New Multilayer Concordance Functions in ELAN and TROVA

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

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

Abstract

Collocations generated by concordancers are a standard instrument in the exploitation of text corpora for the analysis of language use. Multimodal corpora show similar types of patterns, activities that frequently occur together, but there is no tool that offers facilities for visualising such patterns. Examples include timing of eye contact with respect to speech, and the alignment of activities of the two hands in signed languages. This paper describes recent enhancements to the standard CLARIN tools ELAN and TROVA for multimodal annotation to address these needs: first of all the query and concordancing functions were improved, and secondly the tools now generate visualisations of multilayer collocations that allow for intuitive explorations and analyses of multimodal data. This will provide a boost to the linguistic fields of gesture and sign language studies, as it will improve the exploitation of multimodal corpora.

Keywords: Concordance; collocation; multimodality; annotation tool; gesture; sign language

Introduction

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).

Multimedia annotation tool ELAN

The tool for multimedia annotation ELAN has been developed and maintained by the Max Planck Institute for Psycholinguistics for more than ten years now. It has a substantial global user base in both psychology and linguistics. ELAN is especially designed to encode and display the multilayer activity we can observe in visual data, whether stemming from hearing or deaf communication (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes 2006). Annotation files can be combined with numeric files, and annotations can be created on the basis of such time series data (Crasborn, Sloetjes, Auer, & Wittenburg 2006). Most recently, the tool has been expanded to allow for the use of automatic recognition techniques working on audio or video signals in generating annotations (Auer et al. 2010).

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.

Exploration tools ANNEX and TROVA

ANNEX is the tool for online visualisation and exploration of annotation data stored in The Language Archive. TROVA is the online annotation content search engine that allows users to execute complex queries on a selection of annotation resources stored in the archive (Stehouwer & Auer 2011). Individual TROVA search hits can be displayed by ANNEX, but the user is faced with similar limited visualisation possibilities of lists of search hits as in the offline tool ELAN.

Goal of this paper

In the present paper we describe a series of new functions in ELAN that expand the types of multilayer searches that can be performed in ELAN and TROVA, and that give the user more visual control over the resulting hits. Overall, these functions are intended to improve the ease, precision, and quality of the analysis of annotated media files, reducing the

large number of steps that are sometimes needed now (Johnston 2010). All of the features that are described in the sections below have been implemented for both ELAN and TROVA.

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 1.

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.

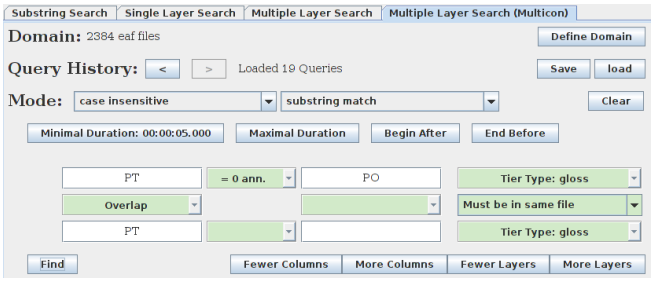

Figure 1. Example of a complex query for gloss annotations on two different gloss tiers

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

Variable-based queries

There have always been three modes for query matching in ELAN and TROVA: substring matching, exact matching and matching based on regular expressions. Although especially the use of regular expressions can cover a wide range of search scenarios, it is not suited for every search task and it can also easily become too complex for many users. For that reason the 'variable based matching' has been developed that allows users to query by using a variable. One use case is that one wants to find every instance of the same exact word (annotation) co-occurring on two tiers ('Find any annotation on tier A that has the same value as an overlapping annotation on tier B, regardless of the actual value of the annotation'). This query schema can become more complex and powerful by combining multiple variables in the query.

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 2, presenting some of the results for a query similar to that in Figure 1.

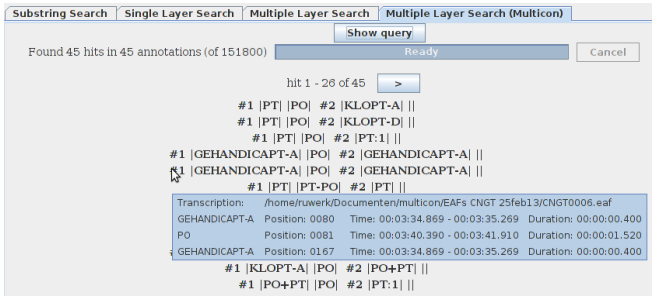

Figure 2. Traditional presentation of the results of a complex query

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 1. 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 1 and Figure 2. 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 3. 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.

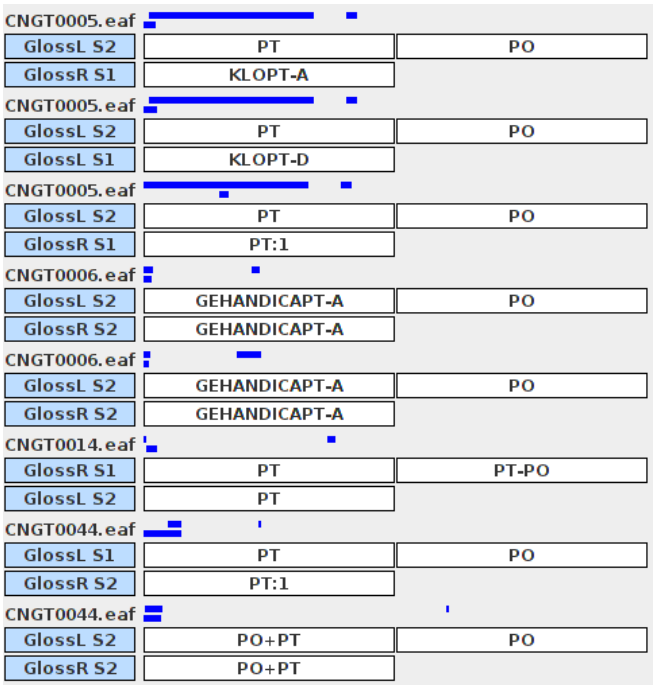

Figure 3. Visualisation of search hits by a matrix view that conforms to the query specification, including the visualization of alignment of annotations

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 4. 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.

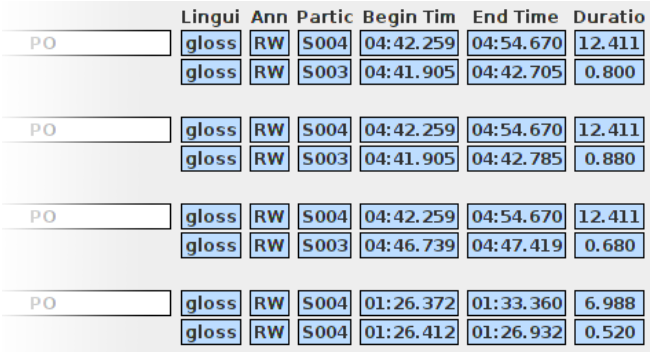

Figure 4. Display of various types of information on search results in columns, here displaying Linguistic Type, Annotator, Participant, Begin Time, End Time, and Duration

Saving and loading queries

Often a (complex) query has to be executed more than once, either over time, in the same set of files that are still being worked on, or in different sets of files, or from different computers. Being able to save and load a query is not only convenient, since the same constructs do not have to be entered over and over again, but it also reduces flawed results caused by errors in the query construction process. This option has now been added to both ELAN and TROVA. The query is now stored as a single string wrapped in an XML file, but a more explicit XML schema for storing this information is envisaged for future development.

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. For instance, one addition that would appear useful is the flexible filtering of search results on the basis of the metadata properties that are displayed in the columns. Two new development efforts that will take place in 2013-2014 with support of CLARIN-NL concern the creation and display of multilingual

annotations and metadata, and the collaboration on annotation files by users in different locations.

Acknowledgments

The project 'Multilayer Concordance Functions in ELAN and ANNEX' (MultiCon) was funded by CLARIN-NL, project number CLARIN-NL-11-018.

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*. http://dx.doi.org/10.1017/S0022226713000054. Published online by Cambridge University Press on 11 April 2013.

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.