Re-linking a Dictionary Universe or the Meta-dictionary Ten Years Later

Ore, Christian-Emil
c.e.s.ore@edd.uio.no
University of Oslo, Norway

Ore, Espen S.
e.s.ore@iln.uio.no
University of Oslo, Norway

More than ten years ago what was called a meta-dictionary was proposed as a central part of the framework for a dictionary laboratory at the University of Oslo (Ore 2001). The framework has since functioned as a pivot in the combined lexical database, text corpus and manuscript editing system for *Norsk Ordbok* (Norwegian Dictionary). *Norsk Ordbok* is published in twelve volumes (to be completed in 2014) and provides a scholarly and exhaustive account of the vocabulary of Norwegian dialects and the written language Nynorsk, one of the two official written forms of Norwegian.

The architecture of the dictionary framework described in this paper was based upon both explicit and implicit assumptions - and some of the latter were not only not consciously considered in the construction phase, they have also led to features or lacks of features in the system where we now see the need for change. In this paper we look at problems related to links between the meta-dictionary and the sources and show how some of the problems are solved.

1. The meta-dictionary?

In the 1990s a huge amount of lexicographical source material (dictionaries, slip archives and texts) was made electronically available by a national digitization project. By then *Norsk Ordbok* had produced three volumes out of twelve. Being a project started in 1930 the future of the project was highly uncertain. Thus the original motivation behind the meta-dictionary was to create a common web based interface to the background material by inter-linking the material to a common headword list as a meager substitute for the edited dictionary. A similar approach has later been taken by the Dictionary of Old Norse Prose in Denmark (ONP).

Fortunately, the *Norsk Ordbok* project was refunded and revitalized in 2001. It was decided that the new project should be completely digital. As a result a new version of the meta-dictionary was designed.

An entry in the meta-dictionary can be seen as a folder containing (pointers to) possibly commented samples of word usage and word descriptions taken from the linked databases etc. Each entry is labeled by normalized headword(s), word class information and the actual orthographical standard used. The working lexicographers view the meta-dictionary as an easy access to systematized source material. The chief editors use it as tool for headword selection and in dimensioning the printed dictionary. The database used in the Cobuild project for lemma selection from the corpus, is an early example of such a database (Sinclair 1987).

The *Norsk Ordbok* is a historically oriented dictionary covering the period 1600 to the present. The time span and the focus on dialects make the background material heterogeneous. The oldest sources are glossaries compiled in the 17th/18th centuries, mainly the results of work done by vicars collecting information about their parishioners' language on request from the government in Copenhagen. For the description of the word inventory of the current dialects surveys and especially local dictionaries form valuable sources. The meta-dictionary constitutes a bidirectional network. Thus the historical or dialectal dictionary linked to the system can be used as an entry point to the entire set of information.

2. Building a dictionary net

The traditional systematic overviews of the use of words in context have been alphabetically ordered paper slips with each word in a small context and the source information. The slip collections have gradually been replaced by text corpora. In the *Norsk Ordbok* project the slip collections are digitized and linked to the meta-dictionary, and a new annotating tool for singular language observations has been developed. A standard TEI-encoded text corpus spanning the period 1850 to present is gradually constructed. The results from corpus queries can be stored and linked to the meta-dictionary.

The old and the local dictionaries and glossaries constitute an important source for historical and dialectal word usage respectively. Traditionally such dictionaries and glossaries have been transcribed to paper slips and stored in the slip collection. In the new system the dictionaries could have been included in the corpus. This may be done with the newer dialect dictionaries. The old dictionaries are written in Danish or Latin and would have introduced a lot of linguistic noise in the corpus. As these dictionaries are important documents in themselves it was decided to treat them as individual works documenting the language view of their time.

The modern dialect dictionaries are given an XML-encoding according to TEI's printed dictionary format. The 17th /18th centuries' dictionaries are represented by printed, annotated text editions of the original manuscripts. These editions have been transcribed and given a TEI markup reflecting their structure, generally not compatible with TEI's printed dictionary format. Due to their systematic character, <div>-elements can be used to organize the text into chunks describing words and thematic sets of words. The “headwords” are clearly identifiable and are marked as <w>-elements. The loose structure implies that there may be more than one “headword” in each text chunk. The TEI-texts are stored as blobs in a relational database. The TEI-texts are chopped up according to the entries (dictionaries) and the text chunks (glossaries) and stored together with the headwords (slightly normalized) in a separate table structure.

In the early version of the system, the linking between these sources and the meta-dictionary were on the <entry> level for the local and on the <div> level for the old dictionaries. There was no information about the keyword in the selected dictionary that was used to create a link. In some cases when a <w>-element was removed an invalid link from the meta-dictionary to the external text sets was left. Today the link is annotated with the actual headword and the person responsible for the link.

The process is automatic with a manual check: a daily job runs through registered dictionaries and looks for keywords in a special metadata field in the database. If the word is found but there is no existing link between the meta-dictionary and this text unit, a link is created, and the record in the meta-dictionary is marked as changed and will be forwarded to an editor for approval. If the word is not found in the meta-dictionary, a new entry is created and linked with the text unit, and this will be sent to the editor for approval (see also Fournier 2001 and Gärner 2008 for interlinking of dictionaries).

3. The meta-dictionary and other dictionary nets

What constitutes a word is an unresolved linguistic question. A traditional monolingual dictionary is a word form oriented index to a set of concepts and meetings where each entry is indexed by a headword and contains a possible meaning hierarchy with samples. Word forms denoting related concepts are connected by cross references. The Wordnet approach is to focus on the concepts and collect the word forms denoting the same concepts in sets of synonyms (synsets). The synsets can then be organized according to a predefined ontology such as in the Global Wordnet Grid (Vossen 2009). The two ways of organizing word information is fully compatible. A word net can be converted to a traditional dictionary and a well organized dictionary rich on semantic references can be converted into a Wordnet.

The current meta-dictionary was pragmatically designed 8 year ago. It has in itself become a valuable lexicographical documentation system. The source material spans both in time and space. Due to the practical purpose, that is, editing a traditional dictionary, the word forms are linked mostly etymologically. Thus an entry covers many concepts as does an entry in a traditional dictionary. However, the meta-dictionary also groups how different scholars have described the meaning of a word from 1600 to the present. The resources comprise the old digitized paper slip collections, the dictionaries and glossaries and stored results from querying the corpus. They all represent collections of systematized language documentation in their own right and premises. The entries in the *Norsk Ordbok* are in fact just yet another source of (scientifically) systematized language information linked to the others. However, the *Norsk Ordbok* system groups the information according to meaning and the dictionary is rich in synonym relations. A future research project is to use the information from the dictionary to create a second set of articles in the meta-dictionary in a Wordnet fashion with semantic relations.

References

Fournier, Johannes (2001). 'New Directions in Middle High German Lexicography: Dictionaries Interlinked Electronically'. *Literary and Linguistic Computing*. **16/1**: 99-111.

Gärtner, Kurt (2008). 'The New Middle High German Dictionary and its Predecessors as an Interlinked Compound of Lexicographical Resources'. *DH2008 conference*. Oulu, Finland, 2008.

Ore, Christian-Emil (2001). *Metaordboken - et elektronisk rammeverk for Norsk Ordbok?*. Gellerstam, Martin et al. (ed.). Nordiska studier i leksikografi. Göteborg. V. 5.

Sinclair, J. M. (ed.) (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

Vossen, Piek (2009). 'From Wordnet, EuroWordNet to the Global Wordnet Grid'. *eLEX2009 conference*. Louvain-la-Neuve, Belgium, October 22-24.

TEI, Text Encoding Initiative. `http://www.tei-c.org/P5/`.

Norsk Ordbok. `http://www.no2014.uio.no`.

ONP, Dictionary of Old Norse Prose. `http://www.onp.hum.ku.dk`.

WordNet. `http://wordnet.princeton.edu/`.