Making Dictionaries for Paper or Screen: Implications for Conceptual Design

Lars Trap-Jensen
Society for Danish Language and Literature
Christians Brygge 1
1219 Copenhagen K
DENMARK

Abstract
With the development of digital dictionaries we can foresee the dictionary as a genre that will gradually change. In this paper, one possible direction of such a change is treated, based on considerations from a Danish project that seeks to combine digitized versions of two existing paper dictionaries with a corpus site into an online reference tool with new facilities. Examples of morphological and syntactic information are shown to illustrate how the digital possibilities necessitate a revision of the existing DTDs/XML schemas traditionally used in paper dictionaries.

1 Introduction

Although digital dictionaries are now quite common alongside traditional paper dictionaries, we have not yet seen many examples of completed dictionaries conceptually designed for publication on screen only. The situation may still be characterized as one of transition in which a dictionary is usually published in two versions, a paper version and a screen version, with only few substantial differences between them. A notable exception, perhaps, is the highly competitive market of learners' dictionaries where the contours of a new development are emerging in which the two products begin to separate.

The project that is the background for the present article is typical in that respect. It is a project that involves electronic versions of two existing paper dictionaries. These are to be made publicly available online with some new facilities; among other things one goal is to provide a closer integration between a dictionary component and a corpus component in order to enable the users to make their own research on the spot and to provide a given reference with additional example material on request. In this connection focus will be on two types of information: morphological and syntactic information as presented in the dictionary entries.

2 Project background

The two dictionaries in the project (www.ordnet.dk) are both monolingual dictionaries of Danish, compiled by the Society for Danish Language and Literature: The *Ordbog over det danske Sprog,* also known as ODS (Dictionary of the Danish Language, cf. *www.ordnet.*

dk/ods), covering the language from 1700 to 1950 appeared in 28 volumes from 1918 to 1956; later, five supplementary volumes have appeared which are to be integrated with the original manuscript as part of the current project. However, being a historical dictionary the ODS will not be revised or in other ways changed as regards content and is therefore of less relevance for the present discussion. Much more relevant is *Den Danske Ordbog* (henceforth DDO, The Danish Dictionary) covering the language from 1950 to the present, which appeared in six volumes from 2002 to 2005. As a modern dictionary, the manuscript was prepared electronically in accordance with explicit rules, and the document structure is fairly orderly and consistent. It is furthermore the first ever dictionary of Danish to be corpus-based, and in the electronic version it will gradually develop away from the paper dictionary as new entries will be added and the original entries or entry elements will be presented in new ways.

The corpus site (www.korpus2000.dk) contains part of the corpus texts on which the DDO was based (texts with special restrictions and spoken language texts have been excluded), covering the period 1988-92, as well as later collections of texts from 1998 to 2002. At present, the corpus contains approx. 56 million words, a number that will grow continuously as more texts are added as part of the current project.

3 DTDs/XML schemas

The manuscript of the DDO was written using an SGML-based dictionary writing system; this was later converted into an XML-based system as part of the current project. As such, however, it is immaterial for the present considerations whether SGML or XML is used, or a DTD or schema. More important is the fact that the end product inevitably affects the way the DTDs/schemas are designed. Let us first consider morphological information.

3.1 Morphological information

Space economy is traditionally a very important parameter for lexicographers in deciding how to present information in a paper dictionary. This is the main reason why it was decided to adopt a condensed style of presenting morphological information in the printed DDO. Also, typographical appearance is obviously an important concern in the preparation of a manuscript for a printed dictionary. As an example, consider the head of the entry **sav** ('saw') shown in figure 1. In the paper version, the morphological information looks as follows:

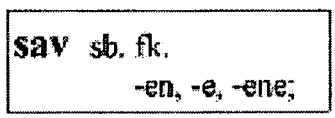

Figure 1. sav ('saw') in print

In the underlying schema the same morphological information is presented as in figure 2

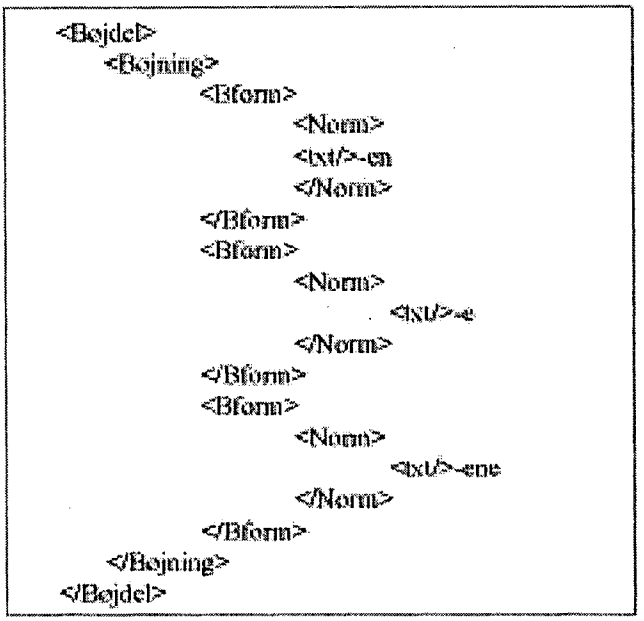

Figure 2. XML structure showing inflection of the lemma **sav** ('saw')

The condensed form in Figure 1 requires a lot of implicit knowledge. We are first told that the lemma is a noun (sb.) in the common gender (fk., as opposed to neuter), and then come the inflectional forms which should read: definite form singular is **saven** ('the saw'), indefinite plural is **save** ('saws'), and definite plural is **savene** ('the saws'). Indefinite singular is the lemma form. This way of presentation was chosen because the dictionary is aimed at human users who have learned a conventional way of ordering inflectional forms. From Figure 2 we can see that the XML elements used for the three forms are identical, a mere text element within the node for (orthographically authorized) inflectional forms. The only clue to the right interpretation is relative position within the linear ordering, and this works well for human users. As stated above, it is perfectly legitimate when preparing a paper dictionary. In the paper dictionary the crucial point is that the authorized form can be distinguished from the unauthorized one – hence the element <Norm> in the XML structure which ensures that the content of this element can be presented differently from the element <Unorm> which is used for commonly found unauthorized inflectional endings. In our case, however, it is obvious to use the morphological information for various additional purposes, e.g. in the look-up part of the interface to ensure that the user gets a match no matter which form of a word is keyed in, or for corpus purposes, e.g. in enabling lemmatization and part of speech tagging of corpus texts.[1] For that purpose, a morphological full form lexicon is needed. With that it is

[1] Apart from our own rather restricted use of it, a morphological lexicon can of course be used for a whole range of NLP purposes in its own right.

ensured that the query will match all and only tokens of the lemma in question. In order to do so, we need to change the schema in such a way that all attested forms of the lemma are registered and uniquely labelled in different elements in the base, resulting in a full form lexicon for all lemmas in the dictionary.

Therefore, one important task is to convert the schema into a format that is suitable for human as well as for language technology needs. It is important that the structure makes allowance for language technology principles of algorithmic processing, i.e. the structure should be explicit, exhaustive and free of unnecessary redundancy. In order to achieve this, the contents of all the existing <txt> elements are to be converted into a full form lexicon in a process that involves the following elements:

1. a large part of the regular inflectional forms can easily be identified automatically as there are no ambiguous endings within the paradigm for each individual part of speech.
2. some of the regular forms exhibit consonant gemination. In itself, this is hardly a problem as it happens according to regular principles. It does, however, entail that they are converted in a separate step.
3. irregular forms (altogether just over 10,000 instances out of a total of 153,000 inflectional forms) may be divided into minority patterns which can be solved in separate steps, and words that involve stem transformation, primarily vowel mutation (Umlaut), which have to be dealt with manually.
4. officially authorized as well as and non-authorized variants included in the dictionary are in this connection unproblematic as they have been tagged differently from the beginning. Their full form and morphological status can thus be inferred directly from the immediately preceding <txt> element.

Once the full form lexicon has been established, it is probably desirable to derive general paradigms by grouping the material into frequent patterns for each part of speech. Thereby the lexicographers' work will be made more efficient as they need only to refer to the paradigm by an ID number when editing new articles, and, more importantly, changes in the paradigm or its representation can be carried out centrally and not in each individual article.

Needless to say, inflectional information can be extracted and presented in a fully flexible way according to the actual publishing need, whether on screen or paper. A presentation as in Figure 1 would still be perfectly possible for a paper version. For the online version, one possibility would be to have a brief, yet fully expanded, presentation given as the default reading, with the complete paradigm including element names as a clickable option together with information about authorized and unauthorized variants etc.

3.2 Syntactic information

The DDO brings information on valency for all verbs in the dictionary, and offers some valency information for other parts of speech as well as other relevant constructional information, e.g. auxiliary verb (see e.g. Lorentzen/Trap-Jensen 2005). Again, the structure is largely determined by the desired typological appearance, and although the presentation has

the form of semi-formal frames, the information is too implicit and incomplete to be used directly as a general resource for language processing purposes. Within the overall conceptual design as a printed dictionary for humans, the DDO notation is, however, fairly well-structured and consistent, and a recent article, Asmussen and Ørsnes 2005, describes how the valency information of the DDO can be transformed into a more generalized notation which allows conversion to a formal representation suitable for NLP purposes.

Comparable with the solution for morphological information, the XML structure behind the printed dictionary uses but a single element in which the whole syntactic frame is placed; only auxiliary verb information is specified in a separate element. To illustrate the basic structure, consider the examples given in Figure 3 (reproduced from Asmussen/Ørsnes 2005: 2).

(1)	NGN/NGT specificerer NGT	SB/STH specifies STH
(2)	NGN spadserer (+STED/+RETNING)	SB walks (+PLACE/+DIRECTION)
(3)	NGN teoretiserer (over NGT)	SB theorizes (about STH)
(4a)	NGN barberer sig/NGN/NGT	SB shaves (oneself)/SB/STH
(4b)	NGN/NGT barberer HÅR af/væk/bort	SB shaves hair off/away/away
(4c)	NGN barberer NGT ned/væk/bort	SB shaves STH down/away/away
(5a)	NGN diskuterer (NGT) (med NGN)	SB discusses (STH) (with SB)
(5b)	NGL diskuterer (NGT) med hinanden	SB (plur.) discuss (STH) with each_other
(5c)	NGN diskuterer om+SÆTN/hv+SÆTN	SB discusses if+CLAUSE/wh+CLAUSE
(6)	NGN planlægger (NGT/at..)	SB plans (STH/that+CLAUSE /to+INF)

Figure 3. Notation of valency in DDO

For all the examples in figure 3, the same XML structure is used, as shown in figure 4.

```
<Valens>
        <Aktant>NGN/NGT specificerer NGT</Aktant>
</Valens>
```

Figure 4. XML structure showing the valency information of example (1)

Since all the information is coded as one string in a single element, valency information rests on several implicit assumptions. Firstly, syntactic function is not stated explicitly, but the human user is able to deduce this from his or her knowledge of basic constituent ordering of Danish sentences. Therefore, the human user will know that NGN/NGT is the sentence subject in figure 4 and NGT the sentence object. Secondly, syntactic, semantic and morphological information is conflated in the notation. For example, a noun phrase can be rendered either by NGN (NP sg., +human), by NGT (NP sg., -human), or by NGL (NP pl., +human). And thirdly, it can only be deduced that the order of alternating elements is based on corpus frequency, e.g. as seen in example 4a in Figure 3 (the elements separated by dashes: sig/NGN/NGT). The

solution proposed by Asmussen/Ørsnes 2005 is one in which the notation gives explicit information about (1) syntactic function, (2) syntactic category (NP, VP, PP etc.), (3) morpho-syntactic restrictions, and (4) selection restrictions. In addition, information on constituent order and the order of alternating constituents needs to be specified in the notation. In a test operation, the conversion of the existing data was carried out semi-automatically, again involving several steps:

1. First, the sentence verb was identified.

2. Secondly, on the basis of word order, sentence subject and object(s) were established automatically in the most common cases. The same could be done for the most common types of other sentence constituents such as adverbial and prepositional phrases. In case of alternating material, each alternation was rewritten as a separate pattern.

3. By means of a small Perl program the patterns were converted automatically in a series of steps, first isolating the most frequent and simple cases, then proceeding by modifying the program with the additional rules necessary to deal with the second most frequent patterns, and so on.

Our experiments showed that conversion of the material in this way can be done automatically for 94-95 % of all the patterns, leaving approximately 5-600 patterns to be solved manually. The bulk of the remaining cases are patterns where selection restrictions have been encoded in the valency notation, resulting in e.g. "hund gør" ('dog barks') rather than "NGT gør" ('STH barks').

4 Perspectives and conclusion

Changing the DTD/schema along the lines suggested or implementing it from the beginning of a digitally conceptualized dictionary will no doubt lead to mutual benefits for both dicitionary and corpus. Explicit and exhaustive morphological information is a prerequisite for correct mapping between corpus instances and the corresponding dictionary entry. In the dictionary component, we can use corpus information to indicate the relative frequency of a particular form. Non-attested forms will be hidden in the default reading, but can be offered as a clickable option for the user who wants to know the potential forms of a word. Similarly, information on unauthorized or rare spelling variants is given in this section. From the corpus point of view, the perspective is an improvement of the existing full form lexicon. For example, information on common spelling errors and unauthorized morphological variants is not readily available and is therefore excluded from the results in the existing corpus site. An improved full form lexicon will ensure a more reliable tagging of the corpus texts.

Correspondingly, the perspective of a more accurate – or better, NLP-friendly – valency notation in the dictionary is that it improves the parsing of the corpus texts. And in the dictionary component, it is utilized to provide more precise frequency information where relevant, for example for competing auxiliary verbs or for constructions with alternating prepositions (e.g. "information on/about"?). Another attraction is the possibility of making corpus queries for a specific syntactic pattern as a direct, clickable option under the entry in question. In a wider perspective, the resulting outline of valency notation may prove useful in its own right, for instance for linguists interested in exploring the syntactic characteristics of verbs (cf. the work documented by Beth Levin in several articles).

References

A. Dictionaries and corpora

Den Danske Ordbog 1-6 ("The Danish Dictionary"), eds. Ebba Hjorth, Kjeld Kristensen et al. Gyldendal Publishers 2003-2005, Copenhagen.

Korpus 2000, www.korpus2000.dk.

Ordbog over det danske Sprog ("Dictionary of the Danish Language"), www.ordnet.dk/ods.

B. Other Literature

Asmussen. J., Ørsnes, B. (2005), 'Adapting valency frames from The Danish Dictionary to an LFG lexicon', in Kiefer & Pajzs, *Proceedings of the 8th Conference on Computational Lexicography, COMPLEX 2005. Papers in Computational Lexicography*, Budapest.

Lorentzen, H., Trap-Jensen, L. (2005), 'Grammatiske oplysninger i Den Danske Ordbog' (Grammatical Information in The Danish Dictionary), in Ruth, V., Fjeld og Dagfinn Worren (eds.) *Nordiske Studier i Leksikografi,NFL-skrift nr.8*, Oslo, pp. 252-267.