Norsk Ordbok 2014 from manuscript to datebase - standard gains and growing pains

ODDRUN GRØNVIK

Norsk Ordbok 2014, University of Oslo
P.O.Box 1021 Blindern, N-0315 Oslo
oddrun.gronvik@iln.uio.no

This paper will present a review of the digitisation process of of a major academic dictionary through the initial phase of the Project Norsk Ordbok 2014 (2002 – early 2005). The hypothesis at the start was that a thorough revision of editorial practice, linked to creating a stringent digitised dictionary writing system, would create a more reliable and consistent dictionary, with clearer procedures for processing source materials and composing entries. An efficient Dictionary writing system (DWS) application would also help train new editors and make them productive in less time than what has traditionally been assumed necessary.
Having publishing the first volume after the project started, and being well under way with the next one, it can be shown that the major goals described below on the whole have been achieved. The paper will discuss some areas in depth, look at the advantages, but also point out some possible pitfalls and some lasting difficulties.

Background

Norsk Ordbok ('The Norwegian Dictionary') was started in 1930 with the aim of providing a scholarly and exhaustive account of the vocabulary of Norwegian dialects and the written language Nynorsk, one of the two official written standards for Norwegian. The model was that of the big academic dictionaries for English, German, Swedish and Danish.

However, Norsk Ordbok differs from most academic documentary dictionaries for European languages in using records of spoken language as well as literature for its source material. Determining an etymology and suggesting a standard form for words documented only through dialect transcripts of necessity forms part of the lexicographical work, and adds to its complexity, but has to be seen as an integral part of Nynorsk lexicographical tradition (Grønvik 1992)

The task was underestimated from the start both in terms of complexity and effort. The work on Norsk Ordbok started with a lengthy phase of material collection and basic material collation. Real editing started in 1947 and had by 2001 reached into the letter h (in volume 4 out of 12 planned volumes). The progress rate through the alphabet had then slowed down steadily, while lexicographical treatment grew more and more detailed. Around 2000, at the then rate of progress, Norsk Ordbok could look towards a final publication date for the last volume after 2060, which could be read as another way of saying that the dictionary would never be finished.

Project refinancing, revision and terms

A project with the aim of completing Norsk Ordbok by 2014, in time for the bicentennial celebrations of the Norwegian Constitution, was started in 2002, with financing guaranteed by the Norwegian Storting (Parlia.ment) and by the University of Oslo.

The project is called Norsk Ordbok 2014 with the abbreviation NO 2014.

The project plan is based on the following conditions (from the funders)

1. NO 2014 must be completely digitised (materials, tools, manuscript)
2. Editorial methods and rules for NO 2014 must be revised to fit (a) digitisation, (b) training a large number of new editors in a very short time, without any loss of academic standard or research quality in the dictionary manuscript.

3. NO 2014 must prove itself useful to linguistic research beyond the purposes of dictionary itself, and it must fit into the larger strategies for research and academic development at the University of Oslo.
4. All digitised materials and research results developed within NO 2014 must be made generally available to the public as soon as possible within the course of production.

The only way of meeting these demands was to edit the the dictionary directly into a relational database, from which an xml file could be generated and modified to produce a correct print page. The chosen software was Oracle, already used in digitising the main language collections of NO 2014.

Critical philological review combined with computational analysis

The present paper deals with how these demands were met through a long term intensive cooperation between the NO 2014 and the Unit for Digital Documentation (EDD) at the Arts Faculty, University of Oslo. The following deserne particular mention: Dr. Christian Emil Ore, Lars Jørgen Tvedt, Dr. Daniel Ridings. Without their inspired commitment, NO 2014 would not stand where it is now.

In this cooperation, the major component was a detailed analysis of editorial practice in volume 1 - 4 by the senior editors (working from inside knowledge of the Norsk Ordbok tradition) and by key staff members at EDD (extracting structure by developing a parser for volume 1- 4, and forcing analysis and discussion of each entry component by programming for maximum data integrity).

This process will be illustrated by four case sketches:

1. Entry structure (linearity of running text versus tree structure)

The formal body structure of an entry in volume 1-4 was supposed to have four levels of sense units, marked by (a) upper case letter in bold (only shown if more than one), (b) Arabic numeral in bold, (c) lower case letter in bold:

A
1
a

There was also the possibility of using (d) Arabic numerals for ordering the meanings of polysemous idioms within a sense unit: 1,2,3 …

In addition, the following markers were used as separators within the sense unit:

// (double slash) for sub-definitions and for fixed phrases and idioms followed by their own definitions,

/ (single slash) for quotations (followed by source references and comments)

; (semicolon) for a part of a definition with a different shade of meaning.

It was intended that elements should be ordered so that double slash marked a stronger division than slash, which again marked a stronger division than semicolon. However, the complex material, set against insufficient editorial rules, left a wide field for individual judgment and improvisation.

The manuscript parsing performed by EDD gave this result for these separators:

Separator can be followed by	**Double slash //**	**Single slash /**	**Semicolon ;**
New definition within sense unit	X	X	x
Idiom (with one or more (numbered) definitions (1, 2, 3))	X	X	x
Quotation or editorial example with comment	X	X	
Sub-definition	X	X	
Introduction to idiom(s)	X		x
Introduction to list of compounds with headword as final element	X		x
Quotation or editorial example without comment		X	x
Introductory comment to quotation or editoral example (mostly style marker)		X	
Cross reference			x
Part of comment after quotation			x
Part of definition			x
Part of definition after idiom			x
Variant information after idiom			x
Etymological information after idiom			x

In short, (a) the entry format was more finely graded than provided for by the editorial rules, (b) all separators had multiple uses in order to cover all needs (c) the descriptive elements of the entry were to some extent created to meet the complexities of the material at hand.

Further, the parsing showed that the marking up of an entry structure to a considerable extent was **relational**, i.e. determined by the relative weight of materials for that particular entry, and not by general criteria for different linguistic categories. The result was a fluid presentation which often read well, but was lacking in hierarchy and consistency above entry level.

With the evidence from the manuscript parsing on the table it was easy to agree on wanting (a) a full revision of field system and entry structure, (b) restraints on the entry structure which ensure an open tree structure, maximum four levels and the necessary restraints to ensure consistency, (c) explicit editorial rules for each entry component.

The result is the new DWS **sense unit**, constructed from (sets of) interlinked tables. An entry body can have an unlimited number of sense units in the **A1a-**structure, but only the sense unit at the end of a tree branch can exploit the format to the full.

The sense unit format now has four major components in fixed order:

1. Main definition followed by examples
2. One or more sub-definitions followed by examples
3. One or more sub-entries for lexicalised phrases
4. (Illustrative) compounds where the headword is the first or the last element.

The real innovation is nr 3, the sub-entry for lexicalised phrases, which in turn has forced us to deal systematically with phraseology as a sub-discipline of linguistics. This development has been pushed forward by the creation of a (so far) ca 20 million word corpus in addition to our older collections.

2. Multiple use of materials and fuzzy documentation

A historic and documentary academic dictionary depends on its use of sources, not only in terms of documentation but in terms of consistency. An important part of NO source materials is word collections from Norwegian dialects, from ca 1600 until today, in manuscript form and as printed books. Another important sub-set of sources are written accounts of tradition and country life, often written in dialect-marked, non-standard language. Finally, NO 2014 also has a large collection of transcribed dialect words from our own informants. For these, and for other unquestionable dialect items in our collections, only the place of origin is given as a source.

The original editorial rules for handling literary versus geographical sources were not stringent enough for the growing collections, and also practice changed over time. Further, a general shortage of coverage for many words could tempt editors to over-exploit sources, by f. i. listing an 17th century dialect form both as a historic form and

as a speech form, or by using a quotation from non-standard language to show a change in the written standard. Nynorsk is a young written language, standardised through consecutive reforms from 1848 until 1981, and that influence from spoken Norwegian on the written standard is considered legitimate (Vikør 2001: 104).

The parsing process (by EDD) revealed multiple and inconsistent use of sources in volume 1-4, especially within the categories older versus newer sources and standard language versus rendering of speech. We needed to create a system that would (a) prevent wrong use of sources, (b) save time for new editors unfamiliar with the language collections.

As a result, a strict classification of all source materials was carried out, where each source was classified for age, genre and use within NO 2014. A database containing a reference bibliography of more than 5000 works for the UiO language collections existed before 2001. This database has been used to mark up each source according to its classification, and it is linked directly to the various source fields in the DWS application. The bibliographical classification is then used to extract specialised sub-bibliographies for f.i. etymological sources, historical sources, dialect sources etc, expressed in the DWS application as fixed menus, so that mistaken use of sources to a large extent is precluded.

Through our dictionary administrative system, the bibliography database is also used to advise editors on whether a word deserves an entry. If f. i. all sources for a word are (bilingual or special) dictionaries, or a word is shown to be a literary hapax legomenon, editing is not recommended.

Our current experience is that the internal control system offered through the bibliography database is popular with the editors because it saves them a lot of time and effort. Getting to know the sources used to be a long and slow process, and consistency in handling sources is hard to achieve. The integration of the bibliography database into the DWS speeds up editing and prevents mistakes.

An important section of NO 2014 written sources consists of dictionaries covering local or regional speech from after 1900, i.e. dialect dictionaries or glossaries. Information from these dictionaries can be listed with a reference to the place where the word is used, or giving the book itself as a source, depending on the category of information used. If this system should prove too complex, it can be tightened up through the bibliography database, but before we do that, we want to see that there is a problem that needs solving.

3. The purpose and logic of cross referencing

Historic and documentary dictionaries are often rich in cross references. Although the practice of explicit cross referencing was well controlled in NO 2014, the purpose and function of cross referencing had never been clearly defined.

In planning the restructuring of the dictionary format we also found that the database designers saw potential and needs for other types of cross referencing than the traditional ones, some of which have been integrated into the DWS application.

Cross references in NO 2014 – from, to, when and why		
	To point in tree structure in entry, i.e.	
	Head word (show head word and homograph number)	**Sense unit (show head word, homograph number and number of sense unit)**
From cross reference entry	a) from less important to more important standard form (where entry is found) b) from dialect variant to standard form	
From entry head	for irregular paradigms where each form has an entry, from inflected form to entry	
From Etymology table	Point to origin of derivations and constituting elements of compounds	
From definition text		Synonym definitions, hyperonyms
From cross reference field after definition		"compare"
From cross reference field after example		"compare"
From Compound list	Pointer from naked compound to (unprinted) database entry and digitized materials	

This table shows the cross reference system built into the DWS application. The "From" column to the left is the place where the cross reference appears. The column titles to the right say what points in the tree structure of an entry you can cross reference to, and for each type it is briefly indicated why this type is included.

Cross references now constitute direct links between entries and sense units in the database. Each entry and sense unit has its own id number. Traditional cross referencing of the "compare" type is minimised. Instead, we encourage editors to put more effort into writing good definitions. Cross referencing is now used primarily to secure that (a) etymological relations are clearly stated, (b) idioms and fixed phrases are defined under only one headword, (c) defining vocabulary is itself defined, and circular definitions avoided, (d) compounds which have to be excluded from the printed.dictionary are linked to their digital entries, which in turn link with the digitised collections that NO 2014 rest on.

The cross references covered by type c, i.e. implicit cross references embedded in definition text, are particularly important in an academic dictionary covering both dialectal variation and a written standard. A case in point is the range of names for common plants, all of which are defined by the official botanical name. The cross referencing system in the database shows cross referencing both ways – at the top of an entry's the tree structure , one can look at a list of entries that contain a cross reference to the entry in question, and thus get a view of f. i. all dialect names of a plant, a bird, together with its official name.

We also see a tendency among editors to use this function of implicit (invisible) cross referencing on the key word, the hyperonym of a definition. This is not something insisted on at present, but it is possible, and it can be inserted at any time. It is logical to use a dictionary lik our to build semantic hierarchies. This is one way in which the project can become useful in linguistic research beyond lexicography proper.

The cross referencing system is not the easiest part of the application to use. But it provides safeguards against cross referencing to non-existent or unprinted entries or sense units, and once a link is correctly entered it stays in place although the entry structure may be changed at either end. It also carries with it the possibility of overviews and insights that paper based editing fails to provide, and we considerate a constitutional part of our DWS.

4. Direct control of sorted and edited materials from the entry back to the digital archives

Ideally a historic and documentary dictionary like NO 2014 should be generated from below, from an exhaustive system of carefully classified individual items of linguistic

information, all from solid and verifiable sources. Further, the language to be described should be fully developed and thoroughly standardised, og of course exhaustively decribed in a huge meta literature from every possible angle.

This is not the case for Nynorsk. The written sources of the language are scanty and diverse, the influence from speech rich and contradictory, the orthography has been revised a number of times, and any real standardisation of the written language can only be looked for after 1945. In order to organise the collections more efficiently and speed up editing, NO 2014 has – together with EDD - created a headword index - the Meta dictionary - in the form of a database to which all electronic sources are linked, lemma by lemma. This database is expandable both as to posts and as to the number of sources that can be linked to it.

A thorough revision of the language collections via the Meta dictionary led to an entry reduction of about 20 % (from 0,7 to 0,55 million), and has proved an essential tool in organising the materials on which the dictionary is based.

The Meta dictionary, as well as our major individual collections, are digitised and freely available on the web, cf. URL below. It is therefore possible for all to check an entry in NO 2014 against available materials and evaluate the product.

It is also possible for editors to go straight from the editing format to the Meta dictionary entry, and look at each item of information as they edit. Once an entry has been generated in the NO 2014 database, a link has been created to the Meta dictionary.

However, No 2014 wants to take care of the work that editors do when they sort materials and structure their entries, and to make this hidden background work visible through the electronic version of the dictionary database (at present available only inhouse). Work on a semantic sorter has been going on for some time and will be implemented in 2005. This semantic sorter will allow linking each quotation and each item of information to its relevant sense unit in the entry. The NO 2014 corpus will be included in this system via the Meta Dictionary.

Conclusion – advantages, pitfalls and points to watch

The process of establishing a digitsl platform for all editorial work with the project NO 2014 has forced the project leadership to look at weaknesses and inconsistencies in the handling of source materials and of editorial practice, and to decide how to handle such problems on a "best practice" methodology. In a word, the digitisation process, combined with revised editorial rules, has forced the creation of a stringent editorial DWS application and more explicit editorial rules, which in turn has resulted in a more lucid and consistent dictionary. Furthermore, current experience suggests that Norsk

Ordbok can be ready in 2014 as planned, in spite of having to train twenty plus editors from scratch in three to four years.

The database system created for NO 2014 makes training of new editors a much easier task. Newly recruited editors become productive after a few months of training, and do not seem to feel daunted by the complexity of the project NO 2014. One third of volume 5 is written by editors who started training in the summer of 2003.

In brief, reworking the format of NO 2014 through passing former practice through the sieve of the DWS database designers has led to:
Clearer delimiting and desvription of linguistic categories
Firmer and more predictable formats
A more consistent and searchable dictionary
A dictonary that is easier to work with
More focus on the job that only properly trained editors can do, i.e. analyse and describe the materials from a linguistic and lexicographical point of view.

The chief pitfall for a project like NO 2014 is to lean back and leave design, solutions and testing to the software designers. One point is that project safety depends on inhouse mastery of the product that has been ordered. That is certainly important – you can't become a good cook if you stay out of the kitchen. But the really important loss would be to miss the intensive and critical overhaul of traditional assumptions and ideas about lexicography, linguistics and the art of categorization, which goes well beyond any individual academic discipline.

Literature

Vikør, Lars S. (2001): The Nordic languages. Their Status and Interrelations. 2. rev. ed. (1. ed. 1993, 2. ed. 1995). Novus, Oslo.
Grønvik, Oddrun (1992): The Earliest Dictionaries of Nynorsk in the Light of Presentday Dictionary Typology. Seventh International Conference of Nordic and General Linguistics 1989. In: The Nordic Languages and Modern Linguistics I-II p xx-xx.

URLs:

Norsk Ordbok 2014: http://no2014.uio.no/tekster/ordboka/index.php

NO2014 nynorskkorpus (tagged and lemmatized): http://folk.uio.no/danielr/tagged-nn-alpha.html