Improving Lemmatisation Consistency without a Phonological Description The Danish Sign Language Corpus and Dictionary Project

Thomas Troelsgård, Jette Hedegaard Kristoffersen
Centre for Sign Language, University College Copenhagen
{ttro, jehk}@ucc.dk

Abstract

The Danish Sign Language Corpus and Dictionary project at Centre for Sign Language, UCC has a dual aim: to build of Danish Sign Language Corpus, and to use this corpus to expand and improve The Danish Sign Language Dictionary. Our goal is a one-to-one correspondence between sign lemmas in corpus and dictionary, but due to limited resources, we cannot include an accurate phonological description of each sign form. In order to secure a consistent lemmatisation in the corpus as well as across the two resources, we thus rely exclusively on sign videos and Danish equivalents. In this paper, we will describe how we use the lemmas of the Danish Sign Language Dictionary, and additional signs found in connection with the dictionary work, as the initial lexical database of the corpus tool. For new signs found in corpus, the actual corpus tokens will serve as preliminary video representations. To facilitate the sign search when lemmatising corpus tokens, we assign several Danish equivalents to each sign, including all equivalents in the dictionary data. Furthermore, we include synonyms found through linking these equivalents to the Danish wordnet (DanNet), although equivalents added in this way cannot be regarded as valid senses of the sign.

Keywords: corpus linguistics, annotation, sign language, language documentation, Danish Sign Language (DTS)

1. Introduction

The Danish Sign Language Corpus and Dictionary project is carried out at the Centre for Sign Language at UCC - by the same project group that developed the Danish Sign Language Dictionary (Ordbog over Dansk Tegnsprog; cf. Kristoffersen and Troelsgård, 2012). In 2015, we began working on a corpus of Danish Sign Language (DTS), the first of its kind. The current project phase has a dual goal: to build a corpus of DTS, and to expand and improve The DTS Dictionary based on this new corpus. For building our corpus we use the iLex system (cf. Hanke and Storz, 2008), a database tool that is developed at the University of Hamburg.

In order to secure consistency across corpus and dictionary, we aim at a one-to-one correspondence between the dictionary lemmas and the corpus lexicon – the set of types used for lemmatising corpus tokens. Unique identifiers of sign types are essential to machine readable text that can serve as the source for linguistic analysis of the sign languages (cf. Johnston, 2010). The lack of a written standard for Sign Languages commonly used by native signers complicates the identification of the lemmas in the annotation process (cf. Zwitserlood et al., 2013). To achieve an unambiguous lemmatisation, some corpus projects, e.g. the German Sign Language Corpus (DGS-Corpus), include a detailed formal description of the sign form, e.g. in HamNoSys (The Hamburg Sign Language Notation System, cf. Hanke, 2004). Other projects represent a sign solely through a gloss – typically a word from the surrounding spoken language, chosen as a mnemonic because it captures (one of) the core meaning(s) of the sign. For the DTS Corpus project we chose to use only glosses because of limited resources. In this paper, we will describe how we try to achieve a high degree of consistency in the corpus annotation and across corpus and dictionary, without having a searchable description of the sign form.

2. Building the vocabulary

In the iLex system, the lemmatisation task is performed by linking every token to a matching type in the lexical database. Obviously, this linking is completed faster, easier and more reliable if the initial sign vocabulary is large and well described, ideally having both a video, a searchable formal form description (e.g. HamNoSys or Stokoe), and one or more spoken language equivalents. Because of limited funding, we decided to leave out the formal description, and go with only videos and Danish equivalents (and/or a prose description of function or use).

2.1 Initial vocabulary

For building our sign vocabulary, we first included the approximately 2.200 lemmas of the DTS Dictionary. As the signs were already analysed regarding form and meaning,, we decided to re-use the definitions of homophony and phonological variation that we use for the dictionary (Kristoffersen and Troelsgård, 2012), and hence (ideally) end up with a one-to-one relation between sign units in the dictionary and in the corpus project.

As a tool for lemma selection for the dictionary project, we built a database containing the signs from a number of older dictionaries and sign lists. We then began analysing video recordings of DTS provided by our group of consultants, adding new signs to the database as we encountered them in the videos. The database was then used as source for the selection of lemmas for the DTS Dictionary. During the following lexicographic work on the dictionary, new signs were continuously added to the database. While building the sign type vocabulary for the corpus, we included all signs from the database that were not already dictionary lemmas. In connection with adding signs to the database, we also added the known phonological variants of each sign according to the variant definition of the dictionary: signs with the same semantic content and variation in only one of the major phonological parameters: handshape, orientation, movement, place of articulation, are regarded as phonological variants of one sign (cf. Troelsgård and

Kristoffersen, 2008). Finally, as a preparation for the corpus project, we made studio recordings of all signs and their phonological variants in the database that were not already dictionary lemmas.

Consequently the initial sign vocabulary in the corpus system consisted of about 7.000 signs (and about 1.000 additional sign variants), all accompanied by a video recording (either from the dictionary or added in connection with the preparation of the corpus project).

2.2 Adding new signs

As soon as we started annotating corpus videos, obviously the need occurred of being able to add new signs to the vocabulary as we encounter them. These signs are lemmatised using temporary "dummy signs", which are regularly checked, and – if they are found actually to be missing in the vocabulary – added to the database, with the actual corpus tokens serving as video evidence. All signs found in the corpus are regarded as future lemma candidates for the dictionary. If a sign is later selected as a dictionary lemma, we will compile a new entry based on an analysis of the corpus tokens, and we will make studio recordings of the sign and its variants.

3. Adding equivalents

As we decided not to include a formal phonological description, it is essential to provide one or more Danish equivalents to each sign. As the 2.200 dictionary signs already were semantically analysed, and described as having one or more sense (each with one or more Danish equivalents, and/or a prose description of function or use), we decided to exploit the possibility in the iLex system of structuring the sign type vocabulary as a hierarchy, and thus we clustered the equivalents according to the word-senses defined in the dictionary. As a result, we work with a three level hierarchy, which we will illustrate through the sign FRUIT, a sign described as having two word-senses, and two phonological variants. The variants differ in handshapes – the movement is in both cases a twist of the wrist, see Figure 1.

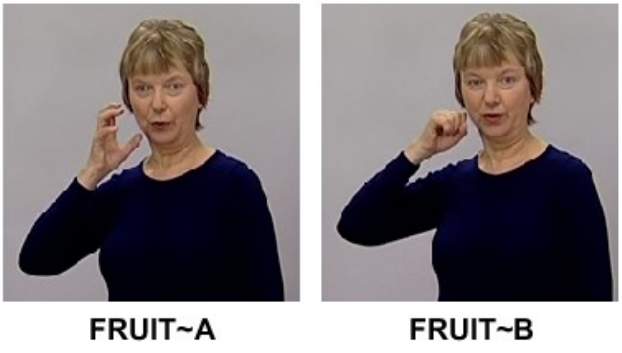

Figure 1: The two phonological variants of the DTS Sign FRUIT.

For the type hierarchy this gives one type at the sign level, two types at the variant level (form), and four types at the meaning level (combination of form and sense), as shown in Figure 2. A more detailed description of the way we use the iLex type hierarchy can be found in Langer et al. (2016).

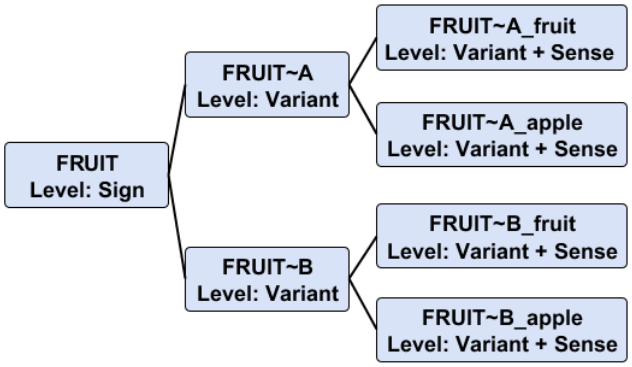

Figure 2: The three-level sign type hierarchy used in the iLex system for the DTS Corpus project.

At the meaning level, we add the first equivalent of the corresponding dictionary word-sense to the gloss as a disambiguator. Furthermore, we use iLex' module for linking types with concepts to assign all Danish equivalents from the DTS Dictionary to the type, thereby making it possible to find the sign through these equivalents. As an example of this linking, we use the sign WOMAN. The DTS Dictionary entry of WOMAN is shown in Figure 3.

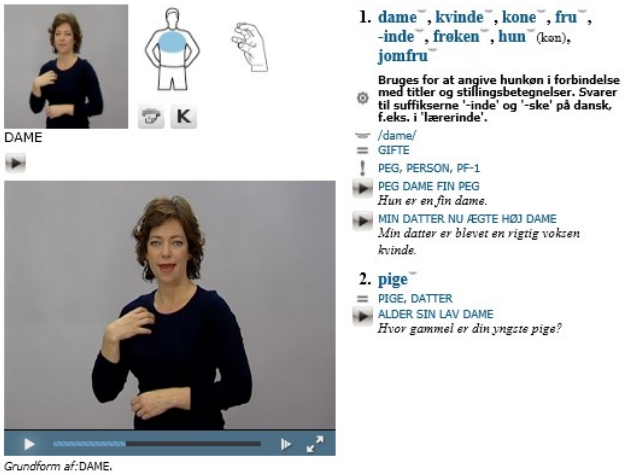

Figure 1: DTS Dictionary entry of WOMAN.

Table 1 shows the meaning level types of the two senses of WOMAN, and the linked (and searchable) equivalents taken from the DTS Dictionary.

Type at meaning level	Linked equivalents
Sense 1: WOMAN_woman	dame (woman) kvinde (wife) kone (wife) fru (madam) -inde (-ess) frøken (miss) hun (female) jomfru (virgin)
Sense 2: WOMAN_girl	pige (girl)

Table 1: The two meaning level types of WOMAN and their linked equivalents from the DTS Dictionary

4. Linking to DanNet

We wanted to add even more relevant Danish equivalents to each sign sense, thereby increasing the possibility of a match when searching signs through words. For this purpose, we chose to use the Danish wordnet, DanNet (DanNet; cf. Pedersen et al., 2009; Trap-Jensen, 2010). A wordnet is a semantic network that clusters closely related word-senses (synonyms and near-synonyms) into so-called synsets, and links these together according to semantic relations such as hyponymy, hypernymy, metonymy, entailment etc. We matched our dictionary equivalents against the DanNet words, and performed a semiautomatic linking between dictionary senses and relevant DanNet synsets. Using these links, we then were able to add equivalents to each word-sense, by including all synonyms of its linked DanNet synsets. Thus, if we consider the sign WOMAN, it is described in the DTS Dictionary as having two senses: 'woman, wife' and 'girl'. The first sense has a number of equivalents in the dictionary data, including *dame* ('lady'), *kvinde* ('woman'), *kone* ('wife'), *fru* ('madam'). If we match e.g. *kone* ('wife'), to DanNet, we get five additional equivalents from the synset of *kone*: *ægtehustru*, *ægteviv*, *frue*, *hustru* and *viv* (all meaning 'wife'). When choosing equivalents for the DTS dictionary, we balanced word frequency against the total number of equivalents, and because of the large number of relevant equivalents for the sense 'woman, wife', none of the five words found through DanNet were chosen as equivalents for the entry WOMAN. Nevertheless, as shown in Table 2, the two most frequent words added through the DanNet matching: *hustru* and *frue* are fairly frequent, and are likely to be used as search words during the lemmatisation of corpus tokens.

Danish equivalents	Frequency	DTS Dictionary
kvinde	3090	present
kone	2573	present
dame	942	present
hustru	579	absent
frue	184	absent
viv	12	absent
ægteviv	2	absent
ægtehustru	0	absent

Table 2: Danish words meaning 'wife', with word frequency count from the Korpus 90 Project[1]

Obviously, the equivalents added in this way cannot be regarded as valid senses of the sign – they are included solely for the purpose of increasing the opportunity of finding a sign though a word-based search. The possible sign-senses – and their appropriate equivalents – can only be deduced through analysis of the actual corpus tokens of each sign.

5. Word-based type search

In the absence of a formal sign description, word-based search is the primary means of identifying the correct sign type while annotating the texts of the DTS Corpus. Through a text search, hopefully the matching sign – checked by watching the connected video evidence – is found (preferably in a matching word-sense), and used for the lemmatisation.

A disadvantage of this approach is that it is impossible to foresee all possible search strings; hence, sometimes searches for signs that are actually in the system do not give any result. In these cases, we lemmatise using special dummy types. Later on, we examine these dummies, in order to decide whether they are instances of existing signs, or of new signs, not yet entered as types in the system.

Sometimes a search results in finding the appropriate sign, but not finding an adequate type at the meaning level. In these cases, we go up one level in the type hierarchy, lemmatising to a type at the variant level, e.g. using the type FRUIT~B, as shown in Figure 2, and indicating that the sign form is right, but the actual sense is neither 'fruit' nor 'apple'.

6. Concluding remarks

For the DTS Corpus project, we do not have searchable formal sign descriptions at hand. Instead, we have chosen an approach where we add many spoken language equivalents to each sign, in order to increase the probability of finding the right sign when lemmatising corpus tokens. Furthermore, we work with a lexical sign base, where every record is represented by a video recording. This secures a correct choice of sign type. Especially when dealing with

[1] Korpus 90 was part of the work on Danish text corpora (cf. KorpusDK) carried out at Society for Danish Language and Literature (DSL, cf. www.dsl.dk). Recent word frequently lists from DSL can be downloaded at korpus.dsl.dk

phonological variants and sign synonyms, the video evidence secures a correct choice.

We believe that this approach is a feasible, second-best solution for sign language corpus projects without resources for performing a detailed phonological description of the sign vocabulary and tokens of their corpus. We also suppose that including the relation links of wordnets might increase the success rate of word searches, as might the inclusion of other spoken language resources, e.g. corpus tools for finding related words.

7. Acknowledgements

The database tool iLex has kindly been placed at our disposal by the developers at Hamburg University. We would also like to thank the DGS Corpus group at Hamburg University for ongoing support of our use of iLex and our corpus building.

The Danish Sign Language Corpus and Dictionary project is supported by funding from the "Diversity and Social Innovation" research fund of UCC (Professionshøjskolen University College Capital, Denmark), and from The Jascha Foundation (Denmark).

8. Bibliographical References

Hanke, T. (2004). HamNoSys - representing sign language data in language resources and language processing contexts. In O. Streiter, C. Vettori (Eds.): From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication. Proceedings of the Workshop on the Representation and Processing of Sign Languages. 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisbon. Paris: ELRA, pp. 1-6.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. Construction and Exploitation of Sign Language Corpora. In Crasborn, O. et al. (Eds.) Construction and Exploitation of Sign Language Corpora. Proceedings of 3rd Workshop on the Representation and Processing of Sign Languages. ELRA, Paris, pp. 64-67.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131

Kristoffersen, J. H. & Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language dictionary. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography*. Oxford: Oxford University Press, pp. 293-318.

Langer, G., Troelsgård et al. (2016). Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In Efthimiou, E., Fotinea, E. et al. (Eds.) Corpus Mining: Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages. 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož. Paris: ELRA, pp. 143-152

Pedersen, B.S, Nimb S, Asmussen J. et al., (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation,* 43(3):269-299. Springer, Netherlands.

Zwitserlood, I, Kristoffersen, J and Troelsgård, T (2013). Issues in Sign Language Lexicography. In *Jackson, H. (ed.): The Bloomsbury Companion to Lexicography*, pp. 259–283. London: Bloomsbury Publishing.

Trap-Jensen, L. (2010). Access to multiple Lexical Resources at a Stroke: Integrating Dictionary, Corpus and Wordnet Data. In Granger, S. & Paquot, M. (Eds.), Lexicograhy in the 21st Century: New Challenges, New Applications. Proceedings of eLex 2009, Louvain-la-Neuve. pp 295-302. Cahiers du CENTAL, Louvain-la-Neuve: Presses Universitaires de Louvain.

Troelsgård, T. and Kristoffersen, J.H.(2008) An electronic dictionary of Danish Sign Language. In Müller de Quadros, R (Ed.) Sign Languages: spinning and unraveling the past, present and future. TISLR9, forty-five papers and three posters from the 9th Theoretical Issues in Sign Language Research Conference Florianopolis, Brazil, pp. 352-362. Editora Arara Azul.

9. Language Resource References

Corpus Resources & Documentation. Society for Danish Language and Literature, DSL. Accessed at: http://korpus.dsl.dk/ [20/02/2018]

DanNet [The Danish wordnet]. Accessed at: http://wordnet.dk [28/11/2017]

DGS-Corpus [German Sign Language Corpus]. Accessed at: http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html [9/1/2018]

KorpusDk [The Danish Corpus] Accessed at: http://ordnet.dk/korpusdk [20/02/2018]

Ordbog over Dansk Tegnsprog [The Danish Sign Language Dictionary]. Accessed at http://www.tegnsprog.dk [9/1/2018]