)</td>
</tr>
</tbody>
</table>

Compared to 2014 the number of persons employed by CLST and associated fte’s increased significantly, which was largely due to the H2020 and NWO Creative Industries projects that were acquired. The number of PhD students decreased from 6 to 4 and was counterbalanced by a notable increase in junior and postdoc temporary staff members (from 8 to 16). It is one of the assets of CLST that it can employ junior and postdoc researchers in short term projects.

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**

As in 2014 CLST spent 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, both at Bachelor and at Master level.

---

9 Joost van Doremalen, Vanja de Lint
9 Ferdy Hubers
10 Maria Eskevich, David van Leeuwen, Emre Yılmaz
11 Ko van der Sloot
12 Aukje Borkent, Jorik van Engeland, Pepijn Hendriks
13 http://www.ru.nl/clst/internships/clst-internships/
Maarten van Gompel won the 2015 CLARIN Young Scientist Award. This award is for young developers or researchers in recognition of their outstanding contributions to the construction of the CLARIN research infrastructure as well as its use.

In March 2015 the CLST website was updated to the new corporate style of Radboud University.

Over the year CLST members were involved in various news items in a variety of media:

Suzan Verberne was interviewed by E-Data & Research, Alles over Rembrandt digitaal doorzoekbaar (February 2015).

Henk van den Heuvel and Catia Cucchiarini were involved in the well attended theme meeting about speech technology for the visually impaired, organized by the Oogvereniging regio Nijmegen-Arnhem in collaboration with the Dutch Language Union and NOTaS. The Dutch Language Union, NOTaS, the MuZIEum and a number of external parties intend to meet again to discuss potential future collaboration. March 28, 2015.

Helmer Strik was interviewed by Mark Beekhuis (BNR Wetenschap), in which Helmer Strik explained about learning Dutch, English and French by playing mini-games incorporating automatic speech recognition that were developed and tested within the framework of the European GOBL project. This interview was broadcasted in the radio programme BNR Wetenschap. April 8, 2015.

David van Leeuwen was interviewed for Radio 1 on the subject of the introduction by Apple of the Dutch version of Siri, a speech app for iPhones and iPads, April 11.

Antal van den Bosch was interviewed by NRC Q about: Chatten in het Chinees? (i.e. Chatting in Chinese?). In principle anyone can speak a language one does not know with the help of translation software available from Microsoft and Google. He was also interviewed for NRC Next (April 14, 2015) and NRC Handelsblad (April 19, 2015).

Helmer Strik was a guest in De Taalshow, a monthly program in the Torpedo theatre (previously known as Parool theatre) in Amsterdam, where he gave a presentation and was interviewed about the use of apps and games for language learning. April 23, 2015.

Helmer Strik: To bridge the gap between science and industry both universities and companies participate in research projects within the framework of the NWO Creative Industries programme. In this context he was interviewed by I. Noordhoek (of CLICKNL games, the Dutch Games Association (DGA)) on the development and use of ‘gamified apps’ in health care. September 23, 2015.

Helmer Strik and Martijn Enter of NovoLanguage were invited to a meeting with the Belgian ambassador C. Hoornaert, the Commissaris v.d. Koning in Gelderland C. Cornielje and the President of the Executive Board of the Radboud University G. Meijer. They presented their spin-off company NovoLanguage. September 23, 2015.

Antal van den Bosch appeared in De Taalstaat on Radio1, where he spoke about language use on social media, September 25, 2015. This was a result of his presentation Nederlanders, ik weet alles van jullie.

DRONGO demo’s experiments ISLA: Catia Cucchiarini, Ferdy Hubers, and Helmer Strik: Hoe goed kom je uit de verf? (Idiomatic Second language Acquisition (ISLA)) (September, 25-26).

Catia Cucchiarini, Louis ten Bosch, and Helmer Strik presented Test your knowledge of Dutch idioms! at De Nijmeegse Taalmiddag, Lindenberg Nijmegen. October 3, 2015.

Antal van den Bosch was interviewed by Marc Hijink for an NRC and NRC Next article on autocompletion on mobile phones ("Mijn telefoon begrijpt me niet"), published on November 28, 2015.

Helmer Strik: the spin-off company NovoLangage was mentioned in an article in 'het Radboud-boekje’ From spin off to start up. November 2015.

Helmer Strik and Martijn Enter were mentioned in an article in the Magazine Mercator Novio Tech, Nr. 2, 2015, about the meeting with the Belgian ambassador C. Hoornaert, the Commissaris v.d. Koning in Gelderland C. Cornielje, and the President of the Executive Board of the Radboud University G. Meijer. December 2015.

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 text15. 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 was implemented in 2015. All demos can be found on CLST’s website16.

Finally, 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.

**PhD Defenses & Publications**

One CLST project resulted in an academic PhD defense in 2015:


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 who will defend his PhD thesis in 2016, and in 2015 for Juliane Kirsch-Schmidt.

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

---

14 [http://www.ru.nl/researchersnight/](http://www.ru.nl/researchersnight/)
15 See [http://cls.ru.nl/swan/](http://cls.ru.nl/swan/)
**Publications 2015**

**Journal contributions:**

- [http://dx.doi.org/10.1016/j.cedpsych.2014.09.001](http://dx.doi.org/10.1016/j.cedpsych.2014.09.001) (Metis)

- [http://dx.doi.org/10.1111/bjet.12320](http://dx.doi.org/10.1111/bjet.12320) (Metis)

- [http://hdl.handle.net/2066/150260](http://hdl.handle.net/2066/150260) (IR)
- [http://dx.doi.org/10.1371/journal.pone.0132245](http://dx.doi.org/10.1371/journal.pone.0132245) (Metis)

- [http://dx.doi.org/10.1080/09588221.2014.889713](http://dx.doi.org/10.1080/09588221.2014.889713) (Metis)

**Dissertation:**

Supervisor(s): Prof. R.W.N.M. van Hout, Dr Catia Cucchiarini & Dr Helmer Strik. Project: FASOP. [Online version].

**Book chapters:**


**Proceeding papers:**

- [http://hdl.handle.net/2066/145192](http://hdl.handle.net/2066/145192) (IR)


Popular publications:


Datasets:

- http://validdata.org/dataset/deaf-adults-ru/ (IR)

- http://validdata.org/dataset/sli-ru-kentalis/ (IR)


Software:

- https://github.com/proycon/python-timbl (Metis)

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

- https://proycon.github.io/colibri-core (Metis)

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

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

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

- https://proycon.github.io/LaMachine (Metis)

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

- https://github.com/proycon/python-frog (Metis)
- https://github.com/proycon/python-ucto (Metis)

- https://languagemachines.github.io/ucto (Metis)

- https://languagemachines.github.io/frog (Metis)

- https://github.com/LanguageMachines/libfolia (Metis)

- https://github.com/LanguageMachines/ticutils (Metis)

- https://github.com/LanguageMachines/toad (Metis)