Rapid processing of ELAN data: quick and dirty numbers for statistical analysis of non-manual features (For poster presentation)

We developed an algorithm that aids in the rapid processing of data stored in ELAN files. ELAN is a tool used by sign language researchers to annotate video data (Crasborn & Sloetjes, 2008). Navigating and searching within ELAN files is clumsy, which may make drawing inferences from the data complicated or time consuming. This problem is apparent when attempting to unearth the phonetic correlates of syntactic and prosodic functions.

As sign languages may convey linguistic information simultaneously on the hands, and on the face and body ('non-manuals' [NMs]), noticing or extracting patterns of simultaneity may be difficult. For instance, brow raise is widely reported in ASL as a marker of topics, y/n-questions, and other grammatical functions (Wilbur & Patschke, 1999). However, there may be secondary articulators that signal that one function is intended over another.

At present, the only reliable way to explore NM overlaps in ELAN is by counting by hand. However, this is undesirable for larger files and for multiple tiers. ELAN does have an overlap counter via its search capability ("Search > Structural Search Multiple eaf > Multiple Layer Search"). It is possible to return the number of overlaps between 2 tiers. However, when searching for 3+ tiers, overlaps are always reported w.r.t. a single, base tier. For example, based off the hypothetical data in Fig. 1, if Tier 1 is taken as the base, a search for overlaps of 3 tiers will return annotations {b, e, g}, as desired. However, running the same search with Tier 4 as the base returns annotations {a, b, d, e, g}. That is, ELAN over-reports overlaps: while annotations {a, e} both overlap with annotation {g}, they don't overlap with each other. As such, one would have to run up to $n!$ searches (where n is the number of tiers) and remove undesired overlaps.

The solution offered by Benitez-Quiroz et al. (2014) is more powerful. They apply a computational model, whereby their algorithm scans ELAN files and learns which NMs reliably distinguish between members of a set of five sentence types. For example, the authors were also able to find that a leftward headturn is the most distinguishing non-manual of hypothetical sentences, which had not previously been reported for ASL.

However, Benitez-Quiroz et al.'s solution is limited in certain ways. The names of the NM tiers, manual tier(s), and sentence type tier are stitched into the program, meaning that their code looks for a particular suite of tier names. This limits the number and nature of research questions explorable through their software. Second, although the authors provide their code, they don't provide a minimal working example. We tried to replicate their findings on our own data, and were unsuccessful. Further, the authors warn that certain analyses may take days to complete.

```
Tier 1   —a—            —b—   —c—
Tier 2         —d—
Tier 3          ——e——         —f—
Tier 4    ————————g————————
```

Figure 1: A hypothetical ELAN window

Who's counting	E(yes)/G(loss)	E/ P(olarity)	G/ P	E/P/G	P/E/G
(a) Handcount	51	52	133	45	45
(b) ELAN	55	52	138	**158**	**60**
(c) Our program	52	51	136	49	49

Figure 2: Overlap tallies of 2 and 3 tiers of a sample ELAN project, (a) by hand, (b) using ELAN's 'Structured search multiple eaf' function, and (c) using our algorithm.

By contrast, our code allows for more exploratory research: it can return the number of overlaps for any number of specified or unspecified tiers. As such, our code will work with ELAN files that have different architectures from the one assumed by the Benitez-Quiroz group. One further advantage is that our code can be executed in a web browser, so data can be processed quickly and on the fly. The trade-off, though, is that our code does not itself find meaningful overlaps and does not provide detailed statistics for what it finds yet. (The raw numbers can be plugged into the researcher's preferred statistical software).

At present, our algorithm is flexible enough to provide the following information:

Scenario 1: Given tiers of interest, our code reports every combination of overlap. For instance, in the hypothetical ELAN file (Fig. 1), our code can tell us that Tier 1 and 4 overlap twice; Tiers 1, 3, and 4 overlap once; and that Tiers 1 and 2 don't overlap at all. Further, the values of the overlaps are returned. So, while Tiers 1, 3, and 4 *do* overlap generally, the specific annotations c and f don't overlap with Tier 4. We imagine that researchers looking to find general patterns would benefit here. For instance, if Tier 4 represents brow position, annotation g represents brow lowering, we might predict that annotation b is a wh-word, while annotation c is not.

Scenario 2: Our code can also run more specific searches for hypothesis testing. A researcher could search specifically for a word and return (a) all of the tiers that overlap, (b) the specific annotation values that overlap, and (c) the frequencies and durations of those overlaps. Points (a,b) allow researchers to see what tiers and annotations overlap with a word like WHO (e.g., lowered brows, head-tilt, etc.). Point (c) allows for the exclusion of rare overlaps and those that do not meaningfully overlap (i.e., potential flukes; perhaps there was only 1 head tilt occuring with WHO across the whole dataset).

We validated our code by hand-tallying overlaps in ELAN for small set of tiers. We recorded cases of 1, 2, and 3 overlaps (Fig. 2a) and compared them to ELAN's search function and our code's results. ELAN produces comparable counts for any combination of 2 tiers tested, as do we. However, ELAN over-reports overlaps (for reasons mentioned above) when 3 tiers are compared. What's more, depending on what order the tiers are compared in, different figures result. Our algorithm produces consistent, comparable results.

References

Benitez-Quiroz, C., Gökgöz, K., Wilbur, R., & Martinez, A. (2014). Discriminant features and temporal structure of nonmanuals in ASL. *PloS one*, *9*(2), e86268.

Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In *Proceedings of LREC 2008, 6th intl. conference on lang. resources and evaluation.*

Wilbur, R., & Patschke, C. (1999). Synt. correlates of brow raise in ASL. *SLL*, *2*(1).