/instituut voor de Nederlandse taal/

Towards a Comprehensive Digital Language Infrastructure for the Dutch Language

Frieda Steurs
Head of Research/Managing Director

EFNIL
October 2023

About us

- Dutch-Flemish Institute for applied linguistics
- 34 staff members from the Netherlands and Flanders
  - 15 linguists
  - 6 computational linguists
  - 6 software developers + 1 webmaster
  - 2 system administrators
  - 4 management office
- Special relation with University of Leiden; teaching, research, servers at the ISSC in Leiden

General objectives

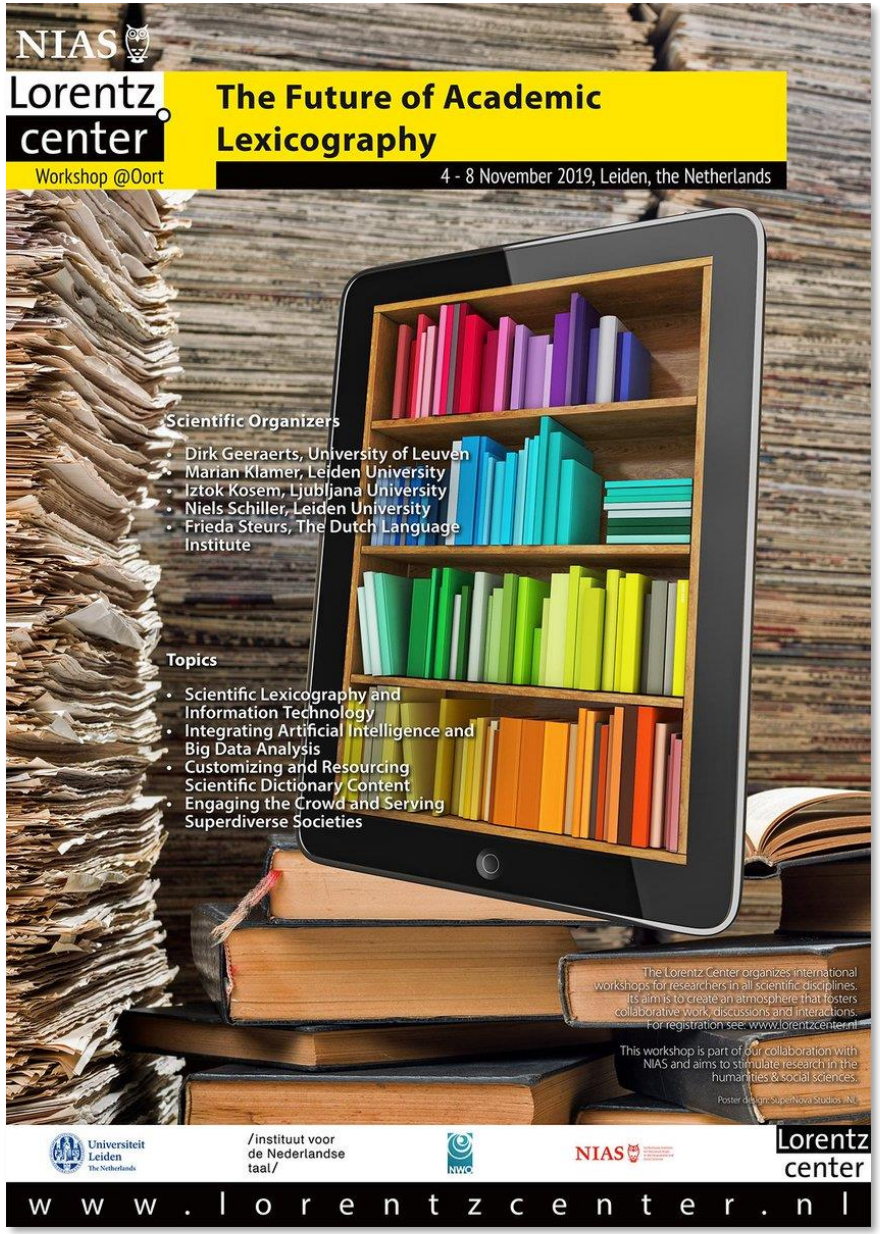
- To be a widely accessible scholarly institute in the field of the Dutch language.
- A central position in the Dutch-speaking world (the Netherlands, Flanders, Suriname and the Caribbean), as a developer, keeper and distributor of corpora, lexica, dictionaries and grammars.
- All the necessary content for the study of Dutch.

Mission

- Promoting the knowledge and use of the Dutch language through applied scientific research.
- Stimulating and coordinating the scientific description of the Dutch vocabulary and grammar in all their varieties throughout the centuries.
- Producing, integrating and disclosing Dutch source material text corpora, dictionaries, lexical digital databases, grammars, and all required technological tools.

Lorentz Workshop 2019

- New challenges after publishing the white paper
- Implementing new workflows, using AI and big data



Academic Lexicography

- Evidence Based Lexicography
- Large amounts of language data
- Challenges:
  - Role in society
  - Scalability of both the analysis and production process
  - Ability to customize and make the content accessible for a diverse audience (including as a resource for information technology developers)

Academic Lexicography

- Systematic extension and improvement of human knowledge

- Systematic, complete and definite description of the entire vocabulary of a language
- Focus on dataprocessing and knowledge creation
- Early adopter of information and communication technology

Academic Lexicography

- Corpus query systems
- Growing body of digital and digitized texts
- Relational databases
- Online publishing
- A new task of information curation
- Scholarly dictionaries from across Europe are being interconnected as Linked Open Data

Challenges for the future

Core business:

- Scientifically underpinned analysis of language use
- Systematic documentation of word meaning and use
- This word knowledge has to be useful and accessible for a wide audience
- Interconnected knowledge infrastructure

Core activities in lexicography

- Data processing
- Content creation
- Dissemination

Online search engines and language-centred applications
(dedicated software for authoring, reading, language learning etc.)

Challenges

Business model and partnerships

- Cooperation with university research groups
- Permanent body on European level (Elexis foundation)
- International cooperation (EFNIL institutes)

Scalability

Deep learning and big data analytics

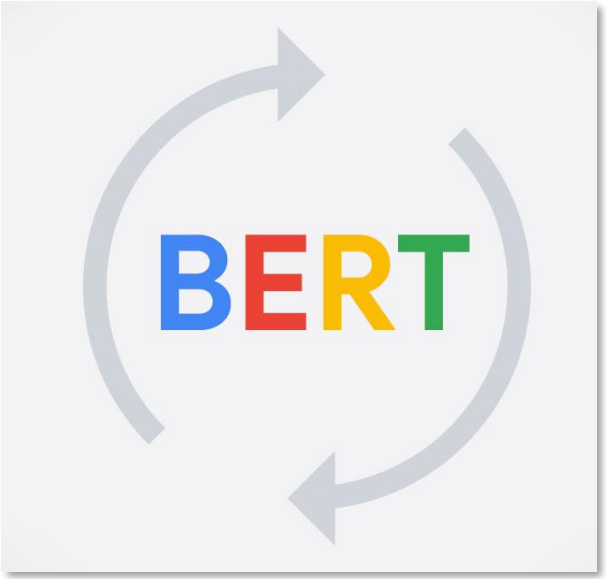
- Changing role for lexicographers
- Interactive integration of human expertise with artificial intelligence: augmented intelligence
- Automating a number of knowledge abstraction steps

Joint team of lexicographers and technologists

- Lexicographers, computational linguists and computer programmers

Research

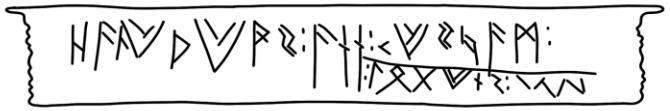
- Automated dictionary writing
- Using existing dictionaries as training data
- Pre-trained BERT embeddings for Dutch
  - Bidirectional Encoder Representations from Transformers
  - BERT embeddings encode word meaning in a continuous semantic space
  - Several sets of BERT embeddings for Dutch are already available

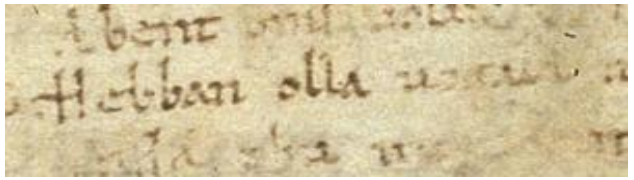


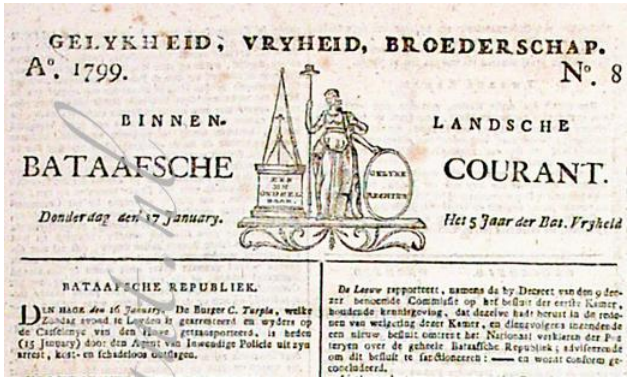
Research (2)

- Deep neural networks and word embeddings
- Create or extend existing knowledge databases
- Research and internships for students in AI
  - Open Dutch Wordnet

Lexicography as an infrastructure



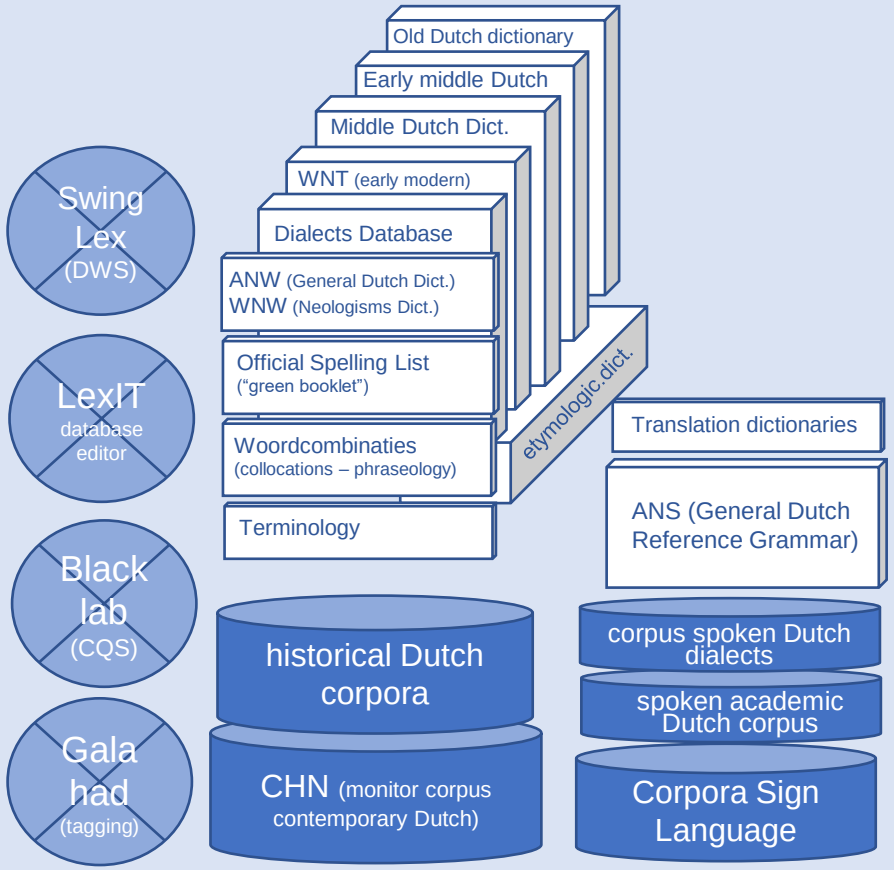




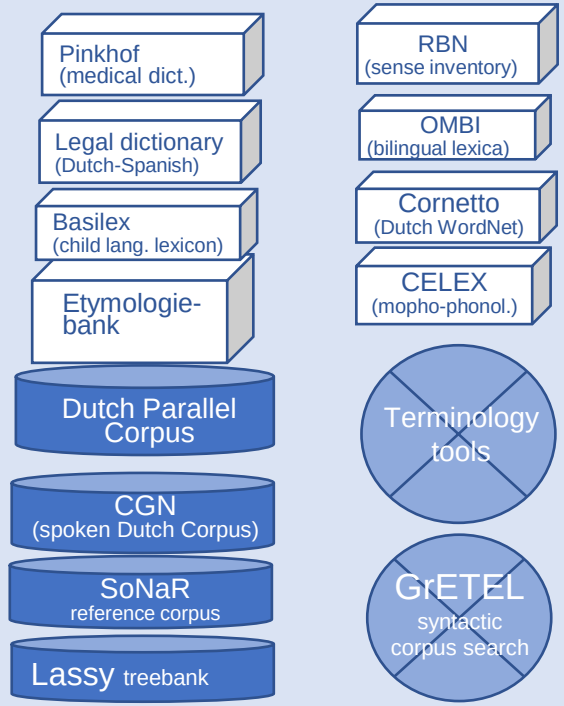
- The lexicographical process as a modular approach
- Linking all the dictionaries and databases into a central lexicon “GIGANT”
  - Computational lexicon of the Dutch Language from the 6th century up to the present.
  - A collection of words and word groups, including named entities, with every possible variant of spelling and form.

INT as language data institute

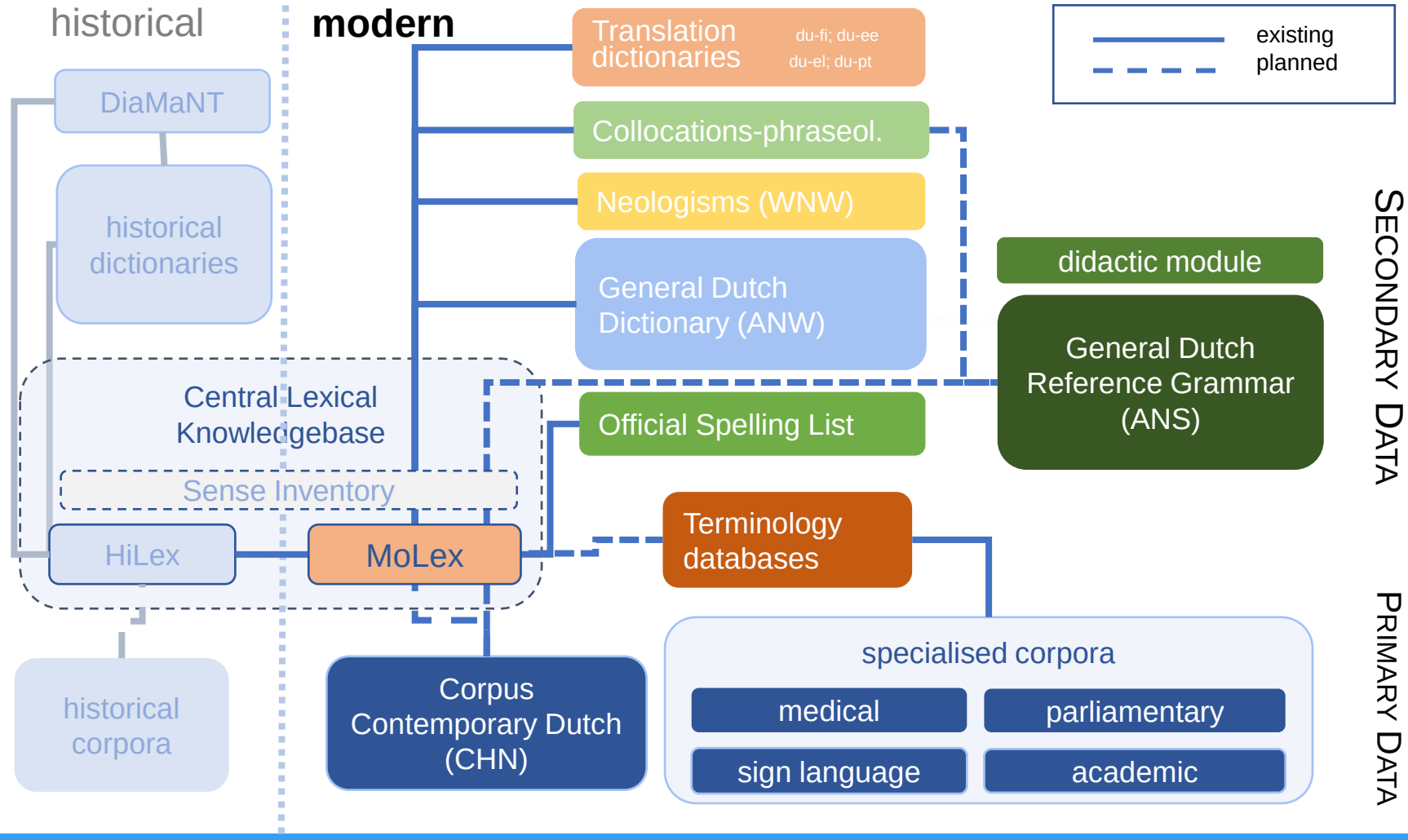
Databases and sofware **compiled** at the INT (selection)



Databases and software **deposited** at INT, **distributed** via the repositories (selection)



Linked databases of modern Dutch



European initiatives

EUROPEAN LANGUAGE EQUALITY

NL AI Coalitie

- The Netherlands AI Coalition is committed to accelerate and connect AI developments and initiatives in the Netherlands.
- The NL AIC is 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 application of AI for prosperity and well-being. We are continually doing so with due observance of both the Dutch and European standards and values. The NL AIC functions as the catalyst for AI applications in our country.

Thank you for your attention!

https://ivdnt.org/