Challenges in linking lexical information and discourse data of signed languages

SignLinC
Deliverable 3

Funded by CLARIN-NL, 2010

Onno Crasborn
Centre for Language Studies
Radboud University Nijmegen

Version 1, February 21, 2011
Introduction

This document briefly characterises the short history of lexical resources and annotated corpora for signed languages. It argues that the two types of resources have typically been developed separately and for separate purposes, leading to resources that are hard to integrate. Moreover, there is no readily available software that integrates the development of lexica and corpus annotation for newly created corpora.

A short history of lexical resources

Sign language dictionaries have been developed for many decades. A major turn was initiated by Stokoe Casterline & Croneberg’s (1965) dictionary on ASL, which employed Stokoe’s (1960) insight that lexical items can be phonologically decomposed just as the words of spoken languages. Since then, paper dictionaries have been developed for many signed languages, and still are. More commonly nowadays, dictionaries are based on databases and published on DVDs and on websites. These dictionaries are without exception bilingual, in that they use the spoken language of the surrounding hearing community (that is also learned by the deaf people in schools) to various degrees for sorting, searching, definitions, and examples of usage.

In most cases, dictionaries have been developed initially to aid hearing second language learners in the acquisition of basic vocabulary. This explains why in many countries, dictionaries started their lives not in an academic setting or in an existing lexicographic institute, but in schools for the deaf, deaf associations, and organisations for parents of deaf children. Nowadays, this situation has diversified quite a bit across Europe and also outside Europe.

Lacking a written culture, it has until recently not been possible to base the development of dictionaries on sign language corpora. This is currently being done in Hamburg for German Sign Language, in an ambitious project that ranges over more than ten years. As sign language corpora of any substantial size themselves are only under development, there have also not yet been any lexical resources independent of sign language dictionaries that were specifically designed to aid the construction of corpora.

A short history of annotated corpora

It is only since the late 1990s that the use of video in standard desktop computers has become possible. Until then, dedicated video cards in high-end desktop computers were needed to capture video from analogue recordings at a maximum resolution of, say, one quarter of PAL resolution (which is 768 x 576 pixels). Before the 1990s, all sign language research was based on the use of analogue video tapes, making a link between annotation and the source media impossible. At the start of the 21st century, very slowly ideas were developed to create corpora of natural discourse similar to speech corpora that were created in the 1990s. While there currently are a few collections of video data that aspire to be linguistic corpora (in Australia, the Netherlands, Sweden, and the UK), none of these has been fully transcribed or annotated at any linguistic level (see also Johnston 2008). There are a few reasons for this, which will be discussed in the next
Current challenges: linguistic and technological

One of the prime challenges in analysing signed languages is that there is no writing system that is in common use by any deaf community at large, let alone that there is a written tradition for any signed language. This is a blessing in disguise in the sense that the written form does not influence the analysis of the actual language use, but for creating linguistic resources this has led to several problems.

First of all, there is no easy to use phonetic transcription system that is met with any consensus among sign language researchers (Miller 2001). While HamNoSys is clearly the best developed and has seen the most revisions over the years, it is still not in common use, partly because of its limited documentation and support. Some would say that its possibilities for encoding the many subtle non-manual activities that occur in natural discourse are too limited. Still, it is the only system that could be compared to the International Phonetic Alphabet for speech.

By consequence, it has become customary to represent sign language by the use of written words in a spoken language that is convenient to use. While this often is the common written language also known to the deaf community, there is no real need for this, and Dan Slobin has advocated that glosses are created in the language of the target audience of a conference or a publication. The problem with glosses is that they presuppose a nice match between lexical items in the signed language and those in the spoken language. At least for western spoken languages, this is far from the case. The lexicon of Dutch, for example, is extremely rich in comparison to that of Sign Language of the Netherlands. Further, their morphological structure is very different. This leads to many problems in finding glosses that match the semantics, the morphology, and the word class of signs. The consequence is large variability in how a sign is glossed. This is not a principled problem per se, but in creating consistent corpus annotations it surely is vital that a given sign is glossed consistently in the same way; this is what Trevor Johnston has called ‘ID glosses’.

Aside from this principled problem, there are no explicit conventions on how to gloss a variety of phenomena such as fingerspelling, sign language gestures, compounds, pointing signs, et cetera. The publication of annotation conventions for recent corpora (see http://www.signlanguagecorpora.org) may lead to gradually increasing consensus in this area in the coming years.

In addition to these linguistic problems, there are also technological challenges to face, especially if we compare the situation to that of creating spoken language resources (whether written or spoken or multimodal). There have only been small efforts in the area of automatic video processing and the detection of sign language activity. Thus, computer-aided transcription is not yet an option for the creation of sign corpora. In fact, as the development of such tools require large manually annotated corpora as training material, little was possible until recently. Ongoing EU projects like SignSpeak (www.signspeak.eu) and DictaSign (www.dictasign.eu) are contributing to this field, but as it stands, highly labor-intensive manual annotation is still the norm.

To create reliable annotations, a link between the annotation tool and a reference lexicon is sorely needed. However, the only tool that currently offers this hard link is
iLex. This annotation and lexicon tool was developed at Hamburg University specifically for the creation of the aforementioned corpus-based dictionary (and previous similar projects). Unfortunately, the tool is not readily available to a wide audience, and it is not well documented and supported. It is therefore vital for projects like CLARIN to support the further development of well-supported and well-documented open source tools like ELAN and LEXUS. The present SignLinC project aims to create some initial links between these two tools, ELAN being the most commonly used annotation tool in sign linguistics, and LEXUS having potential of supporting the development of lexical resources for signed languages.

A role for ISOcat in promoting standardisation

As was made clear above, there are no commonly adopted annotation conventions for signed languages yet, for any level of annotation. This makes ISOcat (www.isocat.org) of extra interest to the sign linguistics community. Many ISOcat categories, related to lexical semantics for instance, can apply to signed languages as to spoken languages. For others, ISOcat in combination with references to it in ELAN annotation documents can form a useful context for discussing and developing new standards. A recently created thematic domain group 'Sign Language' within ISOcat has the ambition to bring together sign language experts with an interest in sign language resources (Crasborn & Windhouwer, 2011) to make proposals that can be discussed in the sign linguistics community. The possibility of using ISOcat data categories in LEXUS would make it extra useful for creating sign language resources.

Conclusion

There are both technological and linguistic challenges in relating corpora of signed discourse with lexical resources. Some of these are hard to overcome as they relate to our state of knowledge of signed languages and the fact that there is no written tradition among deaf communities. On the other hand, many problems can be addressed by publishing and sharing adequate resources so as to promote standardisation, and by continuing development of existing tools. By doing the latter, they can be made more useful to a specific community like sign language researchers which in turn will lead to higher quality resources that can also be used by the language communities themselves and by other subdisciplines in the humanities and social sciences. It is here that CLARIN can make an important contribution.

References