Linking Corpus NGT annotations to a lexical database using open source tools ELAN and LEXUS

Onno Crasborn[1], Micha Hulsbosch[1,2], Han Sloetjes[2]
[1] Centre for Language Studies, Radboud University Nijmegen
[2] Max Planck Institute for Psycholinguistics
[1] PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
[2] PO Box 310, NL-6500 AH Nijmegen, The Netherlands
E-mail: o.crasborn@let.ru.nl, m.hulsbosch@let.ru.nl, han.sloetjes@mpi.nl

Abstract

This paper describes how we have made a first start with expanding the functionality of the ELAN annotation tool to create a bridge to a lexical database. A first lookup facility of an annotation in a LEXUS database is created, which generates a user-configurable selection of fields from that database, to be displayed in ELAN. In addition, an extension of the (open) controlled vocabularies that can be specified for tiers allows for the creation of very large vocabularies, such as lexical items in a language. Such an 'external controlled vocabulary' is an XML file that can be published on any web server and thus will be accessible to any interested party. Future development should allow for the vocabulary to be directly linked to a LEXUS database and thus also allow for access right management.

Keywords: sign language, annotation, Corpus NGT, lexicon, ELAN, LEXUS, controlled vocabulary

1. Introduction

Since the public release of the media files of the Corpus NGT[1] in December 2008, a subset of the data have been provided with gloss and translation annotations in the context of various research projects. During that process, it gradually became clear that glossing is not possible without a lexicon of ID-glosses (see Johnston 2008 for discussion). The creation of that lexicon is described by Crasborn & de Meijer (this volume). The present paper focuses on the facilitation of the process using a combination of the CLARIN standard tools ELAN[2] and LEXUS[3] created by the MPI for Psycholinguistics, a spreadsheet programme, and custom-made Perl scripts.

2. Displaying information from LEXUS in ELAN

As sign language glosses are always in the form of words from a spoken language, it is important that these glosses are consistently linked to lemmata or full forms in a lexicon. As the open source multimodal annotation tool ELAN does not have a lexicon function built in (as opposed to iLex[4], for instance), an effort was undertaken to create a bridge to the open source lexicon tool LEXUS. The first steps for this were taken in the CLARIN-NL project SignLinC (Crasborn, Hulsbosch, Sloetjes, Schermer & Harmsen, 2010), and this functionality has since been expanded.

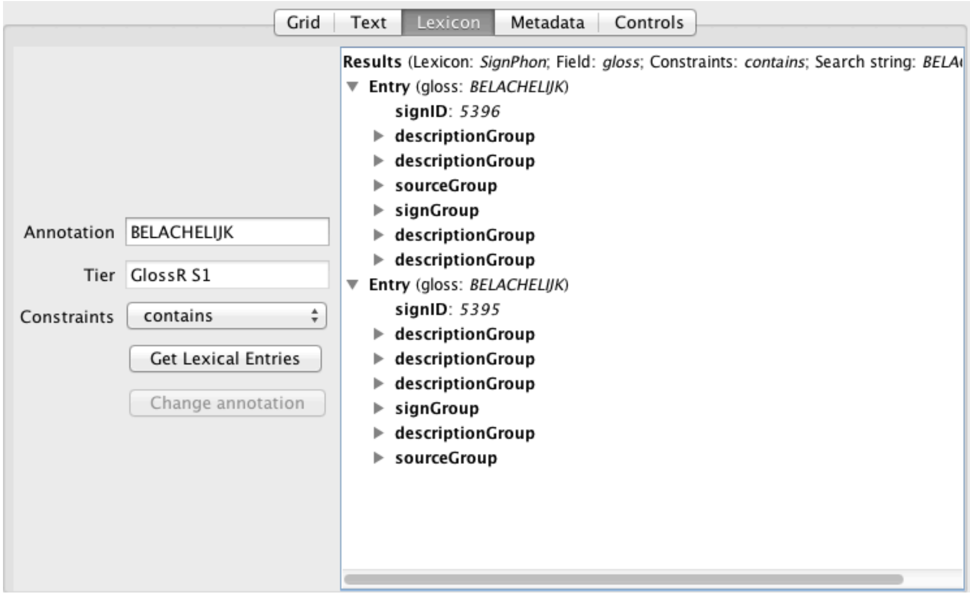
For SignLinC, a lexicon tab in ELAN was created, in which properties of lexical items can be displayed after they have been entered in LEXUS. Figure 1 presents an example of two lexical items in the lexicon tab that both contain the contents of a selected gloss. They have been generated by a lookup in the SignPhon lexicon (which is used here just for demonstration purposes). The resulting hits are displayed with their hierarchical structure, so that the desired information can be quickly selected from a large list of properties. Figures 2-4 illustrate the configuration of this service in ELAN. The actual link to a LEXUS lexicon requires logging in to LEXUS, so that a list of accessible lexica is presented to the user.

In practice, this link works as long as the gloss of a sign is identical to the top-level field in LEXUS: an online lookup is done on the basis of the text string that is in the ELAN annotation. However, there is no such live link between LEXUS and the creation of new annotations. This would not be trivial to develop, in part because ELAN is a stand-alone programme while LEXUS currently is a web-based tool. To avoid the associated complexities, we have created an alternative solution, that ideally will be replaced by a further developed bridge between ELAN and LEXUS. It is MPI's intention to create a stand-alone version of LEXUS. This would facilitate the development of further interaction between ELAN and LEXUS.

3. Defining an external controlled vocabulary

Instead of a direct connection to LEXUS for the creation of new lexical annotations, an 'external controlled vocabulary' (ECV) can now be defined. The ECV file itself is a fairly simple XML file that needs to be published on a web server. It is added to a file in the same way as other controlled vocabularies (Figure 5), and the same options can be applied, including assigning a specific colour to a specific item. Like a regular CV, an ECV can be linked to a 'linguistic type' specification for a tier. Unlike other controlled vocabularies, only the values that are actually used in a file are stored in the EAF annotation document. Also, each item in an ECV is identified by an XML ID which is stored as an attribute of annotations referring to an ECV item. The value of an item is stored in the EAF to have it immediately available for visualisation and for searching purposes,

[1] http://www.ru.nl/corpusngtuk
[2] http://www.lat-mpi.eu/tools/elan
[3] http://www.lat-mpi.eu/tools/lexus
[4] http://www.sign-lang.uni-hamburg.de/ilex



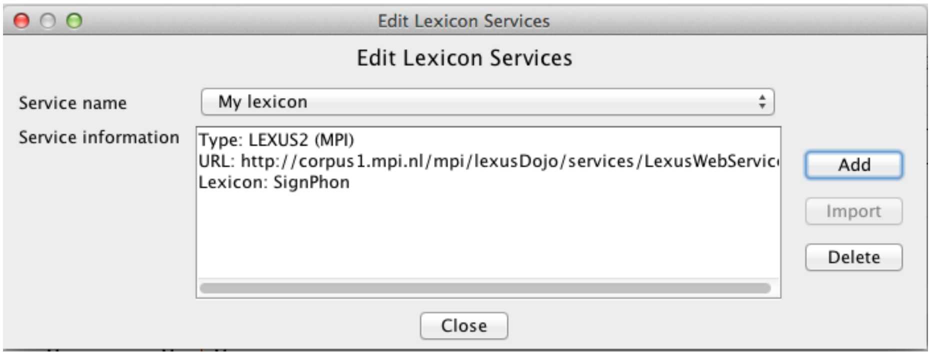
Annotation BELACHELIJK

Tier GlossR S1

Constraints contains

```
Results (Lexicon: SignPhon; Field: gloss; Constraints: contains; Search string: BELA
▼ Entry (gloss: BELACHELIJK)
    signID: 5396
  ▶ descriptionGroup
  ▶ descriptionGroup
  ▶ sourceGroup
  ▶ signGroup
  ▶ descriptionGroup
  ▶ descriptionGroup
▼ Entry (gloss: BELACHELIJK)
    signID: 5395
  ▶ descriptionGroup
  ▶ descriptionGroup
  ▶ descriptionGroup
  ▶ signGroup
  ▶ descriptionGroup
  ▶ sourceGroup
```

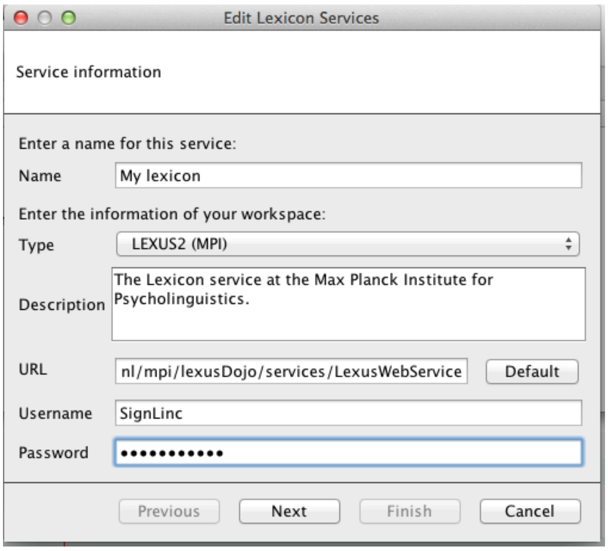
Figure 1: A search for the gloss ‘BELACHELIJK’ yields two hits in the LEXUS database that is specified for the linguistic type of the gloss tier



Service name My lexicon

Service information
Type: LEXUS2 (MPI)
URL: http://corpus1.mpi.nl/mpi/lexusDojo/services/LexusWebServic
Lexicon: SignPhon

Figure 2: Adding a new lexicon service



Enter a name for this service:

Name My lexicon

Enter the information of your workspace:

Type LEXUS2 (MPI)

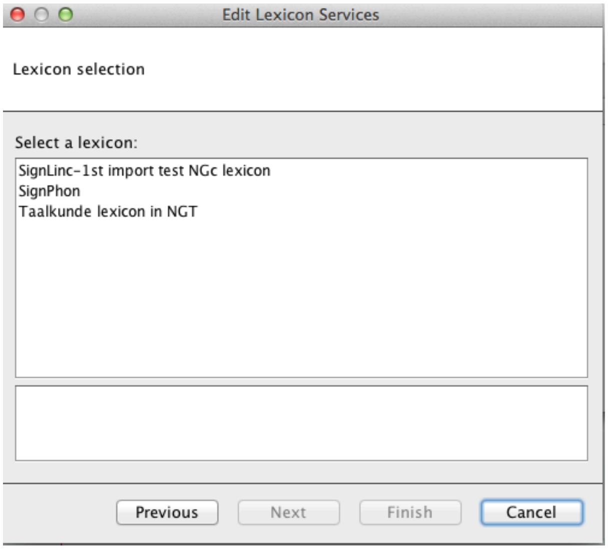
Description The Lexicon service at the Max Planck Institute for Psycholinguistics.

URL nl/mpi/lexusDojo/services/LexusWebService

Username SignLinc

Password •••••••••••

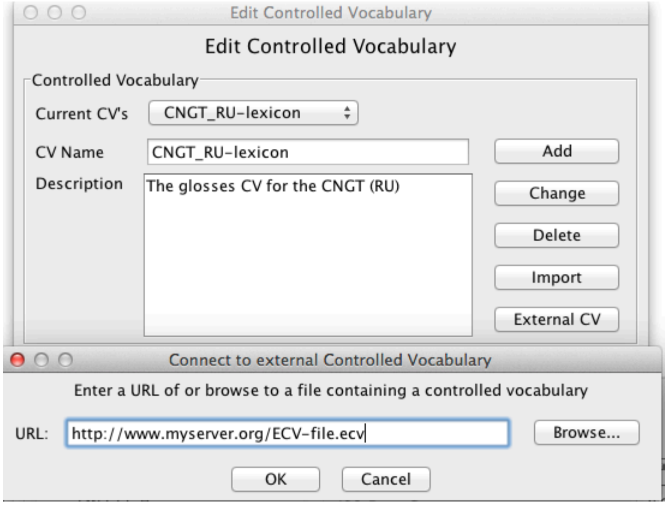
Figure 3: Logging in to LEXUS to create a new lexicon service in ELAN



Select a lexicon:

SignLinc-1st import test NGc lexicon
SignPhon
Taalkunde lexicon in NGT

Figure 4: Selecting a lexicon


Edit Controlled Vocabulary
Edit Controlled Vocabulary
Controlled Vocabulary
Add
Change
Delete
Import
External CV
Connect to external Controlled Vocabulary
Browse...
OK
Cancel

Current CV's: CNGT_RU-lexicon

CV Name: CNGT_RU-lexicon

Description: The glosses CV for the CNGT (RU)

Enter a URL of or browse to a file containing a controlled vocabulary

URL: http://www.myserver.org/ECV-file.ecv

Figure 5: Specifying the URL for an ECV

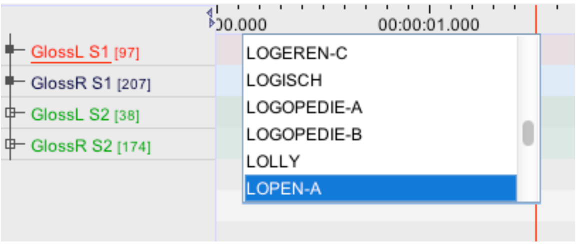
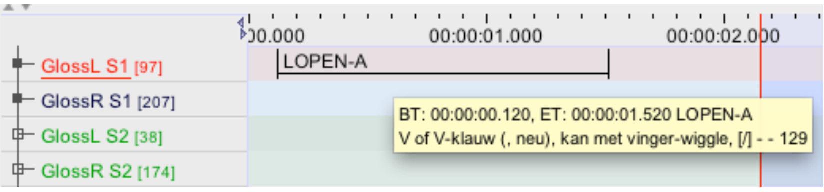
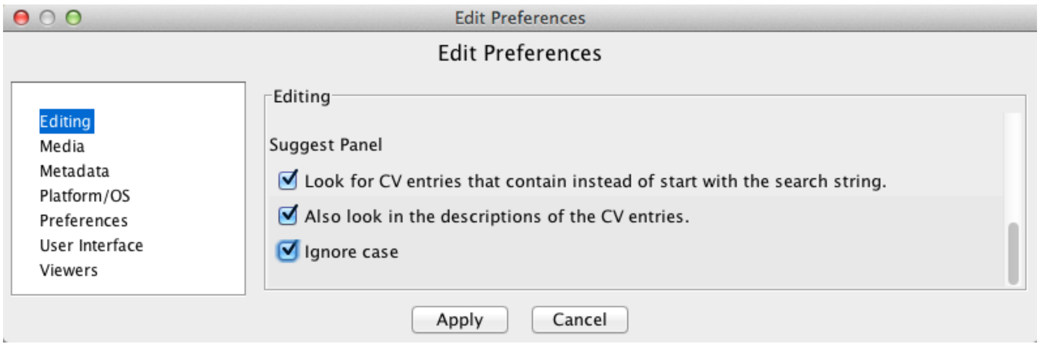


Figure 6: The suggest panel for an ECV



BT: 00:00:00.120, ET: 00:00:01.520 LOPEN-A
V of V-klauw (, neu), kan met vinger-wiggle, [/] - - 129

Figure 7: Inspecting the description properties through the tooltip



☑ Look for CV entries that contain instead of start with the search string.

☑ Also look in the descriptions of the CV entries.

☑ Ignore case

Figure 8: Specifying preferences for the behaviour of the suggest panel

but the XML ID is used to validate and update the values in the EAF in order to keep them in sync with the vocabulary (and eventually with the lexicon).

When a user creates an annotation on a tier with a CV or ECV link, a 'suggest panel' appears, offering the items from the vocabulary as suggestions (Figure 6). These suggestions can be overridden by the user if necessary.

The key advantage of an external CV is that its contents can be centrally maintained for a large set of annotation documents or even multiple corpora, so that there is no risk that the CV list starts to diverge in different documents.

4. Expanding the functionality of the ECV

Subsequent development of the ECV functionality in ELAN has led to two important improvements for the user.

4.1 Displaying additional information for annotations with an ECV link

The description field in the ECV is now visible in the tooltip that is displayed in ELAN when the mouse hovers over an annotation that has a link to an ECV entry (Figure 7). For the Corpus NGT, this description is filled with information from the lexicon, so that users have access to the phonological form of the lexical item as well as lexico-semantic properties. It has thus become easier for the user to inspect whether the selected ID-gloss actually refers to the sign in question. Current development is targeted at presenting this description property also in the drop-down list that users get to see upon creating a new annotation (the 'suggest panel'). Only in that way mistakes in the selection of the right gloss can be prevented, presupposing that the annotator can read the text string that represents phonological description of a sign.

4.2 Facilitating the selection of items from a large ECV

Secondly, users can specify in the preferences what is displayed in the suggest panel: the characters that are typed in to search elements in the ECV to be displayed in the list can be specified to match the start of the item (default), any text in the item, and/or also information in the description field (Figure 8).

It is especially the latter function that represents an important step forward in ensuring that users create correct ID-gloss annotations. As signs are recognised more easily on the basis of meaning rather than form, there is a natural tendency to want to translate the sign in order to create a gloss, rather than to select the ID-gloss from a list. As there are typically different Dutch translations of an NGT sign, this can lead to different glosses for the same sign. By storing potential translation variants of signs in the description field of ECV items, typing in a string like 'area' will also return a gloss like SPACE at the top of the suggest panel, alerting the annotator to the fact that AREA is not the ID-gloss that is listed in the lexicon.

4.3 Impact for the workflow of the annotation of the Corpus NGT

These initial lexicon-like facilities in ELAN have led to a workflow in the annotation of the Corpus NGT where both the glossed part of the corpus and the related lexicon grow at the same time. As soon as a significant number of new ID glosses are added to our Excel table and described in terms of phonological categories, translation variants, and homonyms, the ECV list is updated using a Perl script that runs on the text export of that table, and another Perl script double-checks that all instances of glosses that appear in the new ECV are assigned an ECV reference, and will thus display the description field in the timeline viewer in ELAN. From that point onwards, changes in either the gloss string or in the description field can be made in either the Excel table, and with the first following update will be visible in all instances of that gloss in any annotation document.

5. Conclusion & future developments

The features described in this paper have created a workflow in which ID-glosses can be created on the various gloss tiers in a more reliable way. The mere fact that the list of currently agreed-upon glosses is available upon the creation of a new gloss annotation reminds annotators of the conventions that apply and of the fact that multiple glosses may apply to the sign form at hand. At the same time, the suggest panel still remains a list of words (the essence of an ID-gloss), and does not yet provide phonological or semantic information that can help the annotator in selecting the right gloss. Presenting the information from the description field as part of the suggest panel would form a next major step in improving gloss consistency and reliability.

As far as the Corpus NGT data are concerned, the next step in this development will be that the elementary lexicon on which the ECV is based is converted to a LEXUS database. Only after such a conversion will users be able to access all the information from the lexicon in the lexicon tab. Here, data are presented in a more structured view than in the gloss tooltip in the timeline viewer.

More substantial development of LEXUS and ELAN will be necessary to facilitate the updating of the ECV based on information in LEXUS, or alternatively to merge the functionality of the ECV and LEXUS by letting ELAN generate the items in the suggest panel for a new annotation directly from a LEXUS database. Adding of lexical items to a database and modifying existing items should in the end be an integrated part of the corpus annotation process, creating a coherent set of resources for a language.

6. Acknowledgements

The developments described in this paper were made possible by CLARIN-NL grant 09-003 'SignLinC' and ERC starting grant 210373 'On the Other Hand' awarded to Onno Crasborn.

7. References

Crasborn, O., Hulsbosch, M., Sloetjes, H., Schermer, T. & Harmsen, H. (2010) Linking ELAN and LEXUS. Poster presented at the *Fourth Workshop of the Sign Linguistics Corpora Network*, Berlin. http://www.ru.nl/slcn

Crasborn & de Meijer (this volume). From corpus to lexicon: the creation of ID-glosses for the Corpus NGT.

Johnston, T. (2008). Corpus linguistics and signed languages: no lemmata, no corpus. In O. Crasborn, E. Efhimiou, T. Hanke, E. Thoutenhoofd, & I. Zwitserlood (Eds.), *5th Workshop on the Representation and Processing of Signed Languages: Construction and Exploitation of Sign Language Corpora* (pp. 82-87). Paris: ELRA.