CLST Annual Report 2016

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.

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/more_info/annual_reports/. An action plan for 2017 was formulated which was based on this.

In terms of finances and personnel CLST’s position has remained stable and firm. Many short-term projects were started with new partners and new funding opportunities. Midterm projects now come into focus again.

CLST’s server park was updated and extended in cooperation with the Faculty’s Humanities Lab. 2016 is characterised by an exceptionally large number of journal publications, new websites and software releases.
**Board**

Board members in 2016 were Antal van den Bosch (chair), Louis ten Bosch, Mirjam Ernestus (research director of the Centre for Language Studies), Roeland van Hout, and Nelleke Oostdijk. External members are: Johan Oomen (Ned. Instituut voor Beeld en Geluid) en Anita Cremers (TNO).

Also present at board meetings are Nico Schröder, advisor of the Board, and Henk van den Heuvel, CLST executive director. Final responsibility lies with the Dean of the Faculty of Arts.

**Project Acquisition**

2016 was characterized by the acquisition of various smaller projects granted by funders with whom CLST had not worked before.

Via Suzan Verberne CLST was attracted for a project funded by SIDN\(^1\) called ‘Patient Forum Miner’. This project is directed at the effective extraction of relevant information about diseases and treatment from internet discussion fora for cancer patients. In this project CLST partners with TNO.

In the same area of information extraction for the health domain, CLST obtained a grant from the RIVM\(^2\) to explore three subprojects: 1. Monitoring online sentiment on national vaccination programs against infectious diseases; 2. Contamination in food supplements; 3. Patient empowerment in online support communities.

With Dedicon\(^3\) two projects were acquired. The first deals with a reading app for blind and partially sighted people. CLST provides an automatic speech recogniser to significantly enhance the modalities for using this app. In the second project CLST helps with the automatic translation of image descriptions from English into Dutch, which is essential if pictures in a book must be described aloud.

In the domain of language learning and teaching, Helmer Strik received a grant from Radboud University for a study about the optimal use of an e-Learning system for training spoken English. This project is carried out together with Radboud in'to Languages\(^4\) and CLST’s spin-off company NovoLanguage\(^5\).

With Van Dale Publishers Maarten van Gompel developed a system for enriching dictionaries with data-driven synonyms and collocations using phrase tables in the Moses translation system.

As subcontractor of Telecats\(^6\) CLST acquired a project ‘Spreek2Schrijf’ for the Dutch parliament (Tweede Kamer). CLST will build an application to convert ASR output of parliamentary debates into the more formal written language used in the official reports of the debates.

Based on a proposal with CLST as main applicant CLARIN-ERIC\(^7\) funded a grant for organising a international workshop about transcribing audio recordings of interviews in the area of oral history. The transcription chain covers the digitization of analog recordings up until automatic speech recognition and manual transcription correction and metadating in the context of the Europe-wide CLARIN infrastructure.

---

1. [https://www.sidnfonds.nl/excerpt/](https://www.sidnfonds.nl/excerpt/)
3. Dedicon is a Dutch organization for supporting people with reading difficulties typically due to blindness or visual disorders.
7. [https://www.clarin.eu/](https://www.clarin.eu/)
This year’s acquisitions were short-term projects. Proposals for mid-term projects (3-4 years) will be targeted with more effort in 2017 to secure the longer-term research continuity and financing.

**Running Projects**

Below we present an overview of the research projects in which CLST was involved in 2016, 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
- Spreek2Schrijf project for the Dutch Parliament as subcontractor of Telecats

Language learning and language teaching:

- ISLA (Idiomatic Second Language Acquisition)
- MijnST (My Speech Trainer)
- SMiLE (Serious Gaming for Minority Language Elicitation) (completed in 2016)
- Van Dale Project: Enriching dictionaries with synonyms and collocations

Communication in Health Care:

- CHASING (CHAllenging Speech training In Neurological patients by interactive Gaming)
- Dedicon projects: speech enabled reading app & automatic description translation of a picture bank
- INSPIRE (Speech processing in realistic environments, e.g. in hearing aids and cochlear implants) (completed in 2016)
- EST2 (e-Learning based Speech Training, part 2) (completed in 2016)
- RIVM projects: Language technology for mining social media
- SIDN project: Patient Forum Miner

Resources, tools and research infrastructure:

- Nederlab (Repository of digitised Dutch texts and search interfaces)
- App development in the Gravity project Language in Interaction8
- CARE (Curation of Regional Dialect Dictionaries) (completed in 2016)
- 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) (completed in 2016)
- Cyrillic Transcription App (completed in 2016)
- Tools for the ERC project ‘Unraveling the Language of Perspective’
- Tools for the ERC project ‘Mediate’: open up historic libraries through advanced OCR and database connections
- Transcription Chain for Oral History Research

In this overview we find projects from various funding sources:

- CLARIN projects: CARE, OpenSoNaR-CGN, Transcription Chain
- EU projects: DigLin, FutureTDM, INSPIRE, Mediate, TraMOOC, Unraveling the Language of Perspective

8 [https://www.languageininteraction.nl/apps.html](https://www.languageininteraction.nl/apps.html)
- NWO projects: CLARIAH, EST2, ISLA, Nederlab
- NWO Creative Industry: CHASING, DISCOSUMO, FAME, SMiLE, Transcription App for Cyrillic
- Public/Private organisations: SIDN, RIVM, Telecats, Van Dale

CLST is the coordinating partner of CHASING, DISCOSUMO, FAME! and ISLA at the national level.

Consultancy

Iris Hendrickx was active as consultant and tool developer for Centro de Linguística da Universidade de Lisboa.

Henk van den Heuvel involved 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 in October 2015 and in June 2016. He is also engaged as advisor in the valorization track of the Language in Interaction program⁹.

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 the Board level (as 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, since 2016 he is President of the Board of ELRA. ELRA's missions are to promote language resources for the Human Language Technology (HLT) sector, and to evaluate language engineering technologies¹⁰. CLST was involved in CLARIN-NL and its successor CLARIAH, both of which are large-scale national collaborative efforts to create, coordinate and make language resources and technology available and readily usable in the Humanities and Social Sciences.¹¹ In CLARIAH Antal van den Bosch is technical coordinator of WP3 and in Nederlab, he was member of the Steering Committee representing the technical tracks. He is also WP7 leader in the Language in Interaction program. As of 1 January 2017 he has been appointed director of the Meertens Institute¹², and became project leader of Nederlab in that capacity. CLST is represented in the International Speech Communication Association Special Interest Group (ISCA SIG) on Speech and Language Technology in Education (SLaTE¹³) by Helmer Strik (chair since 2015) and Catia Cucchiarini. 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. She is also member of the Advisory Board of Dedicon. Louis ten Bosch is member of the editorial board of: Computer, Speech and Language. Catia Cucchiarini is member of the editorial board of: Computer Assisted Language Learning.

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

---

⁹ [https://www.languageininteraction.nl/utilization.html](https://www.languageininteraction.nl/utilization.html)
¹⁰ [http://www.elra.info](http://www.elra.info)
¹¹ [http://www.clarin.nl](http://www.clarin.nl), As of Feb. 29th, 2012, CLARIN is also an official ERIC
¹³ [http://www.sigslate.org](http://www.sigslate.org)
**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 2015 and now consists of two webservers, eleven 16-core computing units and over 250 TB disk storage. In 2016 three servers were equipped with extra RAM up to 512 GB. 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. In 2017 part of the servers will be renewed and extended with GPUs for training Deep Neural Nets.

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 further endorses 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**

One temporary staff member left CLST in 2016 and found employment elsewhere. Our secretary Hella Jooren who worked for CLST from the start in 2003 retired in 2016 together with her husband Bert Cranen. One new postdoc employee was added to the staff. Two employees had a temporary employment at CLST in 2016. In the course of 2016 the staff employed by CLST consisted of five tenure staff members, fourteen temporary staff members (of which ten postdocs), four PhD students, and three tenure staff members with a temporary part-time posting from the Department of Linguistics, summing up to 12.9 fte. The institute’s daily management is in the hands of the director (0.2 fte).

The table below shows the distribution of the CLST staff (in fte) over various types of contracts and financing in 2016. Student assistants are excluded from this overview.

<table>
<thead>
<tr>
<th>Scientific personnel</th>
<th>NWO (-like) projects, incl. CLARIAH (fte)</th>
<th>Other external funding (fte)</th>
<th>Sum fte (+nr of employees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure staff</td>
<td>1.4</td>
<td>1.6</td>
<td>3.0 (5)</td>
</tr>
<tr>
<td>PhD students</td>
<td>1.9</td>
<td>0.1</td>
<td>2.0 (4)</td>
</tr>
<tr>
<td>Other temporary staff:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior Postdoc</td>
<td>1.5</td>
<td>0.1</td>
<td>1.6 (4)</td>
</tr>
<tr>
<td></td>
<td>2.7</td>
<td>3.2</td>
<td>5.9 (10)</td>
</tr>
<tr>
<td>Secondments Dept. Linguistics</td>
<td>0.0</td>
<td>0.4</td>
<td>0.4 (3)</td>
</tr>
<tr>
<td>Sum</td>
<td>7.5</td>
<td>5.4</td>
<td>12.9 (26)</td>
</tr>
</tbody>
</table>

Compared to 2015 the number of persons employed by CLST and associated fte’s remained stable. A substantial number of CLST’s projects is short term. This is a characteristic strength of CLST, but for an optimal performance of such projects it needs the continuity of personnel with long-term appointments, Therefore CLST aims to enhance its proportion of employees with tenure contracts.

---

14 Juliane Schmidt  
15 Florian Kunneman  
16 Matje van de Camp and Sander Wubben
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**

CLST spends 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\(^\text{17}\), both at Bachelor and at Master level.

**Public Relations**

In the first half of 2016 there was large public interest in the project Sprekend Nederland\(^\text{18}\). As a result of this David van Leeuwen was interviewed in a number of radio and TV broadcasts.

Helmer Strik and Mario Ganzeboom demonstrated the CHASING game during an Open Day event at Sint Maartenskliniek (24 September) and at ‘ICT in de Zorg’.

During the DRONGO Language Festival in Utrecht several demos were given. Projects presented were DigLin, CHASING, ISLA, and a number of Twitter demo’s. During the event Catia Cucchiarini was awarded the DRONGO network prize.

OpenSoNaR+ is a search interface for the combined Dutch written reference corpus SoNaR and the Dutch spoken reference corpus CGN\(^\text{19}\). It was launched in October which was published in eData & Research\(^\text{20}\).

Another new website is Transcriptor\(^\text{21}\). This website allows to find the correct Dutch spelling of Russian (Cyrillic) names.

Florian Kunneman and Wessel Stoop launched the website www.yourewhathyoutweet.nl (4 November). With this demo users can analyze personality aspects of their Twitter messages as well as those of others.

On 8 December a new exhibition was opened in the MuZieUm. Its title is “Krijg jij het inzicht?” (Are you enlightened?) and it shows the potential of language and speech technology in various applications to support blind and partially sighted people\(^\text{22}\). NOTaS and CLST are among the contributors of the exhibition.

The website www.e-wld.nl was launched on Dember 17\(^\text{th}\) in het Limburgs Museum in Venlo. This website offers a search interface for the digitized version of the dictionary of Limburgian dialects (WLD: Woordenboek van de Limburgse Dialecten). The website was mentioned in various news articles e.g. in NRC next.

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 called Swan uses

\(^{17}\) http://www.ru.nl/clst/internships/clst-internships/
\(^{18}\) https://www.facebook.com/sprekendnederland/
\(^{19}\) See http://opensonnarplus.science.ru.nl
\(^{20}\) https://www.edata.nl/1101/pdf/1101_5.pdf
\(^{21}\) http://cls.ru.nl/transcriptor_uitgebreid/
\(^{22}\) http://taaluniversum.org/nieuws/6769/digitale_hulpmiddelen_maken_blindenwereld_zichtbaar_
autocompletion and automatic translation to help the writer to complete the Dutch text\(^2\). 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 (called You are what you tweet) was implemented in 2015 and 2016. In 2016 Eric Sanders and Wessel Stoop developed a demo to predict parliamentary election on the basis of tweets (called stemming2017.nl). All demos can be found on CLST’s website\(^2\). In 2017 a new call for demos will be issued.

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. The author’s affiliation to CLST, CLS, and Radboud University is mentioned.

**PhD Defenses & Publications**

One CLST project resulted in an academic PhD defense in 2016:


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, 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

---

\(^2\) See [http://cls.ru.nl/swan/](http://cls.ru.nl/swan/)

**Publications 2016**

**Journal contributions:**

- [http://dx.doi.org/10.1016/j.specom.2016.09.003](http://dx.doi.org/10.1016/j.specom.2016.09.003) (Metis)

- [http://dx.doi.org/10.1111/bjet.12249](http://dx.doi.org/10.1111/bjet.12249) (Metis)

- [http://dx.doi.org/10.1080/09588221.2014.927365](http://dx.doi.org/10.1080/09588221.2014.927365) (Metis)

- [http://dx.doi.org/10.1080/09588221.2016.1167090](http://dx.doi.org/10.1080/09588221.2016.1167090) (Metis)

- [http://hdl.handle.net/2066/159818](http://hdl.handle.net/2066/159818) (IR)

- [http://dx.doi.org/10.1016/j.csl.2015.12.002](http://dx.doi.org/10.1016/j.csl.2015.12.002) (Metis)


- [http://dx.doi.org/10.1016/j.bandl.2016.08.006](http://dx.doi.org/10.1016/j.bandl.2016.08.006) (Metis)

- [http://dx.doi.org/10.1017/S1351324916000036](http://dx.doi.org/10.1017/S1351324916000036) (Metis)

- [http://hdl.handle.net/2066/166367](http://hdl.handle.net/2066/166367) (IR)

- http://dx.doi.org/10.1515/iral-2016-0121 (Metis)

- http://dx.doi.org/10.1145/2873067 (Metis)

- http://dx.doi.org/10.1016/j.ins.2016.03.002 (Metis)

- http://dx.doi.org/10.1002/asi.23717 (Metis)

- http://hdl.handle.net/2066/2066/161281 (IR)
- http://hdl.handle.net/2066/2066/161281 (IR)


- http://hdl.handle.net/2066/157380 (IR)

Dissertation:

Proceedings papers:
- http://hdl.handle.net/2066/166368 (IR)

- [http://hdl.handle.net/2066/159583](http://hdl.handle.net/2066/159583) (IR)

- [http://hdl.handle.net/2066/166327](http://hdl.handle.net/2066/166327) (IR)

- [http://hdl.handle.net/2066/166327](http://hdl.handle.net/2066/166327) (IR)

- [http://hdl.handle.net/2066/156870](http://hdl.handle.net/2066/156870) (IR)

- [http://dx.doi.org/10.1007/978-3-319-44564-9_1](http://dx.doi.org/10.1007/978-3-319-44564-9_1) (Metis)

- [http://hdl.handle.net/2066/162489](http://hdl.handle.net/2066/162489) (IR)


- [http://hdl.handle.net/2066/162514](http://hdl.handle.net/2066/162514) (IR)
- [http://dx.doi.org/10.3233/978-1-61499-649-1-51](http://dx.doi.org/10.3233/978-1-61499-649-1-51) (Metis)

- [http://hdl.handle.net/2066/162443](http://hdl.handle.net/2066/162443) (IR)

- [http://hdl.handle.net/2066/160745](http://hdl.handle.net/2066/160745) (IR)
- [http://dx.doi.org/10.21437/Interspeech.2016-109](http://dx.doi.org/10.21437/Interspeech.2016-109) (Metis)
- http://hdl.handle.net/2066/162389 (IR)

- http://hdl.handle.net/2066/162513 (IR)

- http://hdl.handle.net/2066/159844 (IR)

- http://hdl.handle.net/2066/159204 (IR)

- http://dx.doi.org/10.1007/978-3-319-44564-9_29 (Metis)

- http://hdl.handle.net/2066/162505 (IR)

- http://hdl.handle.net/2066/166346 (IR)

- http://hdl.handle.net/2066/162481 (IR)

- http://hdl.handle.net/2066/162473 (IR)
  - http://dx.doi.org/10.1007/978-3-319-47874-6_18 (Metis)

  - http://hdl.handle.net/2066/158805 (IR)

  - http://hdl.handle.net/2066/158815 (IR)

  - http://hdl.handle.net/2066/158107 (IR)

  - http://dx.doi.org/10.1145/2911451.2914667 (Metis)

  - http://dl.acm.org/citation.cfm?id=2914667 (Metis)

  - http://hdl.handle.net/2066/161435 (IR)

  - http://hdl.handle.net/2066/166342 (IR)

  - http://hdl.handle.net/2066/160601 (IR)
  - http://dx.doi.org/10.21437/Interspeech.2016-109 (Metis)
- http://hdl.handle.net/2066/159848 (IR)

- http://hdl.handle.net/2066/159832 (IR)
- http://dx.doi.org/10.21437/Interspeech.2016-48 (Metis)

van Noord, R., Kunneman, F.A. & Bosch, A.P.J. van den (2016). Predicting civil unrest by categorizing Dutch Twitter events. In T. Bosse & B. Bredeweg (Eds.), Proceedings of the 28th Benelux Conference on Artificial Intelligence (pp. 72-79). Amsterdam: BNAIC
- http://hdl.handle.net/2066/162394 (IR)

- http://hdl.handle.net/2066/162432 (IR)

- http://hdl.handle.net/2066/159203 (IR)

- http://hdl.handle.net/2066/159235 (IR)
- http://dx.doi.org/10.1016/j.procs.2016.04.044 (Metis)

Popular publications:


Datasets:

- http://hdl.handle.net/2066/162244 (IR)  
- http://www.ru.nl/clst/datasets/ (Metis)

- http://hdl.handle.net/2066/157263 (IR)  
- http://dx.doi.org/10.17026/dans-28m-mwht (Metis)

Software:

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

Gompel, M. van 2016. LuigiNLP. Nijmegen: LST Software.  
- https://github.com/LanguageMachines/LuigiNLP (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/mbt (Metis)

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

Sloot, K. van der, Berck, P.J., Bosch, A.P.J. van den, Daelemans, W., Gompel, M. van, Weijters, T. & Zavrel, J. 2016. TiMBL. Nijmegen: LST Software. 
- https://github.com/LanguageMachines/timbl (Metis)

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

Websites:

- https://opensonarplus.science.ru.nl (Metis)

- http://e-wld.nl/ (Metis)

Reynaert, M., Sijs, N. van der, Camp, M. van de & Heuvel, H. van den (2016) Transcriptor
- http://cls.ru.nl/transcriptor_uitgebreid/