Integrating corpora and dictionaries: problems and perspectives, with particular respect to the treatment of sign language

Jette H. Kristoffersen, Thomas Troelsgård
Centre for Sign Language, University College Capital
Ejbyvej 35, DK-2740 Skovlunde, Denmark
E-mail: jehk@ucc.dk, ttro@ucc.dk

Abstract
In this paper, we will discuss different possibilities for integration of corpus data with dictionary data, mainly seen from a lexicographic point of view and in a sign language context. For about 25 years a text corpus has been considered a useful, if not necessary tool for editing dictionaries of written and spoken languages. A corpus can provide the lexicographer with:

- frequency lists (e.g. used in connection with lemma selection or ordering of homonyms)
- examples of language use (e.g. used as evidence of particular word senses, or for example sentences (directly or adapted))
- frequent co-occurrences (e.g. used for describing multi-word expressions or valency patterns)

We know of no sign language dictionary that is truly corpus-based, or even edited with extensive use of the tools provided by a corpus, probably due to the fact that larger, fully annotated sign language corpora is a relatively new phenomenon. An example will be the new German Sign Language-German dictionary which is part of the DGS Corpus Project and will be “the first comprehensive corpus-based dictionary of DGS” (DGS-Corpus, no date [online]).

For corpus projects of languages with an established written form, the lemmatisation of tokens is typically based on an existing dictionary. For sign language corpora this is equally appropriate, as argued in Johnston (2008), but the execution is impeded by the inevitable and – at least in the nearest future – manpower-consuming task of tokenising the corpus texts sign by sign. Furthermore, this approach presupposes the existence of a dictionary or lemma list composed using consistent lemmatisation principles and a consistent identification of the lemmas, e.g. through unique gloses or numbers. For a number of sign languages no such dictionary exists, and building a corpus, would imply the simultaneous building of a dictionary, which would make the process even more time-consuming.

As language resources dictionaries and corpora are both valuable tools for many types of users, and combining the two in one interface, or linking between dictionary and corpus interfaces could, if it is done in a clear and preferably intuitive understandable way, afford a synergetic enhancement of the resources.

1. Introduction
In this paper, we will discuss different possibilities for integration of corpus data with dictionary data, mainly seen from a lexicographic point of view and in a sign language context.

Since at least the 1990’s, a text corpus has been considered a useful, if not necessary tool for editing dictionaries of written/spoken languages. A corpus can provide the lexicographer with:

- frequency lists (e.g. used in connection with lemma selection or ordering of homonyms)
- examples of language use (e.g. used as evidence of particular word senses, or for example sentences (directly or adapted))
- frequent co-occurrences (e.g. used for describing multi-word expressions or valency patterns)

We know of no sign language dictionary that is truly corpus-based, or even edited with extensive use of the tools provided by a corpus, probably due to the fact that larger, fully annotated sign language corpora is a relatively new phenomenon. An example will be the new German Sign Language-German dictionary which is part of the DGS Corpus Project and will be “the first comprehensive corpus-based dictionary of DGS” (DGS-Corpus, no date [online]).

For corpus projects of languages with an established written form, the lemmatisation of tokens is typically based on an existing dictionary. For sign language corpora this is equally appropriate, as argued in Johnston (2008), but the execution is impeded by the inevitable and – at least in the nearest future – manpower-consuming task of tokenising the corpus texts sign by sign. Furthermore, this approach presupposes the existence of a dictionary or lemma list composed using consistent lemmatisation principles and a consistent identification of the lemmas, e.g. through unique gloses or numbers. For a number of sign languages no such dictionary exists, and building a corpus, would imply the simultaneous building of a dictionary, which would make the process even more time-consuming.

As language resources dictionaries and corpora are both valuable tools for many types of users, and combining the two in one interface, or linking between dictionary and corpus interfaces could, if it is done in a clear and preferably intuitive understandable way, afford a synergetic enhancement of the resources.

2. Examples of interfaces that integrate corpus and dictionary content
In this section we will take a closer look at some existing interfaces that integrate corpus and dictionary content. The German DWDS (Digitales Wörterbuch der deutschen Sprache, ‘Digital Dictionary of the German Language’) (DWDS, no date [online]) is an example of a combined corpus-dictionary product where, as Asmussen puts it: “Corpus and dictionary are not formally interlinked, they appear side by side, accessible through a joint interface”, (Asmussen, 2012). A standard word lookup in DWDS presents the user with six sub-windows, see Figure 1. The standard lookup shows (letters refer to the labels Figure 1):

- the result of a lookup in DWDS dictionary (A), the result of a lookup in OpenThesaurus1 (B), a tag cloud (“Wortprofil”) based on the DWDS corpus (C), the result of a lookup in an etymological dictionary (D), and, finally, two concordances (E and F), one drawn from the basic DWDS corpus, and one drawn from a newspaper corpus (Die Zeit). The standard view of a search result can be changed by adding or removing included resources, and

---

1 OpenThesaurus is an open source thesaurus project that was initiated in connection with the development of the OpenOffice software. Information about the project can be found at http://www.openthesaurus.de
Figure 1: Partial screen dump of the standard view of a lookup of Sprache (‘language’) in DWDS.

Figure 2: A dictionary lookup at Ordnet.dk, with links to corpus searches.
by expanding or collapsing the views of the sub-windows individually.

Another example is the Ordnet.dk (no date [online]) website, which deals with Danish. This site also has no formal interlinking, but if differs considerably from the DWDS site, as its approach aims an interconnection of several resources rather than a simultaneous access. Thus, there is no universal search facility, but a word search in one of the two included dictionaries (a contemporary and a historic) provides the user with the dictionary content as well as with links to relevant corpus searches (on the key word and on selected collocations), see Figure 2. Similarly, a corpus search result is accompanied by lookup links to the two dictionaries, see Figure 3. In addition to this, all three resources include a link to a list of the most frequent co-occurrences of the word. See Trap-Jensen (2010) for more information on the Ordnet.dk website.

Where the two examples mentioned above are not formally interlinked, the last example, the Danish Sign Language Dictionary (no date [online]) is. The weakness of this dictionary, on the other hand, is that its corpus is what Asmussen (2012) refers to as a “quasi-corpus”, in this particular case, a corpus build entirely of adapted sentences, namely the usage examples of the dictionary. Furthermore, half of these sentences are derived from video recordings of natural signing, while the remainder have been constructed by native signers. The integration of the two resources is quite basic and somewhat similar to the one used in Ordnet.dk; in the dictionary, there are links from each sign entry to a concordance view of all the occurrences of the sign in the collection of example sentences, see Figure 4. In the other direction, the individual signs in the sentences of a concordance are linked to the corresponding sign entries in the dictionary, see Figure 5. This feature is added in order to present additional examples of the use of a sign to the user, and although the corpus is not a “real” corpus, the dictionary site still serves as an example of how corpus (or corpus-like) data can be integrated into a sign language dictionary. A discussion of what type of sentence was considered the most suitable for uncommented use as example sentences in the Danish Sign Language Dictionary can be found in Kristoffersen & Troelsgård (2010). A more detailed description of the dictionary can be found in Kristoffersen & Troelsgård (2012).
Figure 4: The Danish Sign Language Dictionary. Entry for the sign HUE (‘cap’), with a link to a concordance view of all example sentences containing this sign.

Figure 5: The Danish Sign Language Dictionary. Concordance view of example sentences containing the sign HUE (‘cap’). Glosses in the concordance lines are linked to the appropriate sign entries (if they exist).
3. Considerations regarding the integration of corpora and dictionaries

Language use is described differently in a dictionary and in a corpus; whereas the dictionary data are the result of an editing process and often adapted to a specific purpose, the corpus data, be it text examples or co-occurrence statistics etc., are “raw” and have to be interpreted by the user. The difference between the two resources is somewhat comparable to that between an encyclopaedia and an internet search; the former typically being more well-arranged and reliable, while the latter often provides more information, and more updated information, the downside being that it is presented as lots of co-ordinate results with no quality guarantee. How “dangerous” it is to present corpus data to the user depends partly on the nature of the corpus texts, partly on how trained the user is in the use of the corpus. Corpora of languages with a written form typically have written texts as their main source; digital text, e.g. from newspapers or from the internet is easy to obtain, and you will relatively unproblematically be able to build a corpus – apart from legal issues and an expected margin of error in connection with the tokenisation of the corpus texts. Building corpora of spoken or signed languages, on the other hand, requires a manual or, at best, semiautomatic tokenisation process in order to become searchable. Mainly for this reason such corpora are typically smaller than corpora of written text. Large corpora of a written language are often composed of different types of text, balanced in order to obtain a broad and adequate picture of the language use. For corpora of signed and spoken language, such a balancing will probably always be a major challenge; as you are dealing with non-written language, the only “authentic” text types available would be rather special ones like recordings of radio or television broadcasts, or recordings of speeches and conversations, which are rarely performed spontaneously. Thus, existing sign language corpora mainly consist of elicited data or recordings made for linguistic purposes in a more or less unnatural context. Many corpora contain non-edited language, allowing for ungrammatical language use, misspellings (for written language), and, especially for spoken/sign texts, elliptic utterances. As a result of the above mentioned impediments for building a corpus, a corpus user could find him or herself dealing with corpus that is of limited size, with more or less unnatural text types, and containing ungrammatical sentences. For e.g. a lexicographer, this would be a minor problem, as he or she would look at the source critically, but for an inexperienced corpus user, and even more so for a user of an integrated corpus-dictionary interface, who isn’t necessarily aware of the complex character of the resource, it could be quite problematic to extract the needed information, cf. the discussion in Asmussen (2012). For this reason, an integrated corpus-dictionary interface should always, at least ideally, provide the user with the necessary prerequisites for using the corpus in a meaningful way, e.g. by informing about the corpus sources, and by clearly indicating if the user is presented with edited or non-edited text. A corpus-dictionary pair of a specific language can be more or less closely related, or coherent, so to say. Thus, the ideal prerequisite for an extensive integration of a corpus and a dictionary is a situation where the lexicon and definitions of the dictionary are based on the corpus which, in its turn, is linked token by token to the dictionary. On the other hand, as exemplified in Asmussen (2012), if the two resources are not based on the same source texts, an integration can lead to situations where the user is presented with a confusing result, e.g. if a specific word sense is predominant in the corpus, but absent in the dictionary (or the other way round).

4. Future perspectives of integrated sign language resources

The integration of sign language corpora and dictionaries is an obvious field for development in the future. The crucial formal interlinking between corpus and dictionary, which for written language corpora is often insufficient or missing, is typically an innate feature of a sign language corpus project, as a tokenisation is needed in order to make the corpus searchable, and as the tokenisation, in its turn, requires a dictionary or lexicon. The synergy that arises from joining a corpus and a dictionary could even be enhanced by including e.g. grammatical information, or links to external resources. There is however a risk that the user is overwhelmed by the amount of diverse data and possibilities presented in the interface. Another future challenge could be the development of means of accessing sign language corpora that are more appropriate than traditional text-based concordance views. Several sign languages are now documented, or in the process of being documented, through corpora, e.g. Australian, British, Dutch, German, New Zealand and Swedish Sign Language, and, hopefully, we can look forward to seeing some innovative projects integrating corpus and dictionary data in the future.

5. References


