The Dictionary and its Sources: the Ideal of Integration and the Example *Norsk Ordbok*

Kristin Bakken
Norsk Ordbok 2014
P.boks 1021 Blindern
0315 Oslo

Abstract
The paper shows how the digitalization process that the Norwegian historical dictionary, *Norsk Ordbok,* has gone through, not only has improved the editorial routines and the editors' efficiency; the process has also given the dictionary a technical platform that integrates the dictionary with its sources. The different aspects of the integrated application that has been developed for the project Norsk Ordbok 2014 by the Unit of digital documentation at The University of Oslo, are presented in the paper. In conclusion this integration is discussed in a broader perspective. The author comments on the scientific implications of the integration, on the pragmatic aspects of it as seen from the users' perspective, and most importantly she comments on the way this integration has the potential of reconceptualising and rejuvenating the genre of the historical dictionary.

1 Introduction

When the task of reorganizing *Norsk Ordbok* was given to me in 2002, my challenge was simply to speed up production time, thus completing a dictionary which by then had been under way for 72 years. The financial means to reach the goal of completion was granted to us by the Norwegian parliament, but the funding was not sufficient if one did not also improve the staff's efficiency. In the proposition that prepared our case for the politicians, digitalization was mentioned as one of the means to obtain improved efficiency. In the years that have passed since 2002, we have come a long way along the route of digitalization, and we have found that our dictionary making *has* become more efficient. In this paper, however, I will discuss a side effect of the digitalization process that we did not express as an explicit goal in the beginning, but one that we now see have the potential of making us redefine our concept of a scientific historical dictionary. I will go on to discuss the relationship between the dictionary and its sources in the age of digitalization, and I will use our experiences with *Norsk Ordbok* as my point of orientation.

2 The digitalization process of the project Norsk Ordbok 2014

The first phase in our digitalization process is probably typical of any large scientific historical dictionary's endeavours into this field. The starting point was the digitalization of our slip archive. It ought not to be controversial to say that the perusal of the old slip archives provides a time consuming hedge for editors of documentary historical dictionaries. At the

last Euralex conference in Lorient (2004) I paid special notice to Adam Kilgariff when he argued for "The Sketch Engine" in this way:

> The more data we have, the better placed we are to present a complete and accurate account of a word's behaviour. But it does present certain problems. Given fifty corpus occurrences of a word, the lexicographer can, simply, read them. If there are five hundred, it is still a possibility but it might well take longer than an editorial schedule permits. When there are five thousand, it is no longer at all viable. (Kilgariff et al. 2004:106)

The old historical dictionaries which are based on meticulously assembled slip archives, are bound by these archives, and their accessability poses a real problem for the editors who have no means of "summarizing" these data as the corpus based Sketch Engine does. Consequently the editors of such dictionaries not only read 500 slip occurrences of a word on a regular basis, they often read 5000 and more, and note that these occurrences have to be physically and separately handled as they only occur on paper slips. It is no wonder that the editors and thus the old historical dictionaries themselves may "drown" in their own archives, and it is ironical that the bigger (and thus better) archives, the bigger the hazard of getting lost. In Scandinavia one current and illustrative example is the Swedish dialect dictionary that has an archive of 8 million slips, but which is now terminated due to slow progress after 50 years of work and one volume published. The Historical dictionary of Hungarian provides another interesting example, where the slip archive at a comparatively late point in time was substituted by a historical corpus as the immediate basis for the editing of the dictionary (cf. Pajzs 2004).

Therefore, the obvious starting point in making any historical dictionary more efficient would seem to focus on the accessability of the archives. Fortunately, the Documentation project led by the Unit for digital documentation at the University of Oslo had both developed the technical solution and administered the digitalization of the *Norsk Ordbok* archives in the 1990s. So when the new project was initiated in 2002, one major prerequisite for the new efficient regime was already a fact. All our (to a considerable degree hand-written) slips had been photographed and indexed with headwords, so that the material was searchable and accessible through these headwords. This technical solution was chosen as the least time-consuming way of making the archive digitally accessible. The ideal solution would have been to digitalize the content of the slips as text, the way the Austrian dialect dictionary now is in the process of doing (cf. Wandl-Vogt 2005), but this solution was attempted and considered too costly.

Moreover, by 2002 our slip archive and other digitalized sources, had been co-indexed in the so called Meta-dictionary. The Meta-dictionary, developed by Christian-Emil Ore at the Unit of digital documentation (EDD) provides a precise list of all the different lemmas in our archives. Since homographs have been separated and dialectal variants have been united, we know both the number of lemmas we have to consider and plan our dictionary according to, and we know how many instantiations of each lemma there are in our archive. Although the archive in this way is more accessible since screen access is easier than the physical handling of paper slips, the actual perusal of lemma occurrences still have to be done.

Of the two next projects we started in 2002, one was an indirect answer to the problem of material analysis. We started compiling a text corpus to supplement, not to substitute our archive. The corpus was developed by Daniel Ridings at EDD. By now, our corpus consists of 25 million words, and when editing a "large" word, it enables the editors to express hypotheses that serve as starting points when checking the archive afterwards. A corpus provides the patterns and the collocations that give structure to the archive material. I venture to say that the corpus promotes a deductive way of perusing our archive, where we before hade to rely on a basically inductive material analysis. It is, however, important to note that our archive contains a large amount of rare, historically interesting dialect words and meanings that would never surface in a regular text corpus, thus validating the use of our archive as a source to a combined dialect and written standard lexicon.

We went on to specify a database that would accommodate the editorial routines and practises that by then had produced four out of twelve planned volumes of *Norsk Ordbok*. Our goal was to turn *Norsk Ordbok* into a database, with a dictionary writing system and a search system developed to go along with it. Oddrun Grønvik at Norsk Ordbok 2014 specified the structure of the "new" *Norsk Ordbok* down to the last detail, whereas Lars Jørgen Tvedt at EDD developed a database format to accommodate it. Their very close cooperation and mutual readjustment process secured a happy result. Here my main focus is the way our dictionary database is integrated with the Meta-dictionary (our indexed archive), thus making a physical link between the sources and the dictionary entries. In documentary historical dictionaries, where one objective is to demonstrate how words are actually used, this kind of integration represents a very interesting development.

When an editor of *Norsk Ordbok* starts editing a new word, he does so by generating the word from the Meta-dictionary. There is thus a physical link between the Meta-dictionary and the dictionary database. This link has the potential of being activated either from the Meta-dictionary or from the dictionary entry. So far this link is not accessible through the internet, although the Meta-dictionary is. This kind of integrated access through the internet is an immediate goal for us, but depends for now on the outcome of negotiations with our publishing house about the rights to the electronic version of *Norsk Ordbok*. We hope to have settled these issues in the early part of 2006. For now the fact remains that the archive and dictionary databases are integrated today, and that they are exploited accordingly by our editors.

Our application's latest extension is one that originally was motivated by efficiency concerns, but which explicitly contributes to an ideal of integration. This is our new sorting module. When the old editors were faced with the digitalized slip archive on screen, their response was mostly positive, but one objection was voiced immediately. How to sort the slips thus viewed on screen? The traditional mode of sorting word forms and meanings was of course to sort paper slips on one's desk. Actually this sorting procedure has been continued by many editors even after the slip archive was made available on screen, because we have had no direct way of sorting the digitalized material within the application. Dating from this summer (2005) we now have a functional sorting device. When the editor has generated a lemma from the Meta-dictionary, he imports the Meta-dictionary data into the sorter which is

a module attached to the dictionary entry. The same holds for corpus concordances. This allows the editor for the first time to sort corpus and slip archive occurrences in one operation. The sorting device is very flexible. Each editor defines his own sorting criteria, his own labels to these critera, and the number of criteria he wants to apply. Each occurrence is thus given values according to variables in a matrix, enabling the editor to sort the material by these values after the mark-up is done. The primary advantage of this procedure is of course that we have done away with the physical handling of paper slips once and for all, but I will comment on two very interesting side effects as well.

Firstly, we can now *save* the results of our sorting, and we can save *different* groupings of the material. Before, when having sorted paper slips in small heaps according to one criterium, one lost that pattern when starting to sort the material according to another criterium. Compensational and reparational procedures had to be thought of in order not to lose the results of one's analyses. Now the results of the analyses are saved in connection to the dictionary entry. Secondly, and most excitingly, the editors may now make physical links between the sorted material and the dictionary entry, thus connecting all instances of one sorting criterium to the appropriate information category in the database. An example could be that all relevant material instantiating one specific meaning of a word is accessible from the relevant node in the semantic structure of the database article, although maybe only one or two examples are actually used as quotations in the entry itself.

By these means the digitalized project Norsk Ordbok 2014 has not only reached the goal of more efficient, less time-consuming editorial routines, we have also reached a situation where the dictionary is much more integrated with its sources than before, and once here, we do not hesitate to conclude that this situation scientifically is a very interesting one.

3 The integration of a dictionary with its sources, some general considerations

The possibility of erasing the clear-cut boundary between the dictionary and its sources is of course primarily a corollary to the electronic dictionary. The traditional way of integrating sources and dictionary is simply to include quotations in the entries, a practise that defines *Norsk Ordbok* and all other documentary dictionaries. The electronic dictionary provides completely different opportunities as to the degree of how one may lay ones' sources open to the user.

De Schryver (2003:167ff.) has commented on what he calls the dream of including a corpus cum query tool as an integral part of an electronic dictionary. He refers to Varantola who already in 1994 suggested that users should be allowed to access corpus citations directly from the appropriate dictionary sense, and also to others who have ventured similar ideas. However, De Schryver regrets to conclude that with one exception, this dream had not come through by 2003. He makes particular mention of the historical dictionaries in this respect and refers to Cederholm 1996 and Ruus 2002 who have suggested that one could create a dynamically growing historical language database and link the electronic dictionary to a bibliographical database, to some of the actual source texts and to other synchronic dictionaries. These ambitious ideals make De Schryver (2003:169) conclude by saying that we "might be dealing with science fiction in the end".

I think that our experiences at Norsk Ordbok 2014 not only show that these goals are closer at hand than one could expect, they are obviously also very interesting when reconceptualising the historical documentary dictionaries. Any dictionary that wants to be "scientific" in that it wants to include quotations that attest instances of linguistic usage, will in my opinion be more interesting as a scientific enterprise if its users could check for themselves what basis the entry has been formulated upon. If the dictionary entry in the form of definitions related to each other in a semantic hierarchy, is seen as a scientific hypothesis about actual word usage, the result should according to ideals of science be reproduceable or possible to falsify. Unless the users have access to the editors' sources, his interpretation cannot be immediately refuted, except, of course on the basis of linguistic intuition.

From a more pragmatic perspective, it is self-evident to me that the historical dictionary is more true to its own rationale if it is able to lay more of its sources bare to its users. This touches upon a fundamental problem common to all historical dictionaries, namely the problem of limitation. If one should take the goal of documentation seriously, it is logical that these dictionaries are ever-growing, rendering 20, 30 or even 40 volumes. The integration of sources and dictionary within the electronic medium, solves this problem, and I think enhances the relevance of these dictionaries to the public. To the editors of *Norsk Ordbok* who adhere to strict limitations when including quotations, it is very gratifying that these limitations will be remedied for the users who have access to the same and full material as they have themselves.

Our experiences at Norsk Ordbok 2014 prompts me to suggest two ways of integrating the dictionary and its sources. The first presents a dictionary and its corpus or its archive in such a way that the users have access to the same material as the lexicographer and might go back and forth between the two. Thus the user might reproduce the lexicographer's analysis or make an alternative one. The second way is exemplified by our sorting device. Here the dictionary entries are linked to the material such as it was interpreted or processed by the lexicographer. The users are thus more explicitly told which part of the material that forms the basis for each particular sense number or formal variant. So far we see that the second solution is scientifically more gratifying, but it is more costly to prepare, and it is not obvious that it will be the preferred one by the users.

4 Conclusion

The project Norsk Ordbok 2014 has since 2002 come a long way towards its goal of digitalization. We started out by wanting to improve our efficiency, enabling us to complete the dictionary by 2014 as the Norwegian parliament had asked us to. But as one module after another has been developed for us, we now see that our integrated application improves and rejuvenates our work in ways that we initially did not explicitly think of. The electronic dictionary format, which the database solution brought along, has given us the opportunity of tearing down the clear-cut border between the dictionary and its sources. As a historical documentary dictionary we are greatly served by exploiting this opportunity, and I think that the whole genre of historical dictionaries will benefit from development along similar lines.

References

A. Dictionaries

Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet. 1966-. Oslo, Det Norske Samlaget.

B. Other literature

Cederholm, Y. (1996), 'A Historical Lexical database of Swedish. The O.S.A. Project', in Gellerstam, M. et al. (eds) *Euralex '96 Proceedings I-II, Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden.* Gothenburg, Department of Swedish, Göteborg University, pp. 65-72.

De Schryver, G-M. (2003), 'Lexicographers' dreams in the electronic-dictionary age', *International Journal of Lexicography*, 16:2, pp. 143-199.

Kilgariff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004), 'The Sketch Engine', in Williams, G., Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France July 6-10, 2004*, Lorient, Université de Bretagne-Sud, pp. 105-116.

Pajzs, J. (2004), 'Wade through letter A: the current state of the Historical Dictionary of Hungarian', in Williams, G., Vessier, S. (eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France July 6-10, 2004*, Lorient, Univer-sité de Bretagne-Sud, pp. 397-404.

Ruus, H. (2002), 'A Corpus-based Electronic Dictionary for (Re)search', in Braasch, A., Povlsen, C. (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13-17, 2002.* Copenhagen, Center for Sprogteknologi, Københavns Universitet, pp. 175-185.

Varantola, K. (1994), 'The Dictionary User as Decision Maker', in Martin, W. et al. (eds.), *Euralex 1994 Proceedings, Papers submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam, The Netherlands.* Amsterdam, Vrije Universiteit, pp. 601-611.

Wandl-Vogt, E. (2005), 'From Paper Slips to the Electronic Archive. Cross-linking Potential in 90 Years of Lexicographic Work at the Wörterbuch der bairischen Mundarten in Österreich (WBÖ)', in Kiefer, F. et al. (eds.), *Papers in Computational Lexicography. Complex 2005.* Budapest, Lingusitic Institute, Hungarian Academy of Sciences, pp. 243-254.