Åse Wetås

The approach to digitization and experience with working with complex database systems and digital editing platform in the *Norsk Ordbok*

1. Introduction

Norsk Ordbok is an academic dictionary covering Norwegian Nynorsk literature and all Norwegian dialects. The dictionary will provide a scholarly and exhaustive account of spoken Norwegian and of texts written in Norwegian Nynorsk from 1860 and up till today, and is to be completed by 2014, in time for the bicentenary of the Norwegian constitution. From 2002 on the dictionary work has been organized in the time limited project organization Norsk Ordbok 2014 (NO 2014). The project owner is the Department of Linguistics and Scandinavian studies at the University of Oslo. In 2014, the finished work will include more than 300 000 entries, published in 12 volumes.

This article will focus on how the production of the dictionary and the development of lexicographic method in the project organization have profited from digitization, from the construction of a compex relational database system of the digitized sources and from the development of an digital editing system. To those unfamiliar with the background of *Norsk Ordbok* and of recent Norwegian language history, I will give a short introduction. The aim of this short excursion is to present the historical background to why *Norsk Ordbok* covers both written and spoken Norwegian, and also to explain why our technical solutions are constructed in the manner they are.

After this short introduction to the Norwegian language scene, I will use the main part of the article to describe the basic components of the Norsk Ordbok 2014 digital editing system, comment on how the scientific programmers at The Unit for Digital Documentation ("Eining for digital dokumentasjon", EDD) at the University of Oslo, have developed these resources in close cooperation with the lexicographers in the dictionary organization, and on the effects of leap into the digital age. The initial goal of the digitization when it started out, was to improve the efficiency of the dictionary production. However, during the almost ten years of editing the dictionary on a digital platform, we have experienced that the digitization and the close integration of the digital dictionary entries with the source material also have had important scientific implications. Some of these are already described by Bakken (2006), but since then others have been added. I will come back to these implications towards the end of the paper.

2. Historical background

For four hundred years, from the 15th to the 19th century, Norway was under Danish rule. During that period, Danish was the official language of the church and the crown, and it was the written language Norwegian children were thaught at school. The written sources of

Norwegian from this four hundred year period are confined to short vernacular texts (mainly letters, short stories and verse). The rise of national romanticism in the early 19. Century, led to a dissatisfaction with the fact that there was no separate Norwegian language, but the proponents of establishing a written Norwegian did not agree on how to do this. In fact, the main problem was a disagreement on what language to base this new written standard of Norwegian on. Thus, two parallel standards were eventually established. One of them was the Dano-Norwegian or Norwegian Bokmål, based on written Danish and on urban Norwegian middle class 19. century speech. The other one was the Norwegian Nynorsk, based on the Norwegian dialects and Old Norse.

Norwegian Nynorsk was constructed in the 1850s by the Norwegian linguist and lexicographer Ivar Aasen. Aasen had travelled around in Norway collecting data and documenting the Norwegian dialects. Based on an impressive amount of empirical data, he investigated the stuctural dependencies between the dialects and codified a written language that represented the common, underlying norm of all Norwegian dialects – or Proto Norwegian, as another famous Norwegian linguist, Einar Haugen, later described it (Haugen 1966). Aasen used comparative and historical lingusitic method in his work, and his Norwegian Nynorsk consisted of the set of forms that he considered to be the best representatives of the phonological and morphological diversity he found in the dialect samples. He also made comparisons with Old Norse, Danish and Swedish to find related patterns. In 1873 he published the comprehensive normative dictionary *Norsk Ordbog* (Aasen 1873), which contained 45 000 entries. This was a revised and enlarged version of his 1850 dictionary, which he had named *Ordbog over det norske Folkesprog* ("Dictionary of the Norwegian People's Language", Aasen 1850). The title of this early dictionary reveals Aasens main goal – he wanted to codify a written Norwegian that would give the Norwegian people an opportunity to read and write a language that was close to their own everyday spoken Norwegian. In 1885 the Norwegian Parliament stated that Norwegian Bokmål and Norwegian Nynorsk were equal official written standards of Norwegian.

When *Norsk Ordbok* was conceived in the late 1920s, Norwegian Nynorsk was still a langauge in the making, and the written standard was continuously fed by Norwegian dialect words. The proponents of the standard wanted to make a huge academic dictionary based on the work of Ivar Aasen. The immediate goal behind the dictionary was to develop Norwegian Nynorsk further, and to raise the prestige of the new written standard. The collection of data for this new academic dictionary started in 1930. A dictionary board of trained lexicographers instructed and supervised more than 550 volunteers who during these early years collected dialect data from all over the country in order for the dictionary board to build up a slip archive. The learned dictionary board also supervised the extraction of literary excerpts from the Nynorsk literature, both fiction and non-fiction. In addition, they compiled Aasens dictionaries together with a range of other canonical dictionaries from 1870 to 1910, and also added data from glossaries and local dictionaries from 1600 to 1850. By the year 1940 they had established a draft manuscript for this new, academic dictionary.

The editing of the dictionary started in 1946, and the first volume of *Norsk Ordbok* was published 20 years later. Volume 1 covered the alphabet from the letter *a* to the adjective *doktrinær*. The original plan was to make a 2-3 volume dictionary, but in 1966 the chief editor estimated that 8-9 volumes would be needed to cover the whole alphabet. During the first 50 years the editing of the dictionary progressed slowly. At the same time the source material grew larger, and so did the dictionary entries in volumes 2 and 3. All the work was done manually, the slips sorted on the lexicographer's desk and the manuscripts prepared in hand writing.

In the 1990s huge amounts of data from the Norwegian language collections were digitized. At that time, the *Norsk Ordbok* slip archive included some 3 million slips covering dialect data and literary excerpts, and a substantial amount of them were in hand writing. The digitization of the slip archive represents the first step towards the development of a new way of editing an old dictionary. Facsimiles of the slips were made, and these were later indexed with headwords and relevant meta data. As Bakken (2006) states, the ideal solution for the slip archive would be to digitize the whole contents of the slips, but this was considered too expensive at the time. I would like to add that some of the slips are even illustrated with small sketches, and this would also be an obstacle to a complete content digitization.

After 70 years of work, three volumes were published, and the alphabet was covered from the letter *a* and to the verb *gigla*. The progress had been slower for each new volume and the total number of volumes was uncertain. In year 2000 a political initiative was carried out to finish *Norsk Ordbok* in 12 volumes in time for the celebration of the bicentenary of the Norwegian Constitution in year 2014. Thus, in 2002 an agreement was made between the Norwegian Ministry of Church and Cultural affairs and the University of Oslo to split the cost of establishing a time limited lexicographical project organization to do this job. The dictionary organization was reorganized and the process of rationalization through digitization started.

3. The process of moving the dictionary to a digital platform

3.1 The Meta Dictionary

The first step of the reorganization was to find ways to exploit the results of the large scale digitization of the slip archive and the other digitized components of the *Norsk Ordbok* source material. By 2002, an electronic indexing system was established, and this has become the hub of our editorial database application (for a description of the indexing system, see Ore 2000, Runde et al. 2005). The index is named "The Meta Dictionary", and has a lemma list where the entries represent folders with headwords in normalized spelling. Each entry contains links to relevant data in the whole range of databases of source material. As a result of Norwegian Nynorsk being a young written language, the oldest parts of the empirical data are weakly normalized. This goes particularly for texts older than the 1938 spelling reform of Norwegian, where words can appear in several different spellings. This is a well known problem to all lexicographers working with historical and diachronic data, but the mapping system of the Meta Dictionary and the compilation of complex data under normalized headwords represents a way to deal with this problem. Today, the Meta Dictionary contains some 580 000 entries.

The Meta Dictionary index represents a rationalization in itself, in that it compiles data in different formats and from a whole range of sources. In addition it supplies the lexicographers with a digital list of lemmas, each with a specific number of links to the instances. This enables the organization to make calculations on which entries to include and which ones to exclude from the dictionary based on number of instances, as well as to calculate on the relative dimension for each single dictionary entry. It also enables the NO 2014-organization to make a master plan for setting alphabet dimensions for the whole dictionary. These two functions are very important to control the quantitative production in the dictionary organization, and the quantitative production can be measured on a monthly basis. This tool is very valuable to a modern dictionary project with limited resources.

The Meta Dictionary communicates with a set of relational databases. These databases include the old slip archive, which was the first of the collections to be digitized. Today the slip archive contains some 3,2 mill. slips, and for the last ten years editors and assistants have added new slips electronically. The database system further includes the *Norsk Ordbok* draft manuscript from 1940, the complete bibliography database and a “Dictionary Hotel” consisting of local word lists from the last 60 years. The Dictionary Hotel is a database with unlimited space for new hotel guests, and the guests are local dictionaries and word lists from all over Norway, digitized and published online by NO 2014 in agreement with the copyright owners. We also have a corresponding “Dictionary Home” for old glossaries from the 17th to 19th centuries. These are not tied up in any copyright restrictions and we can therefore offer the old glossaries a permanent, digital housing.

The Meta Dictionary further includes source material from the two modern Norwegian prescriptive dictionaries *Bokmålsordboka* and *Nynorskordboka*, and a range of smaller collections. The electronic index in this way represents the hub of the language collections. Source material from an unlimited range of databases can be included, and the index communicates directly with our digital editing system as well as with the dictionary database where all the *Norsk Ordbok* entries are stored. Each digital lemma in the Meta Dictionary contains a list of instantiations, both facsimiles and strings of digitized text.

The electronic linking up of the source material in the Meta Dictionary with the *Norsk Ordbok* entries secures that the interpretation of data and the products of scientific research can be easily reproduced. An unbreakable rule in the process of editing *Norsk Ordbok* is that all entries have this link to the corresponding Meta Dictionary node, and through that to the empirical data behind the entry. Reproducibility is very important to an academic dictionary. In fact, it is a basic ideal of all scientific enterprise that analyses are reproducible and in principle possible to refute. By the close integration of the dictionary entries and the source data, we try to meet these ideals.

3.2 The text corpus

The next step that was taken to rationalize the dictionary production in 2002, was to compile texts for a monitoring text corpus. This corpus, *Det nynorske tekstkorpuset* (“*The Nynorsk Text Corpus”*) is a very important supplement to the traditional language collections, paticularly when it comes to covering high frequency verbs, function words, collocations and idioms. Today, *The Nynorsk Text Corpus* consists of more than 90 million words and it is accessible on our website in both a tagged and an untagged version (www.no2014.uio.no/korpuset). Our experience from the last ten years is that the traditional language collections and the text corpus supplement each other in a very satisfying manner. The slip archive gives the editors information on dialect words that are frequent in spoken Norwegian but often rare in written texts, and it provides an older layer of semantics and an important historical dimension. The text corpus enables them to investigate and document the standard lexicon of the language, high frequency words, compounds and collocations. It also solves the problem of the double subjectivity of literary excerpts, which amongst others Atkins (1992) has pointed to.

3.3 The editing platform

The third step was to make a digital editing platform that could generate words from the Meta Dictionary and accommodate editorial practice. This editing platform was taylor made for us by the EDD. Our senior editor in charge of digital development made the specifications, and the scientific programmers at the EDD developed the database format. The editing staff in *Norsk Ordbok* did the consecutive testing (see Grønvik 2005, Ore & Tvedt 2006). The end result is an editing platform which has automatized a whole range of the routine operations. This includes integrated scroll bar menues of references linked to the bibliography database, and it includes scroll bars of geographical names, grammatical categories and etymological information, to mention some of them. These standarized procedures release time for the editors to focus on the linguistic enterprise – including interpreting the data, establishing the semantic structures and investingating morphological, phonological and syntactical regularities. The interface of the editing platform, the fields and the menues of fixed abbreviations and references also secure a uniform structure of the articles. Our editors are dedicated researchers, and they highly appreciate the opportunity to spend more time on scientific work and less on time consuming routine tasks.

The editing program also includes a sub article component used for editing multi word expressions as integrated sub articles inside the alphabetically structured entries of the printed dictionary. Fixed multi word expressions are very interesting from a theoretical point of view and they are important both to first and second language learners. At the same time they are notoriously difficult to treat in a uniform manner, and they have strange and often unpredictable meanings. The sub entry article component of the editing system is a very valuable tool for covering multi word expressions in an orderly and predictable manner.

3.4 The sorting module

The digital editing platform has been continuously refined during the ten years of the NO 2014-project, and an important addition was the inclusion of a sorting module that enables the editors to import data from the Meta dictionary and the text corpus and sort them in one single operation (see Bakken & Grønvik 2008). This sorting tool is very flexible, and all editors are free to define their own sorting criteria. This represents a huge step forward compared with the old fashioned sorting of physical slips on the editor's desk. All dictionary editors who have worked with a traditional slip archive know that the perusal of paper slips can be a very time consuming activity. The slips often include information on a whole range of items, such as morphology, phonology, semantics and usage. For the editor, this means that the paper slips have to be sorted several times in order to get relevant information for the different parts of the entry. With the integrated sorting tool all results can be saved and altered whenever the editor wants to, and it also makes it easier for the editors' supervisors to evaluate the hypotheses and generalizations that are expressed. Bakken (2006**:** 120) also comments on a third result of the sorting module: “the dictionary is much more integrated with its sources than before […] this situation scientifically is a very interesting one.” I would add that this of course is very relevant to all historical and academic dictionaries. The classification of instances in the sorting tool is considered the editor’s personal sketch for the article, and are not meant for publishing.

3.5 The dictionary database

In the dictionary database, the articles are stored in XML. The editors have the opportunity to link entries electronically, both on lemma level and on the level of definitions within the entries. These links constitute a semantically based growing network in the *Norsk Ordbok* database. This electronic linking will prevent circularity, and it is therefore an important tool for the scientific quality of the dictionary. The semantically based electronic network of articles also adds new and interesting opportunities to survey specific semantic fields and taxonomies. A very illustrating example is the *Norsk Ordbok* coverage of local names for all kinds of wild flowers. In the dictionary each of these wild flower names is linked electronically to the official Norwegian name of the species. Linked together, the sets of articles form beautiful taxonomic networks that could be of interest to researchers in other fields of research (folkloristics, ethnology, ethnobiology, history etc.).

3.6 The 2010 slip archive module

In 2010 we launched a new version of the electronic slip archive module of our database system. This is a remoduled version of the old slip archive module, and it is set up to handle both acoustic data and pictures, in addition to the more traditional strings of text. In addition, the new slip archive module has made it easier to add new data. The 2010 electronic slip archive enables us to harvest and store data in other formats than the traditional. One important example is text samples harvested from the Internet. To meet the demands of reproducibility, we are now able to store static print screen versions of web pages together with the harvested word and grammatical information in question.

3.7 The online dictionary

Our most recent achievement is an online version of *Norsk Ordbok*, freely available on the Internet. The online version was launched in March 2012, and the e-dictionary profits greatly from all the efforts we have taken to improve data integrity since 2002. The articles produced for the printed version of *Norsk Ordbok* are stored in XML, and this is also the basis of the online version. XML-technology is used for making the display format, and this solution was deliberately chosen when the program application was constructed in 2002, to facilitate for an online version of the dictionary alongside the printed version. Our online dictionary is a preliminary version of our online dictionary product, but we hope to refine this by adding new functionality and develop advanced search options.

4. Concluding remarks

Norsk Ordbok 2014 digitized its collections and set up a digital editing platform with the intention to rationalize the production of the dictionary. In this, we have so far succeeded. We now use 15 months, approximately 35 man year of work, to produce each 800 page volume of the dictionary. In the pre-digital production, we used 64 man year of work to produce the same amount of dictionary text. What we also have gained from the digitization process and the refining of the data base system, can be summarized in the following bullet points:

- The digital tools and products are essential in our work to meet the scientific standards of an academic dictionary
- The editors can spend more time and effort on the linguistic enterprise and less on time consuming routine work
- The stringency of the articles is improved, and the uniform system of abbreviated references and labels makes it easy to organize the text and expand abbreviations in the online dictionary. The traditional lexicographical problem of strictly limited space is not urgent in an online dictionary, and this gives us the opportunity to present an even closer integration of the dictionary entries with the underlying source material to the public
- The Meta Dictionary index constitutes the hub of the application and makes it possible to process data from a wide range of sources, both syncronic and historical
- The 90 mill. word electronic text corpus represents an important and valuable supplement to the traditional language collections, particularly when it comes to the coverage of high frequency verbs, function words and multi word expressions
- The uniform structure of the data, the networks of electronic links between the entries in the dictionary database and the high degree of integration all pave the way for the future development of new lexicographical products based on the contents of this database

Bibliography

Aasen, Ivar. 1850. *Ordbog over det norske Folkesprog*. Kristiania: Carl. C. Werner.

Aasen, Ivar. 1873. *Norsk Ordbog*: *med dansk Forklaring*. Christiania: Mallings Boghandel.

Atkins, Sue. 2008. Theoretical Lexicography and its Relation to Dictionary-making. In Fontenelle, Thierry. 2008. *Practical Lexicography*, 31–50. Oxford: Oxford University Press.

Bakken, Kristin. 2006. The Dictionary and its Sources: the Ideal of Integration and the Example *Norsk Ordbok*. In Corino et al. *Proceedings*. *XII Euralex International Congress*. Torino, Italy, September 6th – 9th 2006: 117-122.

Det nynorske tekstkorpuset (*The Nynorsk Text Corpus*): www.no2014.uio.no/korpuset

Haugen, Einar. 1966. *Language Conflict and Language Planning. The Case of Modern Norwegian*. Cambridge, Mass.: Harvard University Press

Norsk Ordbok: www.no2014.uio.no

Ore, Christian-Emil. 2000. Metaordboka. *Ord om Ord* 6: 30-32.

Runde, Ålov, Terje Svardal, Oddmund Vestenfor. 1995. Metaordboka som reiskap for lemmaseleksjon og ordboksproporsjonering. *Nordiske studiar i leksikografi* 7: 326-30.