Improving the exploitation of linguistic annotations in ELAN

Onno Crasborn (o.crasborn@let.ru.nl)
Radboud University Nijmegen, Centre for Language Studies
PO Box 9103, 6500 HD Nijmegen

Han Sloetjes (han.sloetjes@mpi.nl)
The Language Archive, Max Planck Institute for Psycholinguistics, Wundtlaan 1
6525XD Nijmegen, The Netherlands

Keywords: Concordance; collocation; multimodality; annotation tool; gesture; sign language; multilingual metadata; multilingual annotations

1. Introduction

This paper discusses some improvements in recent and planned versions of the multimodal annotation tool ELAN[1], which are targeted at improving the usability of annotated files. ELAN is being developed by The Language Archive, a unit of the Max Planck Institute for Psycholinguistics, and evolved round the turn of the century from a few smaller tools and applets. It is a tool for annotating audio and video recordings and is built on a tier-based data model in which a tier represents a kind of layer that groups annotations of the same type. Tiers can be organised hierarchically so that multi-layered annotation of primary data is supported. The annotation data are stored in an XML format, EAF (ELAN Annotation Format). ELAN is a versatile tool that is used in a variety of disciplines such as gesture research, documentary linguistics, behavioural studies, and sign language research. As it is becoming more and more common to share research data, both in terms of larger corpora and in the form of individual files containing data of specific research publications, the question of how these data can be exploited is becoming more and more important. The usability is addressed from two different perspectives. First of all, the access to annotated files requires that the user masters the language that is used in the annotation document. While the interface of ELAN is multilingual (counting eleven languages in version 4.6.1 released in early 2013, including e.g. Mandarin Chinese, Russian, and English), annotations are typically created in only a single language. The creation of multilingual vocabularies is one step in broadening the potential range of users, and efforts are underway to facilitate such vocabularies. Moreover, specifying the language of the speaker/signer at the tier and/or the annotation level allows for fine-grained specification of the language used by the participant. Secondly, various enhancements have been made to the multiple file search functionality, aimed both at broadening the possibilities for formulating queries and at improving the visualisation of search results.

These two aspects of usability will be discussed in turn below.

2. Multilingual vocabularies

As a tool for linguistic annotation in the broadest sense, ELAN supports any Unicode compatible language. But so far it does not supply dedicated means to specify which language is used in a particular annotation or even on a tier as a whole. Tiers do have an optional property that has traditionally been labelled “Default Language”, but that in fact only identified an input method (like a specific writing system) to use when annotating on that particular tier. Although sometimes there is a relation between an input method and the language annotated, equally often there is not. The number of available input methods is also very limited. Dedicated language identifiers are therefore being introduced on different levels of ELAN documents: on the tier-level they can be used to specify the overall language used in the annotations of that tier and on the annotation level to specify a per instance language. At the annotation level, language identification can be useful for annotating and investigating code switching or for marking loan words.

This improves the already existing possibility of multilingual annotation: doing transcription on one tier and translation into one or more major languages on depending tiers. The availability of dedicated language identification, when used, makes it easier to compare resources across files and across corpora.

A different approach to multilingualism in annotations is one based on multilingual vocabularies. ELAN has since long provided the feature of controlled vocabularies, flexible lists of values that can be applied to annotations. In the context of a previous CLARIN-NL project, SignLinc, this facility has been extended with the possibility to store the vocabulary outside of the annotation document, giving it a more stable and more controlled status (Crasborn, Hulsbosch, & Sloetjes, 2012). In the current effort these vocabularies have been extended such that they can have multiple values for multiple languages. Since an ID identifies each entry of a vocabulary and annotations can link to this ID, it is possible to switch the displayed language without

[1] http://tla.mpi.nl/tools/tla-tools/elan/

necessarily changing the annotation value in the annotation document. The editing of multilingual vocabularies is illustrated in Figure 1. In this example, a (test) vocabulary named 'BegripnaamCV' is being produced to be available in three languages: Dutch (nl), English (en), and French (fr). Colours can be attached to individual vocabulary items to highlight them in annotation documents, such that the same colour will be linked to an annotation irrespective of the language chosen for display.

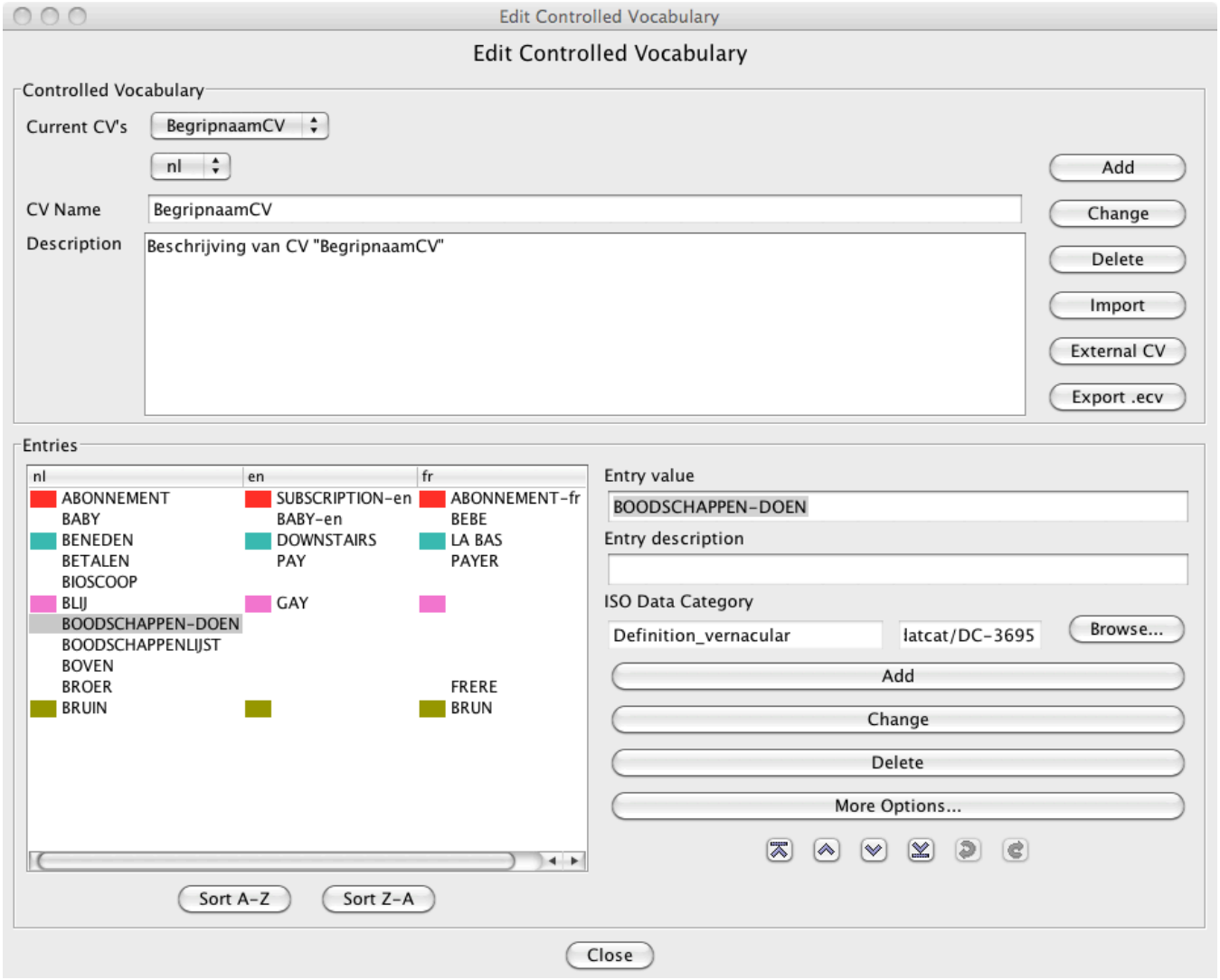

Figure 1. Editing of a multilingual vocabulary

Similarly, for annotations and other elements that can be linked to an ISOcat data category, users will be able to select a language from a set. The language sections of ISOcat allow for storing definitions and descriptions in multiple languages, but currently only the default, mandatory English section is used within ELAN. By introducing a global language setting the user can specify which version of resource she or he would like to be displayed. This pertains to items in an external controlled vocabulary, to elements of CMDI metadata, and to descriptions of those annotations and tiers that have been linked to a data category. Above that, it is possible to override the global language setting by selecting a language per document or even per individual component (viewer), making the full potential of multilingual resources easily accessible to the user.

3. Multi-layer concordances

Concordance tools and the resulting collocations are a core part of corpus linguistics (Sinclair 1991). They give us a view on the use of lexical items in context, on multi-word expressions, fixed complements with verbs, etc. Concordance tools assume that data are ordered in a single uni-layer string, as is common with written text or phonetic transcriptions. Multimodal data have an inherently multidimensional organization: events are not only ordered sequentially on a single axis, but they happen simultaneously, can each have different durations, and show a specific temporal alignment with respect to each other (Knight & Tennant, 2010). Frequently investigated examples are eye gaze and other non-verbal behaviour accompanying speech (e.g. Kendon 1967), the timing of head movements with respect to the manual activity in sign languages and in gesture (e.g. McClave 2002, Crasborn & van der Kooij 2013), and the activity of the two hands in both sign and gesture (e.g. Vermeerbergen et al. 2007).

ELAN annotations can be created on an unlimited number of tiers, which can be nested in complex ways. The latest versions of the tool already offered several search functions allowing for the extraction of certain patterns from larger corpora, but these were limited in several respects. Most importantly, while multilayer searches for complex types of alignment were possible, the visualisation of the resulting 'multilayer collocates' was a complex text string that was hard to parse as it lacked information on the timing of the events on different layers. Search hits either had to be browsed one by one by inspecting and interpreting the accompanying time codes from a tooltip, or could be exported to text files for further processing by other means.

Structured Search in Multiple Files

ELAN currently contains several search functions, most of which are also available in TROVA, the search engine that can be used to query annotation files as well as other types of files, in the online corpora. Search hits in the online annotation files are displayed in the ANNEX browser tool. Searches can be performed within a single file, but also across sets of files that users can compose themselves. Different types of complexity in multi-file searching cater different users and different purposes, where in the simplest version the user can search for strings in the set of all annotations, and in the most complex version the user can specify not only the tiers or sets of tiers on which annotations in the query must occur, but also temporal alignment patterns of annotations on different tiers in combination with sequential patterns within tiers. For an example of a complex query, see Figure 2.

In the example query in this screen shot, a combination of gloss annotations in a sign language document is described: a sequence of the annotations 'PT' (a pointing sign) and 'PO' (the palm up gesture) on any gloss tier, where the 'PT' annotation must overlap in time with another 'PT' annotation on any other gloss tier.

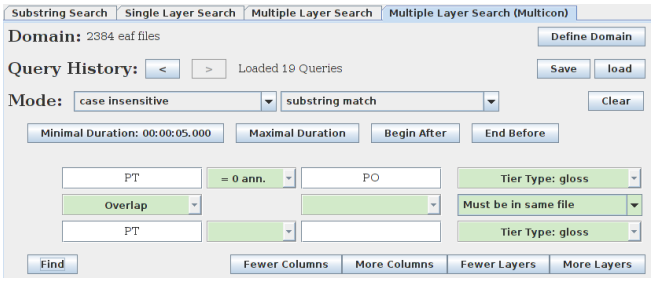

Figure 2. Example of a complex query for gloss annotations on two different gloss tiers

In the following sections, we describe the functionality that has been recently added and that will be published in the spring of 2013.

Multiple search restrictions per layer

In the single layer and multiple layer search functionalities, it is possible to specify per layer in which tier(s) to search. The tier selection can either be the name of a single tier or it can be a collection of tiers based on one of the following three tier properties: tier type (referred to as Linguistic Type in the ELAN annotation format, EAF), participant, or annotator. Although these selection criteria proved to be already quite useful in various situations, they were also too restrictive and not flexible enough in many others. A user might want to search in all tiers of participant A and B but not in those of C (maybe including the subjects while excluding the interviewer), or in the tiers of participant A and B but only in those of a specific type. There are many situations in which a more powerful selection mechanism is required with, as the last resort, the possibility of complete custom selection of individual tiers. Behind the scenes, hidden from the user, all tier selection options result in a list of tiers per search layer that have to be matched with tiers selected in another layer.

Visualising temporal alignment

Finding overlapping patterns on multiple layers (tiers) has been possible for several years, but the results were represented in sequence on a single line per search hit. This is illustrated in Figure 3, presenting some of the results for a query similar to that in Figure 2.

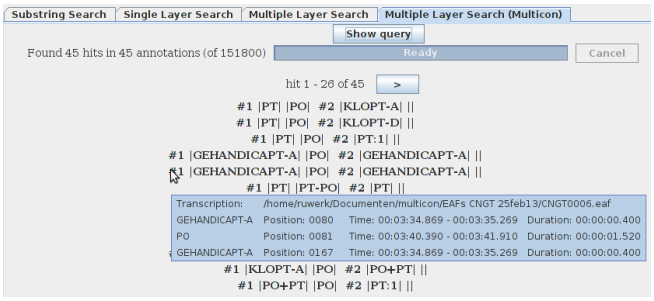

Figure 3. Traditional presentation of the results of a complex query

The user interface allows for the specification of a query for many temporal properties both within and between tiers. Some of these properties can be seen in Figure 2. In a grid of a customisable number of rows and columns, search strings can be entered in the cells and the requested relations between the constituents can be selected in drop-down lists. Examples of these temporal relations between annotations are 'fully aligned', 'overlap', 'left overlap', 'occurring more than 3 seconds after this', etc. The (intuitive) way in which the relations between the query parts are visually represented is not repeated in the display of the hits, as can be seen by the contrast between Figure 2 and Figure 3. As a result, the only way to get an impression of how the annotations in the hits relate to each other in terms of their temporal alignment, is to look at the time information of each annotation, either in the tooltip of a hit or, after export, in a spreadsheet application. This is rather cumbersome way of assessing temporal relations. For that reason, the hit representation in ELAN and TROVA has been enriched by, on the one hand, a matrix similar to the one in the query construction part, which makes the relation between a query part and the corresponding annotation in the hit clear at a glance, and on the other hand a graphical depiction of the annotations on horizontal bars, giving an immediate overview of their relative temporal positions. Both properties of the new hit representation can be observed in Figure 4. Together with the option of hiding the search panel that is normally displayed above the results, a large number of alignment patterns can be displayed simultaneously.

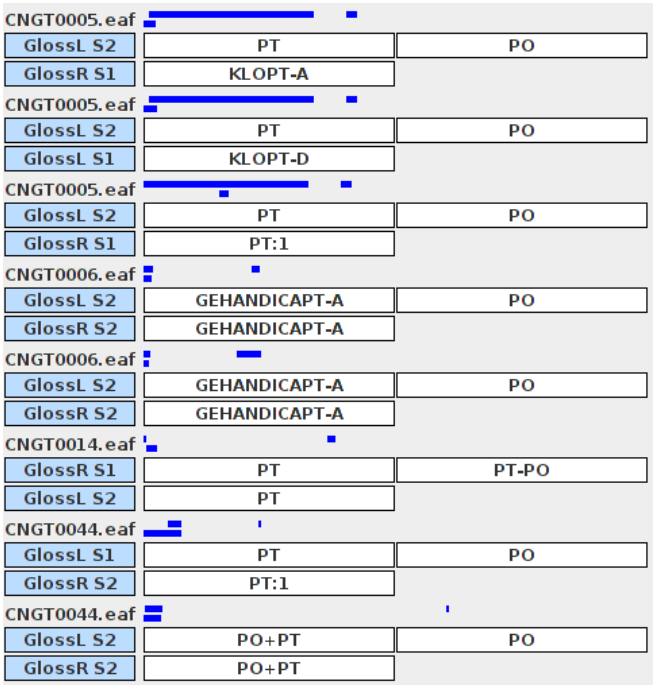

Figure 4. Visualisation of search hits by a matrix view that conforms to the query specification, including the visualization of alignment of annotations

Visualising tier properties

In the traditional display of the hits in the structured search in ELAN and TROVA, only the annotation content is shown, in a style conforming (where applicable) to the Key Word In Context (KWIC) tradition. Additional information like the transcription file name and its path, tier properties, time information, et cetera, are only visible in the tooltips or info tooltips when hovering the mouse over the hits. Options have now been added to the result display to include columns for the above information. The columns can be switched on or off individually. A selection of columns is shown in Figure 5. The initial approach of showing as much of the annotation content as possible is still a valid one, but often more information can fit on the screen, especially with the high-resolution wide screen displays used nowadays. To be able to have the extra information visible directly alongside the annotations in the hits is a huge advantage, as it makes interpretation of the results much more convenient.

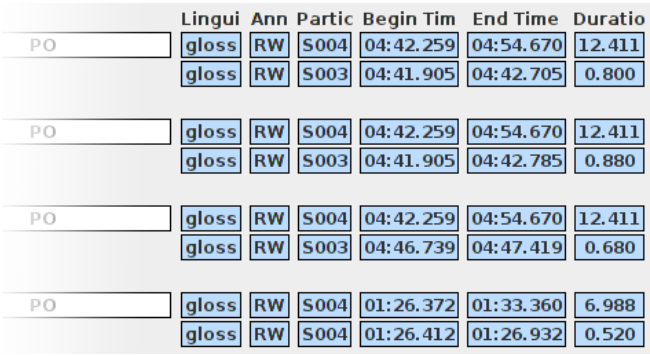

Figure 5. Display of various types of information on search results in columns, here displaying Linguistic Type, Annotator, Participant, Begin Time, End Time, and Duration

4. Conclusion

The developments described in this paper form part of the continuing development of ELAN and related tools by the Max Planck Institute for Psycholinguistics in collaboration with sign language researchers at Radboud University. As in previous projects (e.g. Crasborn, Hulsbosch & Sloetjes 2012), we expect the new functionality to be fine-tuned in the coming year as users start employing it for various purposes and on the basis of various corpora.

Acknowledgments

The work described in this paper is was funded by CLARIN-NL (project numbers CLARIN-NL-11-018, 'Multilayer Concordance Functions in ELAN and ANNEX' (MultiCon), resp. CLARIN-NL-12-020 'Exploiting

ISOcat's Language Sections in ELAN and ANNEX' (EXILSEA)).

References

Auer, E., Russel, A., Sloetjes, H., Wittenburg, P., Schreer, O., Masnieri, S., Schneider, D., & Tschöpel, S. (2010). ELAN as flexible annotation framework for sound and image processing detectors. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Paris: ELRA.

Crasborn, O, & van der Kooij, E. (2013). The phonology of focus in Sign Language of the Netherlands. *Journal of Linguistics*, 49-3, 515-565.

Crasborn, O., Sloetjes, H., Auer, E., & Wittenburg, P. (2006). Combining video and numeric data in the analysis of sign languages with the ELAN annotation software. In *C. Vettori (Ed.), Proceedings of the 2nd Workshop on the Representation and Processing of Sign languages: Lexicographic matters and didactic scenarios (pp. 82-87)*. Paris: ELRA

Crasborn, O., Hulsbosch, M., & Sloetjes, H. (2012). Linking Corpus NGT annotations to a lexical database using open source tools ELAN and LEXUS. Paper presented at the *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, Istanbul, Turkey.

Johnston, T. (2010). Adding value to, and extracting of value from, a signed language corpus through secondary processing: implications for annotation schemas and corpus creation. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. Paris: ELRA

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica,* 26, 22-63.

Knight, D, S. Bayoumi, S. Adolphs, S. Mills, T. Pridmore, & Carter, R. (2006). Beyond the text: building and analysing multi-modal corpora. In*: Proc. 2nd International Conference on E-Social Science.*

Knight, D. & Tennant, P. (2010). Introducing DRS (The Digital Replay System): A tool for the future of Corpus Linguistic research and analysis. *Proceedings of LREC 2010*, Malta.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics,* 32, 855-878.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stehouwer, H., & Auer, E. (2011). Unlocking language archives using search. In C. Vertan, M. Slavcheva, P. Osenova, & S. Piperidis (Eds.), *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage.* Hissar, Bulgaria, 16 September 2011 (pp. 19-26). Shoumen, Bulgaria: Incoma Ltd.

Vermeerbergen, M., Leeson, L., & Crasborn, O. (Eds.). (2007). *Simultaneity in signed languages: form and function.* Amsterdam: John Benjamins.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.* Paris: ELRA.