Integrating Diachronous Conceptual Lexicons through Linked Open Data

Isa Maks
Fac. of Humanities, Vrije Universiteit Amsterdam
isa.maks@vu.nl

Marieke van Erp
Fac. of Humanities, Vrije Universiteit Amsterdam
marieke.vanerp@vu.nl

Piek Vossen
Fac. of Humanities, Vrije Universiteit Amsterdam
piek.vossen@vu.nl

Rinke Hoekstra
Computer Science Department, Vrije Universiteit Amsterdam
Email@email.de

Nicoline van der Sijs
Meertens Instituut
Nicoline.van.der.Sijs@meertens.knaw.nl

Diachronous conceptual lexicons and thesauri describe historical language use through time and provide historians with invaluable insights into the past. As more and more research is done in the area of the automatic processing of historical texts, the need for machine-readable historical lexicons and ontologies is growing. Up till now these lexical resources were not available in digital format, but for this project volunteers have made transcriptions of the printed books. However, the result is in text format, and information and ontologies vary for each resource which hampers simultaneous access to the information contained in them. In CLARIAH WP03, we are working on integrating and enriching several historical conceptual lexicons using linked open data principles The data are modelled according to the LMF[1] and Lemon[2] standards that are specifically tailored to represent lexical and ontological data. The model provides possibilities to take care of the lack of standardized orthography, and to model notions of time, duration and place as properties of word usages thus enabling the heterogeneous resources to interoperate.

In this demonstration, we show how we modelled different resources and converted them to RDF [3] [4]: Brouwers [4], Pland [5], Embodied Emotions [6], Hendriks [7], Meijers [8], and Open Dutch WordNet [9]. These resources are chosen because they cover the language of different time periods, and include references to a language independent semantic typology or - sometimes shallow - ontology.

An important next step was the mapping - both at word and concept level - between the different resources, aiming at mutual enrichment by combining information of the resources. Queries can be used to select data thus dynamically creating digital historic lexicons to be used in computational text

[1] http://www.lexicalmarkupframework.org/
[2] http://lemon-model.net
[3] https://www.w3.org/RDF/
[4] Brouwers, L. en F. Claes Het juiste woord, SDU, 1997 and first edition 1931
[5] http://www.meertens.knaw.nl/pland/
[6] http://access-emotionsandsenses.nl/category/embodied-emotions-2/
[7] Hendriks, Handwoordenboek van synoniemen, 1898.
[8] L. Meijers Woordenschat, 5e druk, Amsterdam 1669
[9] http://wordpress.let.vupr.nl/odwn/

analysis. Moreover, users are enabled to simultaneously query the lexicons. For example, before, a user interested in the concept `rage' would have to query each lexicon separately and aggregate the results herself. Through our alignment and mapping, a single query can be executed which retrieves all concepts from all lexicons that are related to `rage'. Figure 1 presents the model that we developed.

The lexicons will also be linked to Linked Open Data cloud. In this way we can, for example, disambiguate location names (as many of the language usage descriptions in the lexicons are bound to a location) and relate concepts mentioned in the lexicons to external taxonomies and ontologies.

The resulting resource will enable users to trace changes in word meanings and concepts over time. It will also be an invaluable resource for natural language processing of historical Dutch texts.

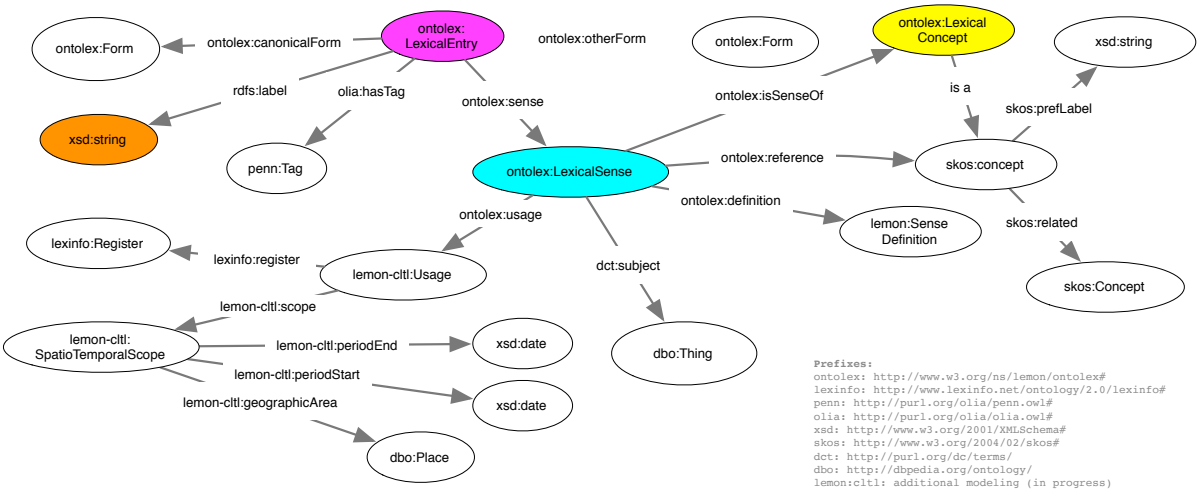

Figure 1: Model for capturing diachronous conceptual lexicons

Acknowledgements

This research was made possible by the CLARIAH-CORE project financed by NWO (http://www.clariah.nl)