Online gateway to language resources

Interviews, novels, newspapers, speech recordings: language resources play an important role in the humanities and social sciences. CLARIN makes these data available to scholars.
Jan Odijk

CLARIN is a European initiative to build an infrastructure for social sciences and humanities researchers who make use of language resources. In the Netherlands, the focus of the national CLARIN project is on humanities researchers working with textual resources. CLARIN-NL will offer scholars the tools to allow computer-aided language processing with a view to addressing one or more of the multiple roles language plays in the humanities. Examples of these roles are 'carrier of cultural content and knowledge', 'instrument of communication', 'a component of identity' and 'an object of study'.

Ensuring accessibility

A key aspect of the infrastructure is that both the resources and the tools to work with them are easy to find and accessible. Moreover, the infrastructure should be tailored to the average humanities scholar, meaning that the required ICT knowledge to use the tools and resources should be moderate. To achieve this goal, resources and tools are being standardised so that they can be used together seamlessly; the tools are equipped with user-friendly interfaces.
In a range of demonstration and curation projects, humanities scholars, ICT experts and data providers collaborate to make resources available in the CLARIN infrastructure and to show the potential of certain technologies or curated data. Curation implies standardisation of resources in accordance with CLARIN requirements, addition of CMDI metadata, ensuring findability and accessibility and making provisions for long-term storage.
clarin.nl

Jan Odijk, Programme Director, and Arjan van Hessen, member Executive Board: "The infrastructure is tailored to the average humanities scholar" *photo Inge Angevaare*

MIMORE A microcomparative morphosyntactic research tool

The MIMORE tool enables researchers to investigate morphosyntactic variation in Dutch dialects by searching three related databases (DynaSAND, DiDDD and GTRP) using a common online search engine. The search results can be visualised on geographic maps and exported for statistical analysis.
With the MIMORE search engine these three databases can be searched simultaneously, using text strings, part of speech tags and syntactic variables. The researchers can combine categories and features into complex tags or use predefined tags. All categories and features are based on the ISOCAT standards. Since all sentences have a location code, the morphosyntactic phenomena found in a set of sentences resulting from a search can be automatically plotted on a geographic map. It is possible to include more than one morphosyntactic phenomenon in one map, thus visualising potential correlations between these phenomena. Also included is a user-friendly export function for external data use, e.g. in a statistical application. *Sjef Barbiers*

PilNar Pilgrim narratives

Churches are being closed and religion is moving to the margin. Paradoxically, we see religion and ritual flourishing and emerging in Europe. One example that stands out here is pilgrimage: the pilgrimage to Santiago de Compostela in particular has become unprecedentedly popular. Pilgrimage narratives, especially travel accounts, have been used as a favourite source for research into ritual and religious dynamics for a long time. In the PilNar project a corpus of modern pilgrimage narratives is constructed. It consists of Dutch texts written after ca. 2000 that present the thoughts and impressions of pilgrims to Santiago de Compostela.
The pilgrimage to Santiago is used as an example of current ritual and religious dynamics. This source has hardly, if ever, been used for contemporary research in the cultural sciences. Previous exploratory research has made it clear that the corpus of stories intended here is an excellent source for research into the profile (or, better, profiles) of the modern pilgrim. *Paul Post*

D-Lucea Database of the longitudinal utrecht collection of english accents

At University College Utrecht (UCU), students and staff speak a wide variety of native languages, but they all use English as the lingua franca on campus. How will the English accent evolve over time, the accent of English native speakers and of native speakers of other languages (Dutch, Italian, Spanish, etc.)?
To answer this question an existing database of speech recordings of L1 and L2 speakers of English is being curated. The recorded speakers are students from the UCU community. These students are being recorded longitudinally throughout their 3-year period on campus, using read and spontaneous speech in their L1 and in L2 English (or in L1 English only). The resulting database is of interest for research and development in linguistics, language education, speech technology, and sociophonetics. *Hugo Quené*

COLUMN

175 years of processing - every day

It will take our standard university server (2 Quad cores, 8 GB RAM, Linux) 175 years to process the daily batch of over 2 million news articles stored by information brokers such as Lexis Nedis. My research group at the Faculty of Arts of VU University Amsterdam is building systems for 'deep reading' of natural-language text to relate today's news to news processed in the past. What happened where and when, and who was involved? But also: who is the source, what is the opinion expressed, is it a factual statement, a denial or speculation? Ideally, our cascade of more than 15 natural-language processing modules should be able to process this batch before the next day's batch arrives.

That's why we use the HPC cloud of the SURFSara infrastructure to have batches processed in parallel by Virtual Machines (VMs) on which the full range of modules has been installed. This makes it easier to deploy more VMs if needed – as long as the infrastructure allows it. Using this configuration, we recently processed 66,000 news articles in one week. We are experimenting with parallelisation of the process and optimisation of the infrastructure at SURFSara. The system also includes a Knowledge Store (KS) for sources and processing outcomes. The KS is installed on Hadoop and Hbase and includes a triple store. For storing the Terabytes of results, we use specific storage units at SURFSara.

Our group is also heavily involved in making research results available in such a way that they can be replicated and reproduced by others. To this end, we use version-control systems such as Github, websites with releases, documentation of modules, processes and data formats, descriptions of experiments and tutorials. We try to standardise the systems and the formats against widely used practices, while at the same developing our own standards for new types of information.

Eventually, we hope to demonstrate that we can handle the news streams in different languages and provide a proper scientific platform both for developing natural language processing modules and for creating data structures for researchers to explore news streams as datasets. The latter will shed light on how many changes in the world are actually reported, how much duplication there is across sources, how much they agree or disagree about the information provided, what opinions and perspectives are provided. Our data structures should provide valuable information and knowledge about the history of the changing world as provided by a wide range of media sources. In the future, we will expand this range beyond written sources to include various multimedia sources, among them structured databases, sensors, audio-visual data, and images.

Piek Vossen
photo Riechelle van der Valk

COLOPHON Publisher: Stichting Uitgeverij E-data & Research Den Haag. **Editorial address**: Postbus 93067, 2509 AB Den Haag, +31(0)70-3494450, edata@dans.knaw.nl, edata.nl. **Editor in chief**: Erica Renckens. **Editorial secretariat**: Lucas Pasteuning. **Corrector**: Pieter Wijker. **Contributors to this issue**: Kees Aarts, Patricia Alkhoven, Inge Angevaare, Sjef Barbiers, Heidi Berkhout, Leen Breurs, Siobhan Bright, Marika de Bruijne, Carl-Christian Buhr, Sally Chambers, Marcel Das, Peter Doorn, Petra Drenth, Tom Emery, Anne Gauthier, Minna van Gerven, Valentijn Gilissen, Henk Harmsen, Hella Hollander, René van Horik, Gerbert Kraaykamp, Kees Mandemakers, Carlos Morais-Pires, Jan Odijk, Johan Oomen, Roeland Ordelman, Marieke Polhout, Paul Post, Hugo Quené, Laurent Romary, Nicoline van der Sijs, Steamwork Graphics, Heiko Tjalsma, Piek Vossen. **Layout**: Colette Sloots. **Production and press**: Quantes, Rijswijk. **Circulation**: 10.000. **ISSN**: 1872-0374. We have tried to approach all stakeholders regarding the images used. Those who believe they may assert rights can turn to us.