What should the electronic dictionary do for you – and how?

Oddrun Grønvik, Christian-Emil Smith Ore

University of Oslo, ILN, Pb 1021 Blindern 0315 Oslo
E-mail: oddrun.gronvik@iln.uio.no, c.e.s.ore@iln.uio.no

Abstract

Language is a common good and a common property. Access to information about language should be fast, easy, and intuitive. The electronic dictionary should therefore be a knowledge base with language as its access point, and with simple, yet rich access to (combinations of) linguistic and non-linguistic facts. One query frame and basic reading and writing skills must be enough to get meaningful results. This solution presupposes (1) a fine grained and systematic database format for dictionary storage and linkage to materials, and (2) a query system offering ease of access for inexperienced users. At the same time, lexicography must be able to prove itself trustworthy by offering access to sources both for usage and for normative decisions. The system described here is used for one academic multivolume dictionary and for standard monolingual students' dictionaries. It is suited to lexicographical projects where source documentation has priority. The focus is on dictionaries integrated with other language resources and produced for the Web.

Keywords: electronic dictionary, relation database, database linking, database entry format, the Meta Dictionary, full form register, indexing source materials, linking source materials to product.

1. Introduction

Electronic lexicography and language analysis is moving from the research and experimentation stage to becoming mainstream. In this setting, attempts are made to work out and present generic solutions. Our argument is that while important steps forward have been made, the present models for generic solutions are too limited, and in particular fail to take into account the importance of documentation as a method for building trust and consensus around lexicographic products.

The issues discussed in this paper are based on our experience with the electronic formats and solutions developed for *Norsk Ordbok* (NO) and the standard one-volume monolingual dictionaries *Bokmålsordboka* (BOB) and *Nynorskordboka* (NOB). We also draw on experience from projects aimed at promoting monolingual lexicography for African languages[1].

[1] the ALLEX Project (1991–2006) which dealt with the African Languages of Zimbabwe, and the CROBOL Project 2006–2011, which dealt with cross border languages involving Zimbabwe, Mozambique and South Africa.

A model for lexicography encompassing

- collecting materials
- analysing materials
- writing dictionary entries
- supervising flow
- presenting the finished product in an optimally accessible fashion is enough in a language community where
- the written standard is fixed and has been more or less unchanged for a long time
- there are plenty of materials documenting the standard through a long time span
- the community is used to using dictionaries
- the community is used to trusting its dictionaries (and there are plenty of them for comparison)

The model above is in short a sufficient model for language communities where there is general agreement on what the written standard looks like and how it is used. Dictionary making can then build on a general consensus concerning the object to be described, which is the lexicon of the language in question.

This is the situation for many of the world's major languages, especially for an important group of European languages.

A recent and very good lexicographical handbook, the *Oxford Guide to Practical Lexicography* (Atkins and Rundell, 2008) presents a model for dictionary projects suitable for dictionary making language communities of this kind. A similar understanding of lexicographical needs underlies Schryver (Schryver, 2011) in his presentation of *TshwaneLex*.

We argue that in many of the world's language communities, this model is insufficient, because it assumes trust, instead of including mechanisms that pre-empt distrust. There are plenty of dictionaries that look trustworthy, and could be produced following this model to the last letter, but are skewed in their selection of materials and lemmata, are incomplete in their presentation of orthography and word senses, and so on.

The reasons for skewed lexicography may be ideological (promoting one particular world view) or in favour of a certain language variant (presented as valid for the whole language community). It may also have to do with the ease of production (imposing a standard and omitting variants for languages which do not have a written standard). The result can very easily be general distrust, not of a certain dictionary,

but of dictionaries and reference works in general. So, as language is so important to people, lexicographers need to be trusted – not as missionaries of a particular cause, but as providers of facts of life.

The only way of dealing with distrust and building trust in linguistic reference works, is to take suspicion and the need for external control for granted, by integrating access to the raw materials (for the whole, and for each entry and sense) into the dictionary model itself. Access to the lexicographical sources has to be easy to obtain and easy to understand, from the Web.

We therefore propose a model encompassing the following stages:

1. Collecting and preparing materials (including referencing and marking)
2. Indexing materials to collect variant forms
3. Generating entries from indexed materials, with a link to the materials
4. Analysing linked materials
5. Generating entry head from a separate full form register
6. Writing dictionary entries, linking materials to each sense
7. Supervising flow
8. Presenting the finished product in an optimally accessible fashion
9. Using a staged search system that first searches the headword register, then other fields

This model is an ideal. In the following we will base our argument on the collective experience of the Norwegian Dictionary 2014 project. Most of the examples are taken from this project[2]. *The Norwegian Dictionary* NO aims at providing a scholarly and exhaustive account of the vocabulary of Norwegian dialects from 1600 to the present and of the written standard Nynorsk since 1853.

A common challenge in editing historical and dialect dictionaries is the heterogeneity of the source material. Since NO covers sources for speech and writing through 400 years, this heterogeneity must be handled both diachronically and synchronically. The source material spans from modern texts, via traditional paper slips to local dialect dictionaries and word lists dating back to the 17th century. The interpretation and use of these materials call for explicit referencing and preferably linking to the source material so that users can check the basis for the editors' conclusions.

[2] The Project *Norsk Ordbok 2014* (The Norwegian Dictionary) to be completed in 12 volumes in 2014.

2. Collecting and preparing materials

In all modern introductions to lexicography the text corpus is presented as the chief electronic source. In our case, the digital sources are of several kinds[3]. Materials in electronic form can include images of for instance manuscript pages, and their transcripts. For languages with a weak standardization or with several orthographies it is not a trivial task to build a lemmatized and POS tagged corpus. To be able to include all texts in a homogeneous corpus one has to encode the text at three levels: The original word form, a standardized word form and a lemma form. The two latter have to be taken from an orthographical standard chosen for the entire corpus. This process is hard to computerize and is therefore very resource demanding. For reference, check the Menota guidelines for medieval Nordic texts (www.menota.org). Norwegian orthography has been thoroughly revised several times during the last 150 years. A POS-tagger developed for modern Norwegian has a very low success rate for text from the first half of the 19th century. Therefore, only the modern part of our text corpus[4] is lemmatized and given a POS mark-up. This is clearly not a problem confined to Norwegian. This is a problem in creating corpora for all languages with changing orthography over time or for weakly standardized languages.

A second challenge is source material which is not running text, e.g. slip archives and older dictionaries and word lists. Including already synthesized information in the source material of a dictionary project obviously requires great caution, and deep philological expertise. The editorial text of old dictionaries may not be written in the language to be documented, e.g. in our case the editorial texts are in Danish or occasionally Latin. When the running text of these sources is made available electronically, the sources are not included as corpus text, but stored and referenced to the indexing system for the electronic language collections, see below.

3. Indexing materials to collect variant forms

For highly standardized languages like the major modern European languages, a lemmatized and POS-tagged text corpus stored in a standard corpus system gives an excellent and coherent access to the source material. For the less standardized languages with many heterogeneous sources a common indexing system is needed to group variant forms according to the standard that will be used for the headword of a dictionary. This is equally important whether the task is collating forms in ancient manuscripts or attempting to standardize a language for the first time.

In the case of NO, a common indexing system called *Metaordboka* (the Meta

[3] The Norwegian language collections, dating back to the 1930s, were computerized in the 1990s.

[4] Texts published after 1938 comprise the modern part of the corpus, about 85 % of the total 90 mill.

Dictionary) (MO) was designed (see Ore, 1999 and Ore & Ore, 2010). The original motivation was to create a common web-based interface to the huge lexicographic materials digitized in the 1990s. MO was later redesigned to become a pivot in the combined source database, text corpus and editing system for NO. An index entry in the MO can be seen as a folder containing pointers to (possibly commented) samples of word usage and word descriptions found in the linked sources. Each entry is labelled with a normalized headword, POS information and the source word form. The linked sources cover the ground from glossaries compiled for the Danish state administration in the 17th and 18th centuries to modern dialect surveys and local dictionaries. The MO has proved itself a very useful tool in the practical editing of NO, as well as an invaluable tool in managing the Norwegian standard language Nynorsk.

For NO the task of collating variant speech and written forms to index forms in the MO includes adding POS information, so that identically-spelt lemmata with different POS get separate entries. Index forms of compounds are marked to show joins, very important in dealing with a compounding language like Norwegian:

headword	**POS**	**Status**	**Nr**
fisk*e*saks	noun fem	recent	1
fisk*e*sal*s*lag	noun masc	OK	4

Figure 1: The Meta Dictionary - normalization categories.

The join marks facilitate searching for end and middle parts of compounds, to keep an eye on productivity, semantic developments etc.

MO is an independent system component that can be linked to many different lexicographical projects. It has in itself become a valuable repository. The old and the local dictionaries are kept in their original form as individual works expressing the language view of their time and author. The bidirectional linking in the system makes each headword in a source an entry point to the entire system (including NO), thus enabling dialect users a unique opportunity to see their dialect in the larger context.

All the collections coordinated under MO as the source material index are searchable in themselves. Some have the standard form of their lemmata as part of their original information, as mentioned above. Many do not, and are standardized only through their link to the MO. Both synchronic variation and diachronic heterogeneity can be a challenge, as shown below:

kjiru, kjuru, kjyru, kjære, tjere, tjære tjøre, tjyru, tjörru

Figure 2: The Meta Dictionary - headword forms found in directly indexed materials for the noun *tjøre,* 'tar'.

The language collections coordinated through MO are under constant maintenance. One index entry can have several thousand items connected to it. In the standardization frame, index entries show standardization level by their status, cf. the Status column shown in Figure 1. Items can be moved from one index entry to another using "cut" and "paste". The MO is a very flexible tool, and looking after it is a specialized skill, closely allied to work with language standardization in general. MO is an important source of information for the Language Council in Norway, the state agency that deals with language issues[5], and is accessible on the Web for the general public.

4. Generating entries from indexed materials, with a link to the materials

An important aspect in trustworthy dictionary databases is that it should not be possible to create entries with no source bound materials showing form and usage. The dictionary databases of the Norwegian language collections do not permit the generation of a new entry unless it is linked to an index entry with adequate materials behind it.

If editors encounter unedited and undocumented lemmata that should be included in the dictionary, they first have to collect and register the documentation in MO, as a corpus text or as one or more electronic excerpts.

In the NO2014 bibliography[6], sources are marked for genre and other qualities. The marking is used to generate advice to editors on whether a lemma merits an entry. If an entry in the MO f.i. is documented only in one work of fiction (a literary hapax), the advice will be not to include it. If it is documented only in older standard dictionaries, the advice will be the same. The editor can overrule this advice, or change it by adding better materials to MO.

5. Analyzing linked materials

We agree with Atkins and Rundell that the linguistic information contained in the documentation for each entry needs to be analyzed, and that the analysis needs to be conserved for future (re)use (Atkins and Rundell, 2008: 98 f.). We do not agree that analysis should be a separate task from editing. The editor needs to do both. This is of particular importance if language standardization is a permanent task. In NO, many lemmata are described in a dictionary entry for the first time.

If the dictionary source material is a giant corpus, ensuring at least 500 usage

[5] See http://www.sprakradet.no/

[6] Yet another independent but linked database, drawing its bibliographical information from the Norwegian National Library

examples of each lemma qualifying for entry (Atkins and Rundell, 2008), running a statistical analysis on them all is an obvious course of action.

This is something we would like to try, but only for a very small part of the 300,000 lemmata to be edited in NO. Since the language we deal with, Norwegian, is a compounding language with a medium rich inflection system, the section of the language collections occurring 500 times or more is much smaller than for English, be it word forms or lemmata. In a corpus of 90 million tokens, only about 1% of tokens occur 500 times or more, and well over 50% of tokens are single occurrences (hapax forms). Of more than 570,000 entries in the MO, fewer entries (i.e. lemmata) than 1:1000 have 500 or more items of documentation, while roughly 50% are (as yet) hapax forms. Many of the hapax forms culled from older materials require careful analysis in themselves, to decide their status and possible affiliation to already identified vocabulary.

What we do have is a corpus function that will give us real numbers of occurrences, with concordances and expanded text excerpts. A search argument like this:

"sus.*"

will produce a frequency sorted list of all word forms starting with *sus-* plus the two following words. It is a very useful function[7], even if numbers are small:

sus i serken	16
sus og dus	14
suset frå pisserenna	10
sus i lufta	9
suste inn i	8
susar av garde	7

Figure 3: *Nynorskkorpuset* - Search result.

Our current solution for analysing data is a database, called “the sorter”. It is separate from, but linked to both MO and NO. In what it offers, it is a great deal less sophisticated than a lexical profiling tool (Atkins and Rundell, 2008: 91–92 and 107 f.), but is undergoing improvement. In the sorter, the editor generates a list of links to all instances linked to the MO entry, served up in a spreadsheet. The instances can be annotated and sorted, spread on several work sheets etc. The sorter has proved suitable as a note block for dealing with fringe materials (old, rare or poorly documented word forms). A sorter can have as many work sheets as the editor wants. The sorters are saved and stay linked to their entries. Sorters (with lists of instances) can also be moved to other entries, if materials are found to be misplaced.

[7] The work of Dr. Daniel Ridings, who is in charge of *Nynorskkorpuset*.

Once sorted, documentation items can be linked directly to the relevant piece of information in the entry, be it dialect form, back up for definition or usage example (comprising both generic examples showing f.i. valence, and full citations). See Figure 4.

6. Generating entry heads from a separate full form register

In a dictionary offering information on spelling and inflection, entry heads traditionally present this information in a condensed form with extensive use of codes and abbreviations. Norwegian is a compounding language, as are most Germanic languages, with a medium rich inflection system[8]. In most Norwegian paper dictionaries compounds have no POS and inflection information since a compound has the same POS and inflection as the final part of the compound. It is assumed that all native Norwegian speakers can analyze compounds. This assumption has proved useful, given the space limitations of a printed dictionary. In an electronic dictionary space is not a problem – nor is it true that all Norwegian speakers can analyze compounds.

However, full inflection tables in the entry head as a first option are not a good idea. They should be shown on request.

The information on POS and inflection has to be accurate, complete and in accordance with school requirements. In the Nordic countries, publicly funded Language Councils are tasked with providing this mass of detail in a comprehensible fashion. Due to the complex spelling rules of Norwegian, with a large number of alternative forms and frequent spelling adjustments, this has been a daunting task. A complete, detailed overview of official standard Norwegian spelling (including all inflected forms) was a by product of the first edition of NOB and BOB. Today, a quality checked database, a word bank, exists for both written standards.

The Word Bank is based on an extension of a spellchecker made by IBM in the 1980s (Engh, 1993). The central idea is to link each lemma to one or more inflection patterns which in turn produce all possible forms. This process will cause the generation of possible but undocumented word forms. These forms are useful for the POS-tagger in which they are used, but not for human users. To avoid generating spurious forms and also to ensure that each set of inflected forms is in accordance with official orthography, additional information is added to the links between a lemma and inflection paradigms. For each link, validity level (unknown, variant form, norm) and the time span for this status, is listed.

[8] Nouns for instance have four forms, eight if genitive forms are included: more than English, less than German.

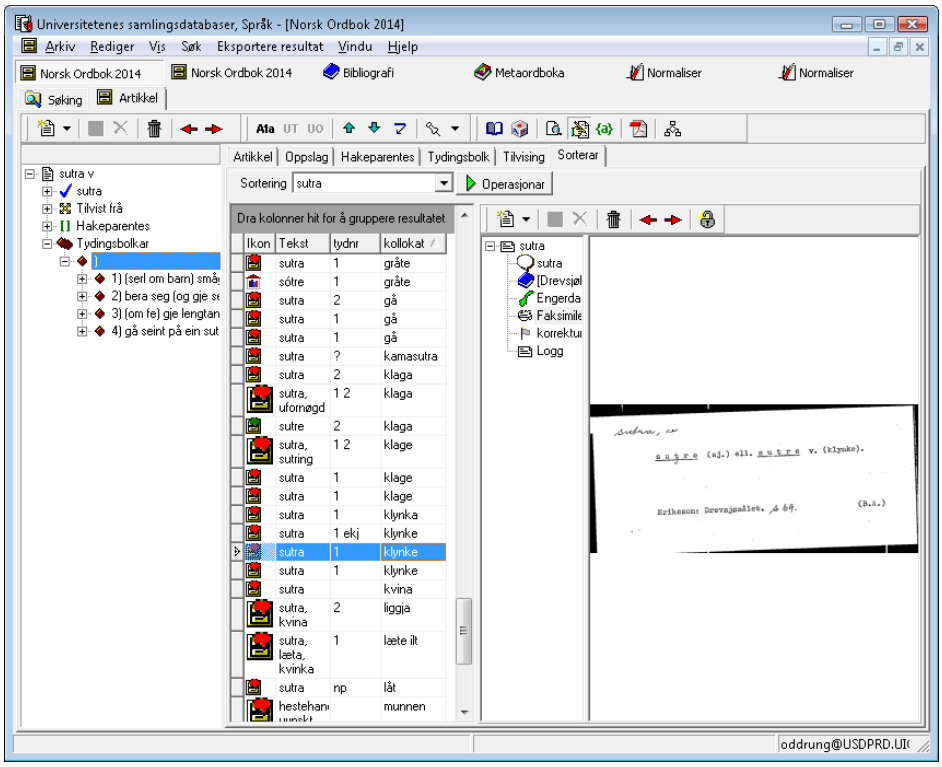

Figure 4: The sorter with list of sources, with entry in tree form to the left and image from slip archive to the right.

Currently, the Word Bank contains information for the time period 1996 to the present. It is possible to generate a valid orthography for any year in this period. An important feature of this system is that it can be used to "wash" lemma lists. The Word Bank has f.i. been used to check the Norwegian part of an Icelandic to Danish, Norwegian and Swedish web dictionary. This exercise turned out to be very useful.

The Word Bank can be used to generate the entry heads of a dictionary. No two Word Bank lemmata have the same set of paradigms and the same status history, but they are not separated with respect to homonyms beyond this point. Separate homographs have a strong tradition in Scandinavian lexicography. Thus one single lemma in the Word Bank may be linked to several lemmata in a dictionary.

Below, we show three examples of how POS and inflection was shown in a standard paper dictionary of Nynorsk from 2005:

rope v1 el. v2

II skru el. **II skrue** v1 el. *-r, -dde, -dd* el. *-tt* el. **II skruve** v1

I søkje el **søke** *-r, -kte, -kt*

Orthographic information in the form of codes and abbreviated forms is no longer acceptable in teaching, and the Web has freed the editors from the need to save space at every turn.

In the new web edition, all headwords of the two standard orthographic dictionaries BOB and NOB are linked to the entries in the Word Bank. The entry head of (web version) is now generated from the Word Bank, in schemas shaped according to school and Language Council requirements. The entry is shown with the headword followed by POS information. A click on the POS information opens a new window with a schema showing the inflection pattern(s) for the word in question (Figure 5).

rope **v1 v2 v3** (truleg frå ty jamfør norr *hrópa* 'baktale')
bruke sterk røyst,**;** skrike, kalle;
varsle med visse ord
rope om hjelp / rope hurra / rope på nokon / rope opp (namn, nummer på ei liste) / rope noko ut / som ein roper i skogen får ein svar; sjå ***skog (1)***

Figure 5: NOB new website - the entry *rope* v with POS plus codes for inflection.

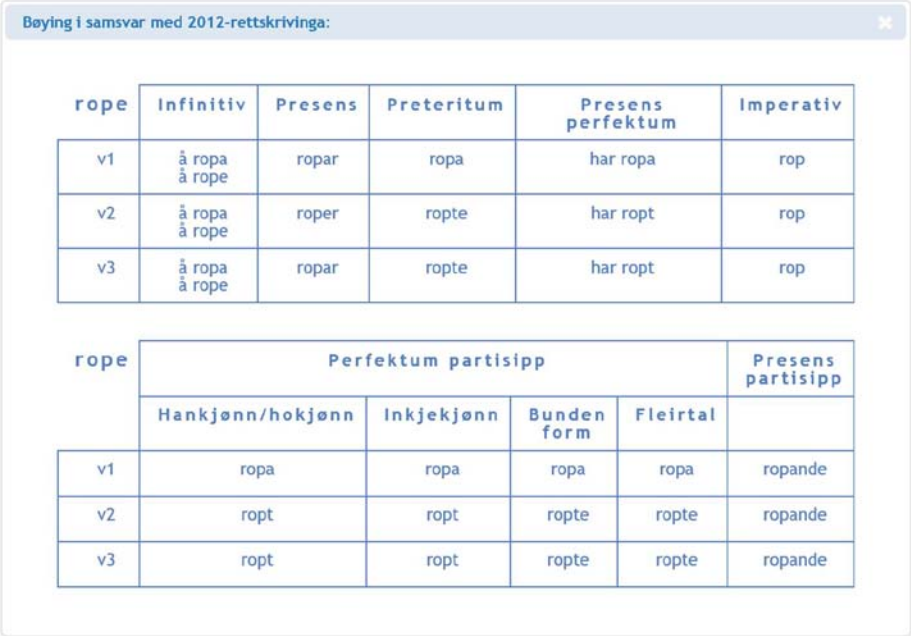

Figure 6: NOB - form showing inflection paradigms for *rope*.

This solution was launched last autumn and has proved a success with users.[9] It is clear that it is complete for each lemma or lemma variant, and it encompasses the entire vocabulary in the dictionary in question. This solution for presenting inflection data can be implemented for any dictionary that is linked to the index MO. As a general feature this solution would be a great improvement for learner dictionaries on the Web.

[9] The evidence for this statement is twofold: The feature is frequently used, and correspondence with users through Ordvakta

7. Writing dictionary entries, linking materials to each sense

Once the materials for a headword are analyzed, the entry gets written. The editorial interface shows the entries in three formats, (1) a tree structure (to the left), (2) a viewer showing the entry as xml text, and (3) a set of forms for editing the entry and managing the MO materials ('entry administration', 'entry head', 'form information', 'sense unit', 'cross reference' and 'sorter').

The sense unit form is where defining and entering usage examples happens. This form also has links to the bibliography and the location register, fields for cross referencing, etc. A particular feature is the compound table which allows editors to give instances of compounds where the sense shown in the definition is applicable. Compounds included in the compound table are linked to MO, which means that their usage is documented.

The sorter is linked to the entry and can be made searchable from the Web. However, it is also possible to link individual items of documentation directly to any node in the entry tree. In Figure 7 a link has been added to the synonym "drynja" (see arrow and boxes). A click on the "Belegg" icon leads straight to the image of the original slip. Currently these pointers to the material are mostly inserted for the benefit of colleagues, and typically added to convince doubters or as aids to the editors' memory. However, there is nothing to stop general access to such links.

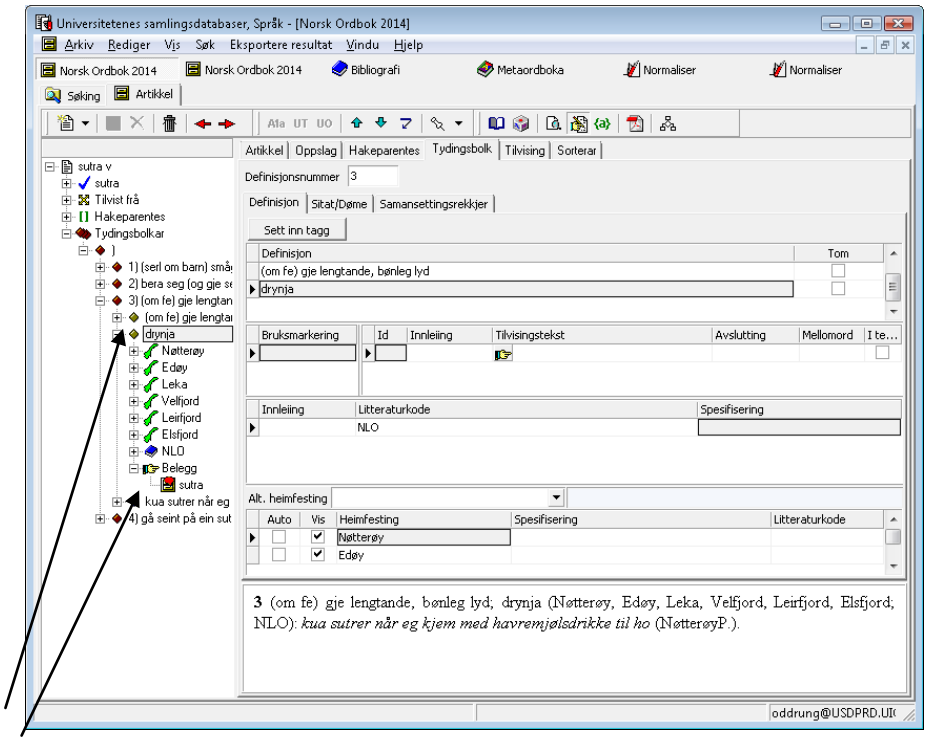

Figure 7. NO editor interface, sense unit form. Arrows show word in definition and its link to the source ("Belegg", red ikon).

Why is it important to have this possibility? Editing a historical dictionary based on materials from Norwegian dialects and Nynorsk, a written standard with a short history, is bound to cause discoveries that break with general preconceptions about language. We mention one in particular: Words associated with "slang", "street language" and other frowned-upon innovations from young people in urban areas often turn out to be dialect variants of words well known in vernacular Norwegian from wide tracts of the country, or standard derivations from such words[10]. Some of them are attested back to Old Norse. When the hoodie turns out to be a preserver of old lexical items, one needs easy access to sources to be believed. Our experience in codifying languages with limited literary documentation and presenting them in dictionaries, has shown us that people very often believe their dialect forms to be unique to their own area. They never use these word forms away from home and will not be aware of their being part of the general vocabulary in the country. In such cases, easy access to documentation is essential.

8. Supervising production flow

Dictionary production is to a large extent a matter of managing time and money. There is no reason why a dictionary project should have poorer progress management than any other kind of project. For ease of administration, the system for supervising production flow is inbuilt in the database package set out in the introduction.

The management devices built into the administrative system is in part a result of what has been known to go wrong in previous large dictionary projects, partly a result of new possibilities when NO in 2003 moved to a digital platform. We will here comment on the management of size, status and storage.

The standard failing of older, paper-bound projects is that entries get longer and longer, and also take longer to produce, so that while manuscript production rockets, alphabet progression grinds to a halt. Our system for supervising size is therefore geared towards ensuring alphabet progression, and proper distribution of entry length within alphabet sections. Editorial work is measured in a given amount of finished manuscript per month. When an entry is generated, a maximum size is suggested, based on the amount of documentation available at the time of generation. Real size is measured against maximum size of the entry throughout editing. The editor can overrule the maximum size for individual entries, but the size of the alphabet section is fixed.

Data concerning production flow is shown in connection with each entry in the form "artikkel" ('entry administration'). Figure 8 shows the subform dealing with size management, with the maximum number of lines and the present line count of edited text outlined.

[10] Examples are verbs *loka* 'hang (aimlessly) around' and *køddа* 'joke, "take the mickey"'.

All change in the dictionary database is logged with name, date and status change. The project management draws out reports every month to see manuscript progress, and while individual progress is always a matter between editor and management; the whole staff knows the exact state of progress per volume in moving manuscript along from draft through several control and correction stages to finished, publishable text. This supervision system combined with the possibility of generating a print version in PDF, promotes both efficiency and job satisfaction, since it is easy to see both from reports and from the dictionary database itself exactly how much one does. As work on NO also counts as scientific production for each editor in the University of Oslo crediting system, an exact count of lines and pages is very important.

The third point concerns the vulnerability of a project as large as NO, where one lost day means the loss of 1.5 man months, and where processed detail can be hard to recapitulate. Dictionary manuscript is stored in the database. Backups are taken every night, and stored. This ensures the project against production losses bigger than that of one working day, but it also means that it is possible to take care of the long version of an entry that needs to be shortened, or reinstatean entry that got deleted by mistake. The XML and HTML presentation of entries is synchronized with the editing. From the XML version, proofs with the correct typesetting are produced as PDF documents.

9. Presenting the finished product in an optimally accessible fashion

The dictionaries BOB and NOB have been searchable as a free web service since 1994. The website was thoroughly upgraded in 2009, with a view to making it visually appealing, especially for school use. The database solutions were thoroughly upgraded in 2012–2013. NO appeared on the Web in March 2012, as a by-product of the printed dictionary. This was possible on a tight budget because the databases have XML-presentation of entries built into the standard production format.

The finished product is the entry as it is presented on the Web, and web lay-out should be as clear as possible. This includes presenting the information most often sought up front, and hiding less popular items behind icons or codes. At the NO website, information on language variants is hidden behind a row of icons above the sense units. The dictionaries BOB and NOB are built on the language collections, but are not directly linked to them. Every entry is, however, directly linked to the Word Bank, and when users look up grammatical information, they are looking into the Word Bank full form lists for that particular lemma.

At present it is not possible to go directly to sources from the web presentation of NO. Source reference detail (bibliography, location) appears to be fixed to the right of the dictionary text.

This does not mean that the sources are inaccessible. In the case of NO, the language collections had been accessible on the Web for more than a decade before the dictionary itself appeared there, and links to the different sources are to be found on the home page of NO. The collections are well known amongst professional linguists and interested amateurs, and represent an important channel to public interest.

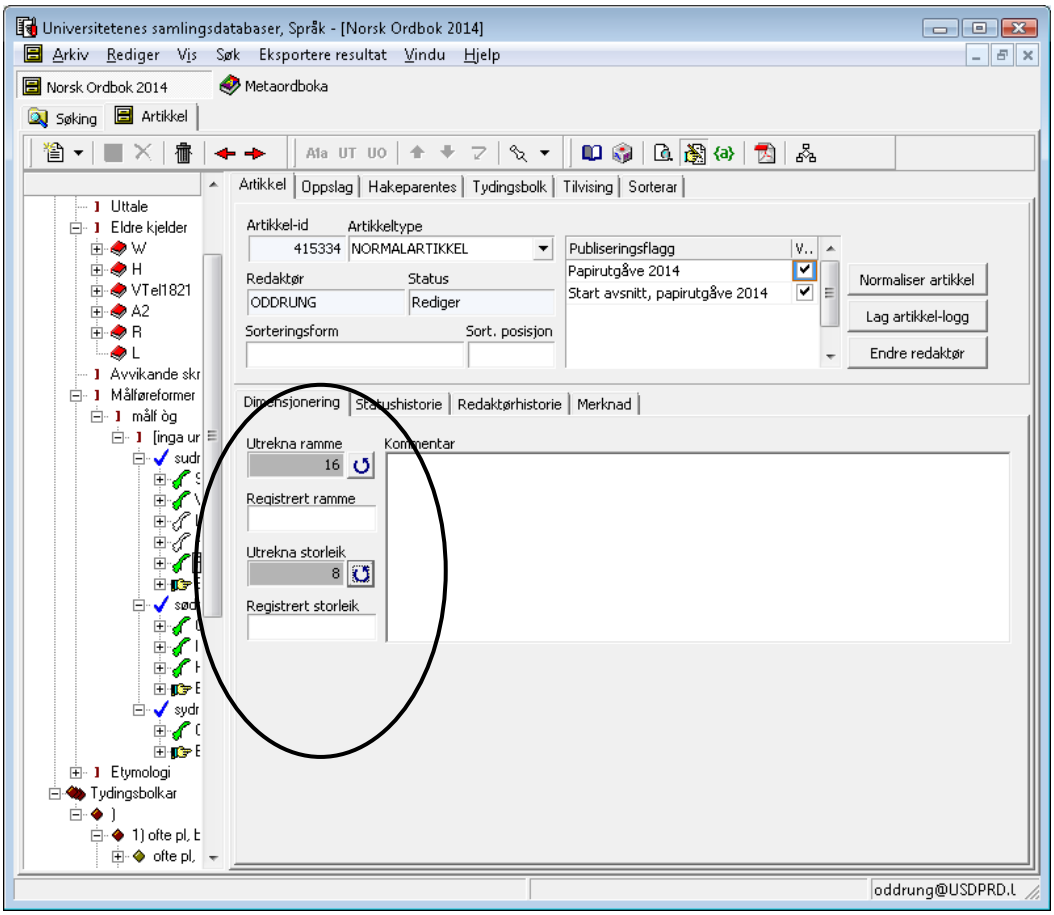

Figure 8: Administration page in the NO editing interface, with maximum number of lines allowed and the estimated number of lines required indicated, cf. section 8 above.

10. Simple search systems for complex databases

A great deal has been written about user-friendly access to web resources, including dictionaries and other language resources. A sort of scale seems to have emerged for solutions. At one extreme one finds the Google-type box where the user writes his search argument and then goes on to refine it, depending on the results. At the other end one finds solutions which require user profiling as a first step[11], so that the database can direct its user to the supposedly most relevant results. In between there are endless possibilities and combinations.

[11] Please note that this kind of user profiling is not the same as planning what sort of user one expects a dictionary to have, as in Atkins and Rundell, 2008 p. 486 f.

The user profiling approach has been promoted by Bergenholz and Tarp [see f.i. Fuertes-Olivera, 2009 p. 132 f.) in connection with their functional theory of lexicography. The idea is to create a lexicographical database as a multifunctional dictionary, with sophistication and detail in the entry increasing according to active choices in self profiling made by the user. They see the lexicographical database as a knowledge base containing multiple dictionaries from which virtual dictionaries, specialized according to the user's (self-described) profile and assumed needs, can be queried.

In the age of Google one may ask if this is a good idea. Our impression is that people tend to use Google and other search tools as data mining tools. A general search is iteratively narrowed until the required information is found. Under this assumption an electronic dictionary should be wide open to Google and other search engines. It is important that when a Google hit is clicked, the user reaches a web page which make the context clear and which offers the user a more detailed search in the dictionary.

On the other hand an electronic dictionary should offer its own search interface. We have seen that complex search forms scare away users. A simple search field should be standard. One can, however, include advanced search strategies in a simple field.

For the two standard monolingual dictionaries BOB and NOB a four step search strategy is implemented. First of all, an auto-complete function is attached to the search field. This gives a quick overview of possible headwords. Combined with wild-characters (truncated searches) this serves as an excellent tool for crossword and Scrabble. Multiword expressions (treated as sub-entries) are included in the headword search. If a headword search does not produce results, the search continues to the full form lists in the Word Bank. If there are no hits there, the search continues to the full text of the dictionary.

We think that queries going through several set stages could be useful in searching NO as well. One possible combination would be 1 headword field, 2 definition field, 3 usage example field (comprising both standardized examples and citations). Another possibility, for advanced searches, would be to extend the search to the source material linked through MO.

Active editors of the NO system have access to the whole of the category system in the linked databases, can put together their own searches, and store results as lists or export them as excel workbooks. The editors have had this possibility since 2003 and they use it actively in support of editorial work, or other information needs. However, this would be beyond the needs of the average dictionary user.

11. One database format - several dictionaries

The database system created for NO was in 2011–2012 utilized for the one volume standard dictionaries BOB and NOB, without any adaptations to the software. This

was not only possible, but completely painless, because the database for NO was created as a maximum format, catering for all the documentation and verification needs of a large academic dictionary with the task of working its way through heterogeneous language collections for the first time, and with a high academic standard to its referencing system, dealing with both written and spoken sources.

Before designing this maximum format the project tried going the other way, i.e. using and expanding existing software designed for a smaller dictionary. It didn't work because the framework was too cramped. We learnt from this experience for instance that speed in a very large and rich database system has to be planned for right from the start, as keeping the highways free is an important aspect of information architecture.

The NO database has four types of entry: standard, prefix, suffix, and cross reference. In addition there is an entry format for multiword expressions, for use within the standard entry. The smaller dictionaries did not have these types of entries, but they could be identified by text criteria (suffix entries having head words starting with a hyphen etc.).

The fact that the database system already had well defined, different formats for different types of entry, simplified the work with NOB and BOB. Two examples: (1) Affix entries do not have usage examples. What they do have are little lists of derived words demonstrating the use of the affix in question. Those derived words now exist in the dictionary database as a sort of minimal entry: they were picked out, got their full form entry in the Word Bank and are linked to the affix entry. (2) In between the usage examples of the NOB, there were also a number of multiword expressions masquerading as usage examples with a comment added. All usage examples with explanations attached were picked out and about 5000 selected for the multiword expression type of entry, with minimal textual adjustments.

12. Some comments on information architecture

When computers and ICT in general were introduced into lexicography several decades ago, computer specialists, as well as many lexicographers, started to talk about dictionaries as databases or knowledge systems. This is not really true, since dictionaries are written as structured texts for human users. Lexicographers used these terms metaphorically while the ICT-specialists saw the potential of extracting information into a relational database from what appeared to be highly structured texts.

The introduction of SGML and later XML technology represents a compromise. The use of XML in dictionary writing systems requires that every dictionary entry has to have a tree-like structure defined by a formal grammar. This is handy for most new dictionaries, but in order to fit older dictionaries into such a structure, a thorough

editing and restructuring of the text may be required. This fact was borne out by the revision process necessary in order to move NO on to a digital platform in 2003 (Grønvik, 2005).

It is often argued that the XML approach is superior to relational databases. This is in many ways a false debate. Most dictionary writing systems (DWS) are a mix. The entries are stored as XML-documents in a relational database and edited in an XML-editor. This gives flexibility, and it is easy to store many different dictionaries in a single system. XML is a format for manipulating and storing structured texts. It is not designed for active linked data. Thus, in the case of NO, where one has a set of heavily interlinked resources, the XML-approach is not sufficient. It is better and easier to decompose the entry text into a relational table structure to ensure data integrity. It is easy to produce XML from a relational database and in the versioning system the entries are stored as XML-documents. XML technology is also used for publishing PDF and HTML for the Web.

13. Conclusion

Everyone must be in favour of generic solutions for dictionary making, provided that the generic solution really covers every need. But a generic DWS must take into account the need to link dictionary text to sources through the database system itself. The need for control and verification is general, and in many cases essential, in showing that the dictionary really is the consensus product its editors set out to make it.

Once done, source linking is also very labour-saving. A click on the screen replaces a trip to the library or searching through archives and bookshelves. In Norway, the Word Bank is freely available for download. With a full form register and a truly generic DWS that can stay linked to its sources, many dictionary writers should find themselves in clover, and dictionary users will be able to see what their own dictionary is built on.

14. Acknowledgements

It should be emphasized that the very rich information architecture for lexicography described above has been shaped in response to input from a large working environment of lexicographers at the University of Oslo, and from important external users, all of whom are hereby thanked and acknowledged. All software development has been done by the Unit of Digital Documentation at the University of Oslo (EDD).

15. References

Atkins, S.B.T. and Rundell, M (2008): *The Oxford Guide to Practical Lexicography* Oxford; Oxford University Press

BOB = Wangensteen, B. et al. *Bokmålsordboka. Definisjons- og rettskrivingsordbok*. (1986-. 4. paper ed. 2006) Oslo; Universitetsforlaget. New web edition 2013. http://www.nob-ordbok.uio.no/

Engh, J. (1993): Linguistic Normalisation in Language Industry. Some Normative and Descriptive Aspects of Dictionary Development. In: *Hermes, Journal of Linguistics*. no. 10 p. 53-64. http://download2.hermes.asb.dk/

Fuertes-Olivera, P.A. & Bergenholtz, H. (red.) (2011): *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum.

Grønvik, O. (2005): Norsk Ordbok 2014 from manuscript to database - Standard Gains and Growing Pains. In Papers in *Computational Lexicography Complex 2005*. Budapest: Linguistics Institute, Hungarian Academy of Science. s. 60-70

Menota guidelines for medieval Nordic texts (www.menota.org).

MO = *Metaordboka* (The Meta Dictionary) http://www.edd.uio.no/perl/search/search.cgi?tabid=571&appid=7

NO = Hellevik, A. et al. (eds) *Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet* (1950–) Oslo; Det norske Samlaget (Web edition: http://no2014.uio.no)

NOB = Hovdenak, M. et al. *Nynorskordboka. Definisjons- og rettskrivingsordbok.* (1986-. 4. paper ed. 2006.) Oslo; Det norske samlaget (New web edition 2012. http://www.nob-ordbok.uio.no/

Nynorskkorpuset http://www.muspro.uio.no/NO2014nynorskkorpus/conc_enkeltsok.htm

Ore, C.-E. Metaordboken - et rammeverk for Norsk Ordbok In *Nordiska studier i leksikografi 5. Rapport från Konferens om lexikografi i Norden, Göteborg 27-29 maj, 1999*. Göteborg: Nordiska föreningen för leksikografi.

Ore , C.-E., Ore E., *Re-linking a Dictionary Universe or the Metadictionary Ten Years Later* Presentation at *Digital Humanities 2010*, King's College London, UK. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-786.html

Schryver G.M. de (2011): *Why Opting for a Dedicated, Professional, Off-the-shelf Dictionary Writing System Matters.* http://tshwanedje.com/publications/asialex2011.PDF

Setelarkivet (The Norwegian language collections - Nynorsk) http://www.edd.uio.no/perl/search/search.cgi?tabid=436&appid=8

TLex Lexicography and Terminology Software http://tshwanedje.com/tshwanelex/