The Diachronic Semantic Lexicon of Dutch as Linked Open Data

Katrien Depuydt, Jesse de Does
Instituut voor de Nederlandse Taal
Rapenburg 61, 2311GJ Leiden, The Netherlands
katrien.depuydt@ivdnt.org, jesse.dedoes@ivdnt.org

Abstract

This paper describes the Linked Open Data (LOD) model for the diachronic semantic lexicon DiaMaNT, currently under development at the Instituut voor de Nederlandse Taal (INT; Dutch Language Institute). The lexicon is part of a digital historical language infrastructure for Dutch at INT. This infrastructure, for which the core data is formed by the four major historical dictionaries of Dutch covering Dutch language from ca. 500 - ca 1976, currently consists of three modules: a dictionary portal, giving access to the historical dictionaries, a computational lexicon GiGaNT, providing information on words, their inflectional and spelling variation, and DiaMaNT, aimed at providing information on diachronic lexical variation (both semasiological and onomasiological). The DiaMaNT lexicon is built by adding a semantic layer to the word form lexicon GiGaNT, using the semantic information in the historical dictionaries. Ontolex-Lemon is a good point of departure for the LOD model, but we need extensions to be able to deal with the historical dictionary content incorporated in our lexicon.

Keywords: Linked Open Data, Ontolex, Diachronic Lexicon, Semantic Lexicon, Historical Lexicography, Language Resources

1. Background

Even though Dutch lexicography[1] can be dated back to the 13th century with the glossarium Bernense, a Latin-Middle Dutch word list, we had to wait until the 19th century for a more systematic and academic description of Dutch language. Two important dictionary projects were initiated by Matthias de Vries: a scholarly dictionary of Middle Dutch language, the *Middelnederlandsch Woordenboek* (MNW) (Dictionary of Middle Dutch) and the *Woordenboek der Nederlandsche Taal* (WNT) (Dictionary of the Dutch Language).

The MNW was compiled by E. Verwijs and J. Verdam and published between 1885 and 1929; a list of sources and a volume on dike building, water management and related terms by A. Beekman were added between 1927 and 1952. De Vries himself worked as editor-in-chief on the WNT, for which he made the design in 1852, until his demise in 1892. The first fascicle of the dictionary was published in 1864. The dictionary was finished in 1998, followed by three supplemental volumes in 2001.

Both dictionaries cover Dutch language from ca. 1250 until 1976. They were based on a corpus of quotations, written on slips of paper, and published in print. In 1995, the WNT was also released on CD-ROM, with a final release of the complete dictionary in 2003. The MNW was published on CD-ROM in 1998, accompanied by a collection of historical texts.

The former Instituut voor Nederlandse Lexicologie (Institute for Dutch Lexicology), founded in 1967 to host the WNT, decided to complete the description of historical Dutch by means of two separate projects, the *Vroegmiddelnederlands Woordenboek* (Early Middle Dutch Dictionary; 1988-1999), covering Dutch language from 1200-1300 and the *Oudnederlands Woordenboek* (Dictionary of Old Dutch, 1999-2009), covering the oldest Dutch language period from 500-1200. Both dictionaries were born digital, based on a closed corpus, in digital format. Having four scholarly dictionaries of Dutch in digital format opened up opportunities for further exploitation of the contents of these dictionaries.

1.1. Online Dictionary Portal (gtb.inl.nl)

The first step was to publish the dictionaries online in a dictionary portal (gtb.inl.nl)[2], which had its first release in 2007 (Depuydt and De Does, 2008). This application mainly supports semasiological search; most users use it to look up the meaning of a word. There is no dictionary of Dutch which describes the complete language period in the way the *Oxford English Dictionary* does for English, so combining all four dictionaries in a portal was the closest we could get to providing a diachronic lexicographic overview of Dutch language. A major challenge was to give the user optimal access to the dictionary information, without compromising the uniqueness of each individual dictionary. For this module, not only the dictionary software application was designed and built, but a lot of work went also into semi-automatic processing of the data to make the dictionary content suitable for searching. The data was converted into TEI XML. Easier access to the dictionary content was provided, among other things by adding a modern Dutch equivalent to each entry in the dictionaries. This does not only enable combined searching in several dictionaries by one single query, it also relieves users of the burden of having to search by one particular historical spelling of a lemma.

[1] For an elaborate description of the history of Dutch lexicography, see Mooijaart 2013.

[2] The first component is the online historical dictionary portal (gtb.ivdnt.org), of which the first module was released in 2007 by bringing the WNT online. In separate steps, the MNW, VMNW and ONW were processed and added and the data and application have had several updates.

1.2. GiGaNT: a Diachronic Morphosyntactic Lexicon

Also in 2007, work started on the design of the computational lexicon module GiGaNT[3] (Groot Geïntegreerd Lexicon van de Nederlandse Taal; large integrated lexicon of the Dutch language). A computational lexicon gives structured information on vocabulary and has to be suitable for use by computer software. GiGaNT provides information on words, their inflectional and spelling variation, and is aimed to cover Dutch language from the 6th century until present-day. The original aim of GiGaNT was to build a lexicon to support annotation of historical corpus material with part of speech (PoS) and lemma, so as to make these corpora better searchable. However, it can also be used to exploring new corpus material in order to harvest new material, not yet described in the available dictionaries. The lexicon has already been made available in a lexicon service, used for query expansion. A good example is the way a user gets suggested potential variants of a search word in the online historical material of the KB (Dutch Royal Library), in www.delpher.nl or in the Dutch national project (www.nederlab.nl) where a historical corpus is being compiled and put online. The lexicon is also used in Nederlab to establish the link between text material and the online historical dictionaries. Using the historical dictionaries as a primary resource for the GiGaNT lexicon was a logical thing to do. It is a very efficient way to build a historical computational lexicon. Each dictionary contains quotation material for which in each quotation, there is an occurrence of the dictionary entry in a particular form, so automatic detection of the correct word form belonging to the dictionary entry is comparatively easy.

1.3. DiaMaNT: a Diachronic Semantic Lexicon

The infrastructure as described above, offers users the means to find out the meaning of a historical word, and gives information on potential spelling and form variation, by means of which searching historical text is made easier. Having the option to search via a modern lemma form also simplifies searching in historical dictionaries and text. From the point of view of INT, another advantage is that it contributes to the structuring of the lexicographical description of the Dutch vocabulary. It gives a more systematic view on what is described, and allows easier detection of inconsistencies and gaps.

To take the infrastructure to the next level, however, would mean finding a solution to resolve one more aspect of the historical language barrier, which is not related to historical variation in form, but to historical variation in vocabulary and meaning. How can we give users the means to search in historical texts for a concept for which he or she only knows the modern Dutch term? In its most simple form, given a certain word, a user ought to get suggestions for potential synonyms of that word, combined with information on the time period in which a particular word was used. And it would even be better if we were able to allow users to look for words with a specific meaning. And can we offer historical linguists better means to study diachronic semantic variation in a systematic way?

This is why in 2015, work on the third module of the infrastructure was started, the diachronic semantic computational lexicon of Dutch (DiaMaNT, Diachroon seMantisch lexicon van de Nederlandse Taal). The main purpose of this lexicon is to enhance text accessibility and foster research in the development of concepts, by interrelating attested word forms and semantic units (concepts), and tracing semantic developments through time. In the lexicon, the diachronic onomasiology, i.e. the change in naming of concepts and the diachronic semasiology, i.e. the change in meaning of words, will be recorded in a way suitable for use by humans and computers. The onomasiological part of the lexicon is designed to enhance recall in text retrieval by providing different verbal expressions of a concept or related concepts (slager → beenhouwer, beenhakker, vleeshouwer (synonyms for 'butcher'); boer → landman ('synonyms for 'farmer'). The diachronic semasiological component (which charts semantic change), aims to enhance precision by enabling the user to take semantic change into account; the oldest meaning of apple for example is 'a fruit' (so apple is also used for pears, plums etc.). The lexicon is built by adding a semantic layer to the word form lexicon GiGaNT, using the semantic information in the historical dictionaries, i.e. the definitions from the dictionary articles from which the word form lexicon is built.

2. DiaMaNT as Linked Data

An important impulse for the deployment of DiaMaNT comes from the Dutch CLARIAH project[4]. One of the aims of the technical infrastructure of this project is to offer a generic linked open data graph, populated with entities relevant for the humanities like persons, locations and concepts, for network analysis, data annotation and linking purposes. The concept-entity graph has to provide the basis of a Dutch thesaurus for semantically related terms over time and DiaMaNT is the core of this graph. The lexicon will also be part of the CLARIAH infrastructure for linguistic resources, which enables federated search scenarios in which information from corpora, treebanks and lexica can be combined.

Publishing as Linked Open Data (LOD) facilitates this type of interoperability and integration of lexical resources (Chiarcos, 2003). The LOD paradigm provides a framework that facilitates information integration, and thus, interoperability, by ensuring that entities can be addressed in a globally unambiguous way using Unique Resource Identifiers (URIs), that entities can be accessed over HTTP, and that the descriptions of entities and links between them can be represented according to the W3C Resource Description Framework (RDF) standard (Berners-Lee, 2006).

The CLARIAH context was an important argument to go for a lexicon development strategy which would allow intermediate releases of the lexicon. So far, a project internal release has been done of the lexicon, containing synonym information extracted from the dictionary definitions. The

[3] The situation is now that two modules (based on MWN and WNT) have been released and work on the modules based on ONW and VMNW is scheduled for 2018.

[4] www.clariah.nl

basic LOD model of the lexicon has also been designed. And some exploratory research has been done into the potential distributional semantics offers for lexicon development and deployment.

2.1. DiaMaNT Source Data

The lexicon adds a semantic layer on top of the word form lexicon GiGaNT. Both DiaMaNT and GiGaNT have the historical dictionaries of Dutch as a base. The elements from the dictionaries used to create the computational lexica are: entry (historical form and modern Dutch equivalent), PoS, quotations and definitions. These elements are encoded in the TEI XML underlying the online dictionaries. The number of entries, quotations and definitions in the four dictionaries is given in table 1.

The core of the lexica is the corpus of quotations, present in the dictionaries. They illustrate the spelling, the morphological variation and the meaning of an entry as described by the lexicographers. Every quotation in the dictionaries has metadata, describing the provenance of the quotation. The quotations are dated and in all dictionaries but the WNT, also location information is provided. Each occurrence of the main structural elements in the TEI XML has its own persistent ID. In both GiGaNT and DiaMaNT, these persistent ID's are retained. In GiGaNT, the occurrences of an entry in each quotation have been detected and stored, together with the quotations and their metadata. Each word form has been given the correct analysis (lemma and main PoS). This means that in some cases, dictionary entries that in fact describe several lexical entries from the point of view of a computational lexicon, were thus split up.

Since DiaMaNT provides a semantic layer on top of GiGaNT, the word forms of GiGaNT are included in the lexicon in order to make the lexicon more suitable for text retrieval by query expansion. The aim is to develop a thesaurus (diachronic wordnet), where synonym clusters represent the concepts for which lexicalisations are described in the dictionaries. In the current prototype, a first semantic annotation layer on top of the entries and senses in the dictionaries consists of synonyms automatically extracted from the dictionary definitions from MNW and WNT. It is not yet a unified semantic resource, but both MNW and WNT entries are interlinked by a manually verified set of correspondences that go beyond the homograph level. The temporal information is provided by the metadata that comes with the quotations providing the lexicographical evidence for the definitions from which the synonyms are extracted.

2.2. Ontolex-Lemon

A standard for the representation of lexical data in RDF is Ontolex-Lemon[5], developed by the Ontology Lexicon (Ontolex) community group (Ciminiano et al., 2016). The model is designed to give linguistic grounding to ontologies, by linking the ontology to lexical entries with grammatical and/or semantic information. The Ontolex community group is currently working on a module dedicated to lexicographical data[6] (Bosque-Gil et al., 2017; McCrae et al., 2017). Even though DiaMaNT is not a mere conversion of historical dictionaries into RDF, there is enough traditional dictionary content in DiaMaNT to be confronted with similar issues, like how to deal with sense hierarchy, how to model diachrony, etc. (cf. Khan et al., 2016, 2017; Bosque-Gil et al., 2017).

We will not describe the complete LOD model for DiaMaNT. Instead, we want to focus on those components that are essential for our lexicon building approach, and for which we had to define extensions to the model.

The main objective for the implementation of the data model for the lexicon is to do justice to the character of the underlying scholarly lexicographical work. The core of our lexica is the corpus of attestations from the historical dictionaries. By analysing the corpus material, using their expert knowledge, lexicographers provided a careful description of the meanings of each word in the dictionary. According to Kilgarriff (1997) "the scientific study of language should not include word senses as objects in its ontology. Where 'word senses' have a role to play in a scientific vocabulary, they are to be construed as abstractions over clusters of word usages." For him "the basic units are occurrences of the word in context (operationalised as corpus citations)." The senses from the dictionaries we use in our DiaMaNT lexicon, and the ontological layer we add to it, remain an interpretation of historical language that came down to us via text. This motivates the extensions we propose to the Ontolex-Lemon model. Senses, lexical entries, lexical forms, and temporal information are linked to attestations. Keeping the complete description of the senses, including the hierarchy, of the lexical entry, is also motivated by the desire to contextualise. Likewise, provenance information concerning the data processing for the DiaMaNT lexicon is included in the lexicon.

We will now give a brief the description of how attestations, sense hierarchy and provenance are modeled for DiaMaNT.

2.3. Attestations

Figure 1 shows how we link evidence ("attestations") to lexical categories which we conceive as interpretations for which the dictionary quotations (or corpus references) provide evidence. The main elements of the lexical entry (*LexicalEntry* itself, *Form* and *LexicalSense*) are assigned to the superclass *LexicalPhenomenon* (the name is maybe not very elegant, *Observable* might be another option).

In this way, the dictionary quotations can be seen as a partially semantically tagged corpus.

Table 2 shows part of our efforts to "put the corpus into the dictionary" (Kilgarriff, 2005) by means of the standoff corpus annotation approach of NIF[7] ontology, and to define a suitable metadata model on top of Dublin Core[8]. Unprefixed class and property names are extensions we had to resort to. The extensive quotation metadata is the main ingredient for the temporal and spatial dimensions of the lexicon. In contrast with the lemonDIA model (Khan, 2016),

[5] www.w3.org/2016/05/ontolex

[6] www.w3.org/community/ontolex/wiki/Lexicography

[7] persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html

[8] dublincore.org/

dictionary	lemmata	definitions	quotations	tokens
ONW	9.268	12.619	30.025	1.056.926
VMNW	25.946	102.202	194.366	6.463.868
MNW	74.773	144.714	400.619	13.078.231
WNT	467.288	553.672	1.667.835	51.246.034
Total	577.275	813.207	2.292.845	71.845.059

Table 1: Content statistics of the historical dictionaries

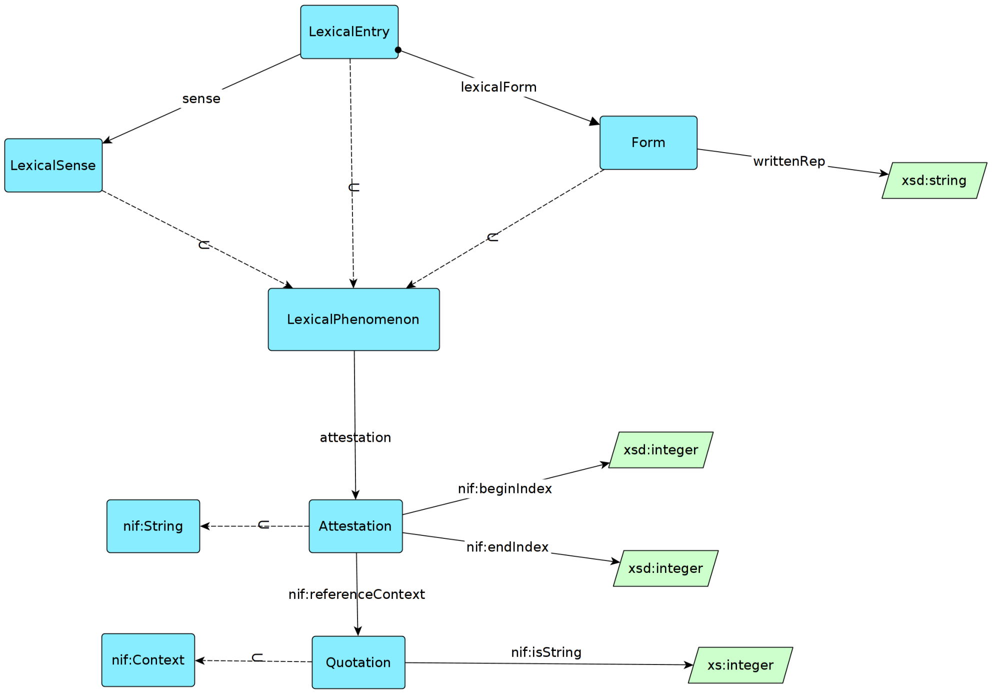

Figure 1: Attestations

the CLARIAH DICOLOD project[9] (Maks et al., 2016) and Cimiano et al., 2013, in which lexical senses are assigned to a time period, we aggregate this information from observed usage.

2.4. Senses, Subsenses and Definitions

Many scholarly dictionaries have a hierarchical subdivision of the sense section, mostly (but not exclusively - grammatical distinctions also play a role) based on semantic criteria. One might wonder whether it makes sense to model this subdivision in the more strictly structured semantic lexical infrastructure we work towards.

Despite the somewhat fuzzy semantic significance of the hierarchy, we think it makes sense to include it in the lexicon. Human perusal of, for instance, the result of a query over the data which presents an unstructured list of senses, immediately prompts the desire to know their position in the article hierarchy. Moreover, we have a usage and evidence-based view of meaning. A sense hierarchy implies a semantically motivated hierarchical subdivision of the evidence (set of quotations and their metadata in the entry). NLP applications like word sense disambiguation profit from the possibility of defining a coarse-grained division. Although the hierarchical information requires postprocessing to make it optimally suitable for this purpose, discarding it would entail unwarranted loss of information.

We briefly describe the sense-related part of the model. In agreement with the core Ontolex model, we use the *reference* property to refer from a lexical sense to a concept in an external ontology[12], and synsets are modeled by sharing Lexical Concepts. We encode the sense hierarchy by means of a (non-transitive) property *subsense* (in the Lemon namespace) and, like Bosque-Gil et al. 2017, an integer-valued data property *senseOrder* is attached to the sense nodes. Attaching the order information in this way implies that a sense cannot be shared among lexical entries, which is not a problem in our setting, as we

[9] github.com/cltl/clariah-vocab-conversion

[10] A terminus post quem is the earliest possible date something may have happened,

[11] A terminus ante quem is the latest possible date something may have happened.

[12] For instance, the Dutch National Species Register, www.nederlandsesoorten.nl/

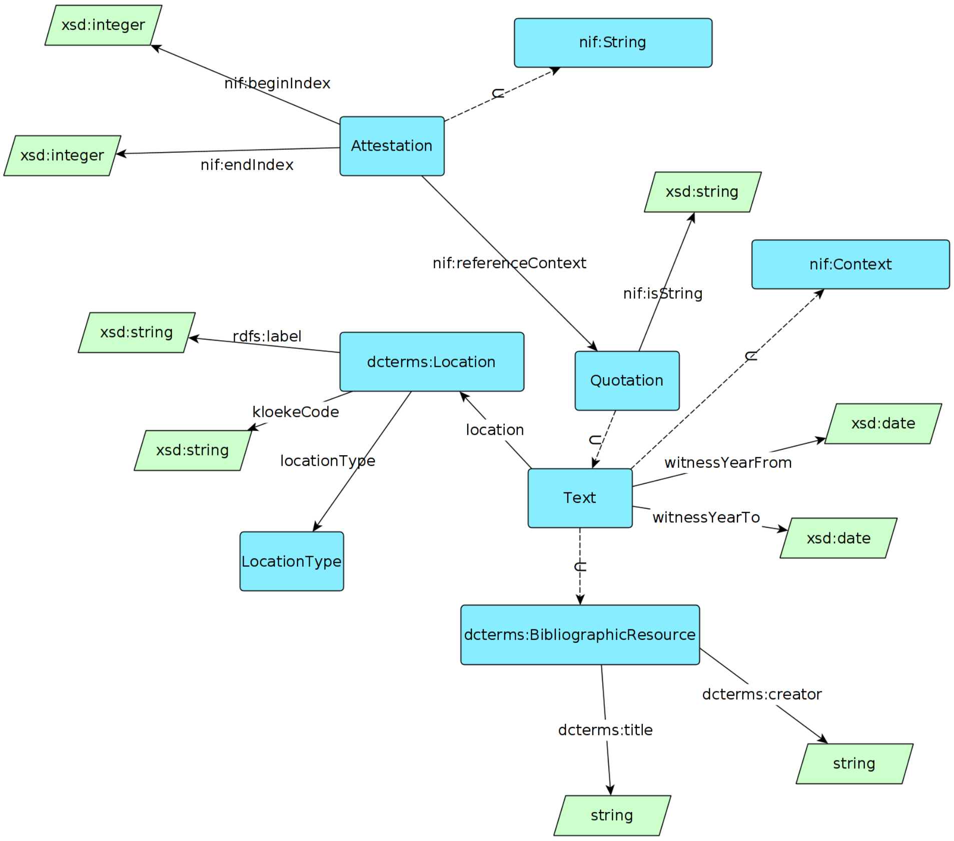

Figure 2: Attestation metadata

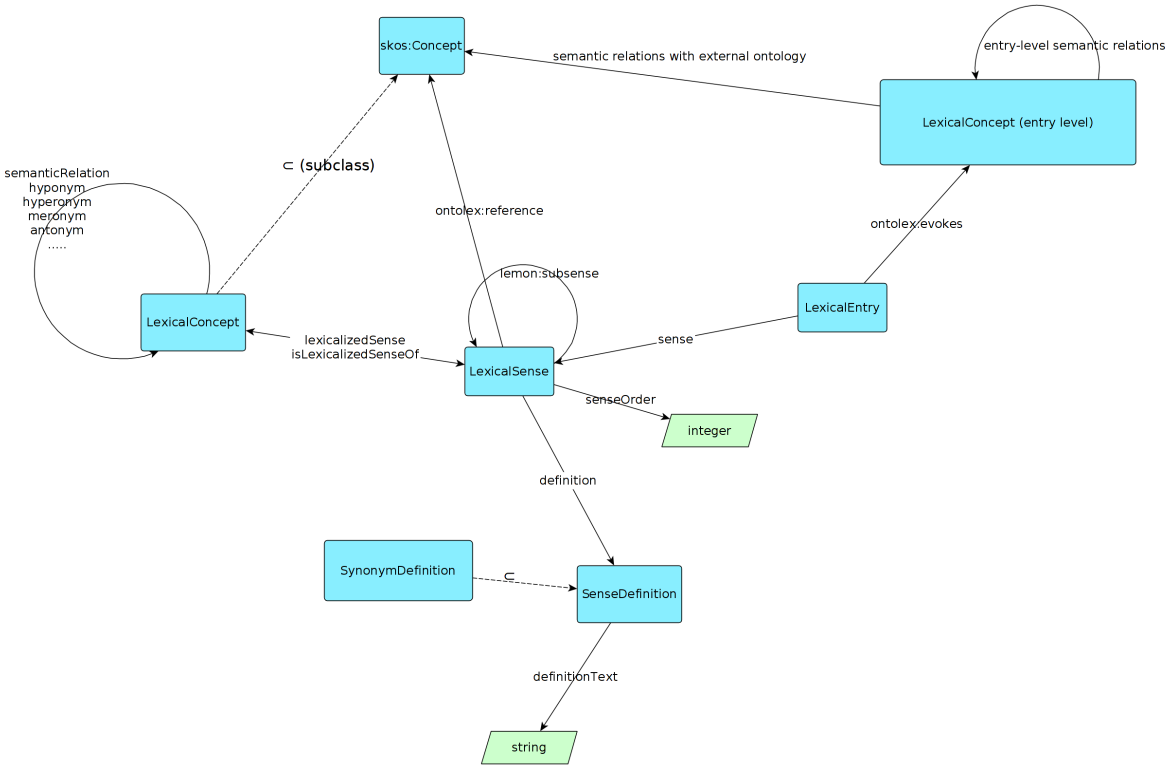

Figure 3: Sense and subsense structure

Metadata property	description
witnessYearFrom	Terminus post quem[10] for the document from which the evidence is obtained
witnessYearTo	Terminus ante quem[11]
LocationType	Some element in an enumeration containing levels like Country, Province, City, etc

Table 2: metadata properties

model (near)-equivalences between senses from different resources by links between the associated lexical concepts. The alternative options of modeling the hierarchy by means of RDF collections or containers, or the *senseSibling* property proposed by (Khan et al., 2016, 2017) generate a huge amount of extra triples, especially given the extensive hierarchy (maximum depth of 9 levels, with up to 760 "senses" per article[13]). We chose to re-reify definitions (current Ontolex dropped the *SenseDefinition* class of its predecessor Lemon and proposes skos:definition, which is a data property) in order to be able to attach provenance (and other information) to them. We further propose that the (Lemon) *SenseDefinition* class can be subclassed, according to different types of lexicographical definition. We are in the process of transforming automatically extracted synonym definitions into semantic links. We use the subclass *SynonymDefinition* to represent the synonym references extracted automatically from the dictionary definitions.

2.5. Provenance

Scholarly lexicography provides evidence for the assertions made. The user can assess the reliability of the interpretation on the basis of the evidence. When dealing with enriched data, equal standards should be adhered to. The PROV ontology[14] provides us with mechanisms to provide information about the provenance of the added layers of information. For the core lexicographical data, provenance is specified in a more succinct way by referring to the id's of data elements. For those enrichments which have been added automatically and only partially verified manually, it is important to distinguish the verified and the unverified instances. By restricting results to resources associated with agents from the subclass *Person*, a user can exclude the unverified part of the lexicon.

3. Conclusion and Future Work

The lexicon model has been tested by converting the dataset to a medium-size resource of about 40M triples and deploying it in a SPARQL endpoint using Jena TDB version 3.1.0[15]. In several realistic usage scenarios, both as a standalone resource and in combination with other resources (DBpedia, Open Dutch Wordnet, distributional thesauri), performance is quite acceptable for non-distributed queries (although the engine used is rather sensitive to the ordering of subqueries). Query formulation is not too cumbersome for users with some knowledge of SPARQL. The main remaining challenges (apart from the development of the lexicon content) are to improve performance on federated queries over several endpoints and to implement a user-friendly query interface for non-technical users.

4. Acknowledgements

This research was made possible by the Instituut voor de Nederlandse Taal and the CLARIAH-CORE project financed by NWO (www.clariah.nl).

5. Bibliographical References

Berners-Lee, T. (2006). Linked Data. Retrieved from https://www.w3.org/DesignIssues/LinkedData.html.

Bosque-Gil, J., Montiel-Posoda, E. and Aguado-de-Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: The K Dictionaries Case. *GLOBALEX 2016 Lexicographic Resources for Human Language Technology*, 65–72.

Bosque-Gil, J., Gracia, J. and Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in Ontolex. In *CEUR Workshop Proceedings* (Vol. 1899, pp. 74–84).

Cimiano, P., Mccrae, J., Buitelaar, P. and Montiel-Ponsoda, E. (2013). On the Role of Senses in Ontology-Lexica. In A. Oltramari, P. Vossen, L. Qin, & E. Hovy (Eds.), *New Trends of Research in Ontologies and Lexical Resources. Ideas, Projects, Systems* (pp. 43–62). Springer.

Cimiano, P., McCrae, J. P. and Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report, 10 May 2016.* Retrieved from www.w3.org/2016/05/ontolex/.

Depuydt, K. and De Does, J. (2008). United in Diversity: Dutch Historical Dictionaries Online. In E. Bernal & J. De Cesaris (Eds.), *Proceedings of the XIII Euralex International Congress (Barcelona , 15-19 July 2008)* (pp. 1237–1241). Barcelona: Institut Universitari de Lingüística Aplicada Universitat Pompeu Fabra.

Hirst, G. (2009). Ontology and the Lexicon. *Handbook on Ontologies*, 269–292.

Khan, F., Díaz-Vera, J. E. and Monachini, M. (2016). Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web. In I. Draelants, C. Faron Zucker, A. Monnin, & A. Zucker (Eds.), *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage Co-located with 13th Extended Semantic Web Conference (ESWC 2016)* (Vol. 1595, pp. 37–46).

Khan, F., Bellandi, A., Boschetti, F. and Monachini, M. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using Ontolex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon.

[13] gtb.inl.nl/iWDB/search?actie=article&wdb=WNT&id=M089102

[14] www.w3.org/TR/prov-o/

[15] jena.apache.org/documentation/tdb/

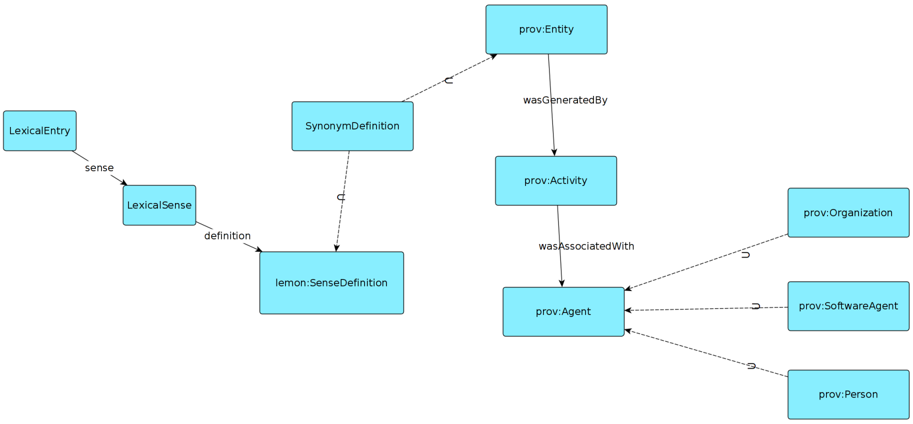

Figure 4: Provenance information attached to the automatically extracted synonym definitions

Kilgarriff, A. (1997). “I don’t believe in word senses”. In *Computer and the Humanities* 31 (1997), pp. 91-113.

Kilgarriff, A. (2005). Putting the Corpus into the Dictionary. *Proc. MEANING Workshop.* Trento, Italy. Retrieved from https://www.kilgarriff.co.uk/Publications/2005-K-Meaning-PCID.doc.

McCrae, J., Aguado-de-cea, G., Buitelaar, P., Cimiano, P., Declerck, Th., Gómez Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E, Spohr, D. and Wunner, T. (2010). *The Lemon Cookbook*. Retrieved from lemon-model.net/lemon-cookbook.pdf

McCrae, J. P., Bosque-gil, J., Gracia, J. and Buitelaar, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Elex 2017 proceedings*.

Maks, E., van Erp, M. G. J., Vossen, P. T. J. M., Hoekstra, R. J. and van der Sijs, N. (2016). Integrating Diachronous Conceptual Lexicons through Linked Open Data (pp. 1–2).

Moerdijk, F. (1994). *Handleiding bij het Woordenboek der Nederlandsche Taal (WNT)*. ’s-Gravenhage: Sdu Uitgeverij Koninginnegracht.

Mooijaart, M. (2013). A History of Dutch Lexicography. *Trefwoord, Tijdschrift Voor Lexicografie*, 1–34.

Moreau, L., Groth, P., Cheney, J., Lebo, T. and Miles, S. (2014). The Rationale of PROV. *Journal of Web Semantics*, *35*, 235–257.