Frieda Steurs

The role of the Dutch Language Institute (INT) in the digital age

Abstract

The *Instituut voor de Nederlandse Taal* (or Dutch Language Institute) is the place for anyone who wants to know anything about Dutch through the centuries. The institute collects new Dutch words, updates important reference works such as the *Algemene Nederlandse Spraakkunst*, the main standard work on Dutch grammar, and creates terminology lists to make professional jargon accessible. The institute also takes a central position in the Dutch-speaking world (the Netherlands, Flanders, Suriname and the Netherlands Antilles) as a developer, keeper and distributor of corpora, lexica, dictionaries and grammars. With these sustainable language resources, all the result of scholarly methods, the Dutch Language Institute provides the necessary building blocks for the study of Dutch. In this presentation, we will focus on the structure and development of the central digital language infrastructure and plans for the near future to improve our processes using the most recent insights into computational and corpus-driven linguistics and AI.

1. The Dutch Language Institute: a treasury of Dutch language materials

In 2016, the Institute for Dutch Lexicology was turned into the more broadly oriented Dutch Language Institute (INT). This change went hand in hand with the renewed terms of reference of the General Secretariat of the Dutch Language Union, which was to focus on policy tasks, leaving the executive tasks to the INT. For the Dutch Language Institute, this transfer of tasks provided the opportunity to broaden its own activities. The institute became the central point of contact regarding the keeping and maintenance of digital language materials and the safekeeping of data collections related to any variations of Dutch. This evolution reflects the strongly altered landscape of linguistic research: large language infrastructures are digitally set up and contain corpora, dictionaries and other specialised lexicons and databases, grammar and so much more. The institute develops and provides data for dictionaries, (computational) lexicons, corpora and tools. The dictionaries are accessible online. Software and computational linguistic tools are available open source.

The INT has a central position in the whole of the Dutch-speaking world (the Netherlands, Flanders, Suriname and the Caribbean) as a developer, keeper and distributor of scholarly and sustainable language resources. The institute is well equipped for this task having a large international network for the exchange of information with like-minded institutions. The Dutch Language Institute also provides the necessary building blocks for all language applications aimed at the development and improvement of businesses and public organisations. We intend to strengthen this role in the coming years, which is why we are focusing on the sustainable distribution of any language materials, with an emphasis on:

1) Dutch vocabulary, both historical and contemporary, both in standard language and dialects, both in general language and professional language;
2) new technologies and techniques to make the internet accessible for linguistic research and for the ongoing maintenance of constantly updated, extensive corpora of contemporary Dutch;
3) a contribution to the accessibility of historical text material (coming from inside and outside the INT), in which considerable variations in spelling are no longer a search impediment and ways are offered to detect and circumvent variations in word use;
4) the use of and contribution to new computational linguistic or language technology techniques to help information retrieval from language materials;
5) the formal structuring of linguistic information, making it suitable for computational linguistic applications;
6) a further expansion of spelling information;
7) the realisation of facilities for third parties to contribute interactively to the description of the Dutch language and the optimisation of the central digital data infrastructure for this purpose;
8) becoming a point of contact for all language teachers and building an infrastructure of language materials that are useful and necessary support for teaching Dutch to various types of language learners.

2. The INT in the digital age: CLARIN services

The institute has responded to new developments in the humanities, especially in the field of digital humanities. In order to fulfill this role, the INT maintains a digital infrastructure for Dutch, paying attention to language variation (terminology, dialects, etc.). Both academic and non-academic parties can make use of this infrastructure. The INT sees a clear overlap between its own activities – the central data infrastructure – and recent developments within the e-humanities. With its own expertise, the INT contributes to the digital future of the humanities in the Netherlands and Flanders. On the one hand, knowledge and products are delivered which support other scientific organisations, and on the other hand collaboration with the e-humanities enhances the quality of the central data infrastructure for Dutch. In the next few years we will work closely together with centres for digital humanities at various universities and with networks such as the KNAW Humanities Cluster (Netherlands), Digital Humanities Benelux, and the WOG Digital Humanities (Flanders). The INT functions as a CLARIN[1] centre for Flanders and informs Flemish researchers about the latest developments in the field of linguistic sources and the wider linguistic infrastructure in Europe (CLARIN ERIC).[2] This allows any researcher to learn more about access to repositories, standards, metadata, available corpora, methods to encode their own corpus material, and storage facilities, etc. Researchers and students affiliated with universities and other research institutes can log in with single sign-on (SSO) to use tools and materials. These can be found through portals. This also makes it easy to keep track of ongoing and previously conducted research, which stimulates the cultivation of (international) contacts with fellow researchers.

The portals enable the online use or downloading of tools and data. Researchers have the option of using a personal workspace. Moreover, they can safely and sustainably leave their own research data and research tools in the infrastructure upon finishing their project. Crucially, CLARIN guarantees that tools will be updated and that materials will remain available and researchable through the use of persistent identifiers.

In 2021, we became a CLARIN K-Centre, the K standing for knowledge, focused on Dutch. In this role, the INT also shares its knowledge with non-Dutch researchers.[3] We provide extensive information about Dutch: linguistic properties, language advice, available tools and resources, etymology, and dialects, etc.

Also in 2021, we succeeded in having Belgium join the CLARIN resource network. The Belgian CLARIN consortium CLARIN-BE is led by the INT. Dr Vincent Vandeghinste, senior staff member of the INT, is the national coordinator for CLARIN-BE.

3. CLARIN + DARIAH = CLARIAH

CLARIAH[4] is a large research project in the Netherlands funded by the National Science Foundation. Researchers in the humanities joined forces and combined CLARIN with DARIAH[5] research groups and funding. CLARIAH develops, facilitates, and stimulates the use of digital humanities resources and infrastructures. We offer these resources to researchers and other professionals in an insightful and user-friendly way.

[1] CLARIN stands for Common Language Resources and Technology Infrastructure.

[2] https://www.clarin.eu.

[3] https://kdutch.ivdnt.org/wiki/K-Dutch.

[4] https://www.clariah.nl/.

[5] DARIAH stands for Digital Research Infrastructure for the Arts and Humanities.

This includes tools, particularly software applications and services aimed at digitising, annotating, analysing, and reporting research data. These tools can help researchers to:

- ✓ Perform research tasks faster, more efficiently, and more accurately;
- ✓ Search, edit, analyse, and present large amounts of data;
- ✓ Pose research questions that could not be answered before, for new scholarly insights.

Not only tools but also data sets are made available: these data sets range from handwritten seventeenth-century texts to radio and television recordings as well as social media reports on current developments. They also contain databases with structured data on historical economic parameters, linguistic phenomena, people, and locations, etc.

The work packages in CLARIAH are well distributed across different scientific disciplines and specialisms to develop its digital resources. There are teams with work packages for linguistics, socio-economic history, media studies, textual sources, and (shared) technology.

Some examples of CLARIAH projects[6] are:

NAMES: Dutch corpus of person name variants

Spelling variations, variants, and digitisation errors in person names are serious obstacles for search operations in historical documents. The NAMES project aimed to standardise 564,000 different surnames and 190,113 different given names with the help of the CLARIAH tool TICCL.

NEWSGAC: News Genres Transparent Automatic Genre Classification

How genres in newspapers and television news can be detected automatically using machine learning in a transparent manner to capture the shift from opinion-based to fact-centred reporting.

Bridging the Gap: Digital Humanities and the Arabic-Islamic Corpus

This project harnesses state-of-the-art digital humanities approaches and technologies to make pioneering forays into the vast corpus of digitised Arabic texts. This is primarily done along the lines of two case studies: Islamic jurisprudence and Arabic literature on proselytism.

[6] https://www.clariah.nl/projects.

CLARIAH Flanders

Being a Dutch-Flemish institute, the INT also participates in the CLARIAH Flanders research project funded by the Flemish Research Foundation. CLARIAH-VL is the Flemish contribution to the European research infrastructures DARIAH and CLARIN. Through its partner institutions, CLARIAH-VL helps organise a series of training events such as workshops, summer schools, and lectures. To support the free exchange of knowledge, CLARIAH-VL encourages its members and presenters to make any teaching or training events available to the general public by publishing them under open licenses and sharing them with the community (whenever they are legally allowed to do so).

4. Inclusion and diversity in the digital age

The Dutch Language Institute focuses on developing materials for the Deaf community and for language users with limited literary skills.

4.1 Working for the Deaf community: SignOn – Sign Language Translation Mobile Application and Open Communications Framework[7]

People who are deaf or hard of hearing face the challenge of interacting with others in real-life situations and are often excluded from accessing information in society. The EU-funded SignON project aims to develop a mobile application that will translate between different European sign and verbal languages. The application, lightweight software running on a standard mobile device, will interact with a cloud-based distributed framework dedicated to computationally heavy tasks. The application and framework will be designed through a co-creation approach where users will work together with the SignON researchers and engineers. The application will be easily adaptable to other languages (sign and spoken) and modalities and will ultimately promote equitable exchange of information among all European citizens.

A large part of the consortium consists of Dutch and Flemish partners, and both the Flemish and Dutch sign language and Dutch play a major part in this project.

4.2 Low literacy and language learners

The Dutch Language Institute has a corpus with data from two newspapers written especially for language learners and people with low literacy: the Wablieft news-

[7] https://cordis.europa.eu/project/id/101017255.

paper[8] (Flanders) and WAI-NOT newspaper[9] (the Netherlands). We use these materials to create new applications for language learning.

At the same time, we cooperate with Oefenen.nl[10], an online environment where adults can practise to improve their basic skills and knowledge. By creating appropriate language materials, we help them study at their own pace.

5. New developments in lexicographic insights: insights in the development of AI for NLP

We participate in the Netherlands AI Coalition (NL AIC), a public-private partnership in which the government, the business sector, educational and research institutions as well as civil society organisations collaborate to accelerate and connect AI developments and initiatives. The ambition is to position the Netherlands at the forefront of knowledge and applications of AI for prosperity and well-being. We continually do so with due observance of both Dutch and European standards and values. The NL AIC functions as the catalyst for AI applications in our country. In 2020, a **workshop on AI for Innovation** was organised by the ministries of both the Netherlands (OCW) and Flanders (EWI) .

The topics covered were:

- AI applied within research in particular on *natural language processing*;
- Smart Industry (Digital Innovation Hubs to introduce AI to companies and public services);
- AI & Legislation (Human-Centric AI);
- Data Sharing (structures and solutions for data sharing);

The Dutch Language Institute provided the input for the first action point.

6. Conclusion

Because current developments in the domains of computational linguistics, NLP, and AI are important to the Dutch Language Institute, it participates in new projects and workshops and implements these new technologies in its work on the digital language infrastructure.

[8] http://www.wablieft.be/nl/krant.

[9] https://www.wai-not.be/page/10.

[10] https://oefenen.nl/.