SignLinC: requirements and desiderata for the CLARIN infrastructure

SignLinC
Deliverable 7

Funded by CLARIN-NL, 2010

Onno Crasborn
Centre for Language Studies
Radboud University Nijmegen

Version 3, March 8, 2011
Introduction

The following wishes for the CLARIN infrastructure have emerged from discussions between partners and from experience during the project with ELAN and LEXUS. They are categorised in three groups: requirements, which are vital for CLARIN; desiderata, which are useful extensions of the infrastructure for broad groups of users; and wishes, which are specifically interesting for the sign language field.

Requirement: link from Language Archive to LEXUS

- A true hyperlink from the lexicon metadata description in the corpus archive to LEXUS needs to be created (for SignLinc: http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI1075267%23). Currently, the metadata for lexica in the corpus browser barely fulfill a function, as it is not clear what and where LEXUS is when browsing the corpus archive, nor who actually has access – unless this is described in text in the metadata description.

Desiderata: LEXUS

- A stand-alone version of LEXUS would be most welcome. Even though our researchers are online 90% of the time, some do work with poor network connections, or ‘in the field’ without any network connection (making recordings at a deaf club, for example, or in family homes).
- The link between lexica and multimedia corpora could be further improved by enabling the initiation of searches in annotated corpora starting from a lexicon entry. Further, lexical entries could be enriched by making it easy for users to select from the search results a number of typical instances of the lexical item in connected discourse or different utterances that illustrate different meanings of the lexical item.
- On a larger scale, harvesting annotated corpora for instances of a lexical item and including statistics on those results in the lexical entry would give the user much more information about the use of lexical items, without having to manually perform searches per entry.
- Allow for an import function for other than Toolbox files, if only for tab-separated text or CSV files, so that one can make a quick start with using LEXUS. This, too, could be vital for the actual adoption of LEXUS by a significant user group.
- Further development of the LEXUS interface is clearly necessary. The integration of movies could be improved by using thumbnails for movie files for instance, or enabling the user to set the size of the media window in the preferences.
- Allow for an “Export ECV” function that generates an External Controlled Vocabulary that ELAN Linguistic Types can refer to.

Desiderata: ELAN
• Extend the functionality of the lexicon connection in ELAN: allow user to flexibly switch between lexicons, without having to specify a lexicon connection for a specific Linguistic Type.

Desiderata: integration between ELAN and LEXUS

• Dynamically load an External CV on the basis of a LEXUS lexicon; or in other words, compose the drop down list with ECV suggestions on the basis of a lexicon query, rather than a query in the ECV file.
• Jump to the LEXUS lexicon on the basis of a selected Entry from that lexicon in the lexicon tab (for instance: open LEXUS in browser upon double-clicking a search result).
• Allowing for the creation of a new LEXUS lexicon based on the types of annotations occurring on a specific tier(s) in ELAN. This would likely dramatically increase the user base of LEXUS, as users would immediately be able to start enriching their lexical resource based on what actually occurs in their corpus.
• Adding entries to a LEXUS lexicon based on a selected annotation. In this way, it would be easier to enrich lexical resources on the basis of discourse data, and the initial instance on which the addition of the lemma was based could be referred to in a field in the lexicon.

Desiderata: ethical issues

• The public availability of all metadata encoded in CMDI is desirable from the point of view of the promotion of data exchange and re-use, but it causes difficulty for privacy-sensitive metadata properties of individuals. This especially holds for children and for members of small communities, in which the smallest piece of information can already reveal the identity of individuals. This holds for many deaf communities, for example. For the SignLinC project, this has not been of direct concern for the new lexical data, but for sign language corpora that would be connected to lexical information in LEXUS this is clearly an issue. The practical solution is to store all sensitive data in password-protected PDF documents which are only accessible to the corpus managers, for instance. However, this prevents researchers from including the information in that file in metadata searches. We urge CLARIN to develop alternatives for this problem, such as enabling enriched copies of CMDI files that can be protected.

Wishes

• More LMF structures for signed languages. With ELAN, too, the actual use of the tool by many research groups in the same field has led to the exchange of ETF (template) files, which is currently stimulating a drive towards standardization. All of the desiderata listed above will also contribute to the attractiveness of LEXUS for existing users of ELAN.