Topic: Curation dialect dictionaries with search interface and visualisation

ID: T008 (+T009, T097)

Author: Henk van den Heuvel, Nicoline van der Sijs

What exists:

Largely curated versions of Woordenboeken Limburgse, Brabantse en Gelderse dialecten (WLD, WBD, WGD) Local dialect dictionaries: circa 50 local dialect dictionaries, curated at the Meertens Institute by interns and volunteers, supervision Nicoline van der Sijs. The collection is still being extended, the goal is to include as many old and younger dictionaries as possible, the total being somewhere around 100-120. A query interface for WLD: http://www.e-wld.nl A query interface for WBD: http://www.e-wbd.nl A similar query interface for WGD is planned for 2018 The eWND website with query interface: http://www.meertens.knaw.nl/ewnd/

What must be adapted / extended / created anew:

1. Curation of WALD (Woordenboek Achterhoekse en Liemerse dialecten and visualisation in e- WALD.nl (search portal will be externally funded)
2. Curation of other local dialect dictionaries : around 50-70
3. Developing Machine Learning tools for "Dutchification" of keywords to foster further harmonization and interoperability in search queries. The curation of the local dictionaries and the addition of Dutch equivalents (essential for search options through all data), is now done manually and takes up much time, for all Dutch entries are added manually by interns or student assistents. However, this has yielded a large amount of golden standard data on which we can develop a semi-automatical method for this task, based on machine learning translation methods like TIMBL, MOSES or NMT (Deep Learning). This must then be followed by lemmatization, for which INT can provide additional historical dictionaries of Dutch. The method must be generic, and applicable to similar cases. The problem is: how to solve unpredictable spelling variation by adding a standard Dutch equivalent. We do not foresee a method for fully automatic unsupervised keyword dutchification yet, but the process could be made much more efficient if it is reduced to manual postprocessing of unclear cases.
4. For the public and specific queries it is relevant that search portals e-wld.nl, e-wbd.nl and e-wgd.nl remain. On the other hand for research, the objective is to build a combined search interface. This will be done by INT by extending their portal for the Hercules project DSDD.
5. Add visualisation options (geomaps) to (extension of) e-WLD to illustrate dialectal variation. Develop visualization components (a.o maps) that can be deployed both in the individual dictionary websites and in the combined portal to chart dialectical variation (T097)
6. Explore accessibility of the data through Linked Open Data in collaboration with experts at Digital Humanities Lab (Marieke van Erp) and VU (Isa Maks)

Why important for CLARIAH (scientific impact):

Unique and costly collected dialectal material for Dutch Dialects will, after curation, be persistently stored and be made accessible in a search portal where the data can be explored in combination which gives new possibilities (1) for comparative research between larger dialect regions, (2) for analyzing larger word fields to study semantic relationships in a geographical perspective, and (3) for analyzing the history and distribution of specific sets of word forms. Furthermore, the dialect data with the added Dutchification can be converted to a search lexicon, which can be used for searching and retrieving dialect text corpora of older Dutch corpora (that still have many dialect traits).

Targeted/Actual users: linguists in general and dialectologists in specific

Actual use (quantify!): estimate 50

Social Impact (concrete examples):

Public exposure. See http://www.e-wld.nl which we plan to extend to include the other dictionaries. It gives users immediate feedback on local dialect words and their geographical distribution. In addition, the data can be useful in setting up the framework for new local dialect dictionaries. For the e-wld there has been publications in e.g. NRC-NEXT, De Standaard, Neerlandia We would also like to organize a workshop within CLARIAH about the research questions that could be addressed in the envisaged search portal where the dialect databases are linked.

Lead: Henk van den Heuvel, Nicoline van der Sijs, Frieda Steurs

Proposed PM estimation (try to justify): Total 13 PM

Proposed Participants + PMs:

- RU, CLST:
  - Curation of dictionaries by assistants: 3 PM
  - Technical curation and metadata: 2.5 PM
  - Developing Machine Learning tool for "Dutchification" of keywords to foster further harmonization and
  - interoperability in search queries: 3 PM
  - Add visualisation options (geomaps) to e-WLD etc to illustrate dialectal variation: 1 PM
- Meertens
  - Data hosting: pro memorie
- INT:
  - Integrating the eWND-data model : 1 PM
  - Extending data and search in Hercules DSDD-portal: 2.5 PM

Time Plan

- Data curation: 2019-2020
- Visualisation options geomaps: 2019
- Prototype Dutchification tool: end 2019
- Finalise Dutchification tool by end 2020
- Curate dialect resources with Dutchification tool and correct: 2020-2021
- Integrating e-WND model: end 2019
- Data search in DSDD-portal: end 2020
- Dialect data storage at INT: end 2019, end 2020, half 2021