A bioinformatics solution to inter-rater agreement for forced time-alignment of data from underresourced languages

Matthew Stave (DDL, Lyon, matthew.stave@cnrs.fr) ; François Delafontaine (DDL, Lyon, francois.delafontaine@cnrs.fr)
Ludger Paschen (ZAS, Berlin, paschen@leibniz-zas.de) ; Frank Seifart (ZAS, Berlin, frank.seifart@berlin.de)

ANR-DFG DoReCo (2019-2022)

A collection of 50+ languages from documentations of small and often endangered languages.

- Morphological annotation (for 30+ languages).
- Phonemic alignment of all languages with MAUS.

Evaluation of the MAUS alignment

- 6 transcriptions (from 3 languages) for 4125 words
- Automatically aligned with MAUS (Kisler et al. 2017)
- Manually corrected by 4 annotators

Languages

Anal (Ozerov 2018) – 1341 word sample

- Sino-Tibetan language of north-east India

Resígaro (Seifart 2009) – 657 word sample

- Arawakan language of Peru

Vera’a (Schnell 2015) – 2127 word sample

- Austronesian language of Vanuatu

Calculating inter-rater agreement

for annotations with different boundaries and different segmentation.

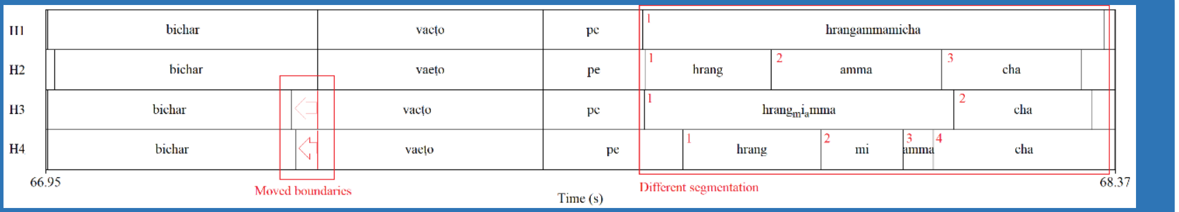

Figure 1: Correction of word alignment in Anal from four annotators (H1-H4)

Without knowing which unit corresponds to which, comparison is difficult. Current methods (Cohen 1968; Krippendorf 2004) don’t match units between sequences:

- Atomization – segmenting further into an equal number of constant segments
- EasyDIAg – relying on overlap and categorization (Holle & Rein 2015)
- Staccato – using overlap to determine « nuclei » (Lücking et al. 2012)

One method does (Gamma, Mathet et al. 2015) with matching and agreement measurements in a single process; but it isn’t easy to access

The Needleman-Wunsch algorithm

An algorithm originally developed for matching DNA sequences (Needleman & Wunsch 1970).

- Makes two sequences correspond by finding the alignments with the maximal number of matching units from all possible alignments

Implemented using the “pairwise2” function from Python’s “Bio” library”:

```
# pairwise2.align.globalms(annoA, annoB, 2, -1, -1, -1)
```

Results

What the implementation returns is pairs / matches of the same units from two different annotations.

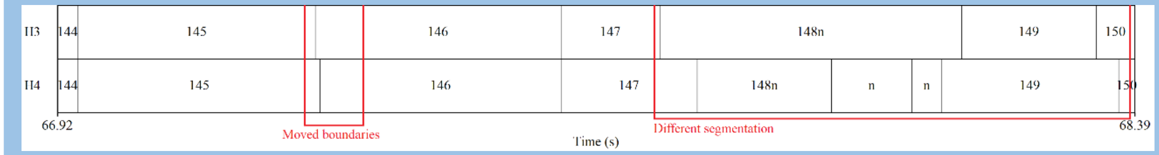

Figure 2: Matching units of annotators H3-H4 (from figure 1, lines 3-4)

- Numbers indicate matched units; “n” unmatched units; numbers with “n” matched units with edited content
- Evaluation of the matching script by inter-rater agreement with a manual correction of the automatic matching result

Table 1: Score between automatic and manual unit matching

Language	Kappa-score	Accuracy
Anal	0.98	98.25%
Resigaro	0.9885	99.05%
Vera’a	0.9727	97.6%
Total	*0.98*	*98%*

- Most disagreement is with pauses,
- The algorithm treats them like any other unit.
- Better results by removing / weighing?

Measuring the segmentation

How to compare unit boundaries? Available measures for MAUS-human alignments (totals).

General unit matching:

85,08% matched; 6,80% edited; 8,12% unmatched

General overlap:

91,88% of matched units perfectly overlap

Average distance of moved boundaries:

111ms for onsets; 133ms for offsets

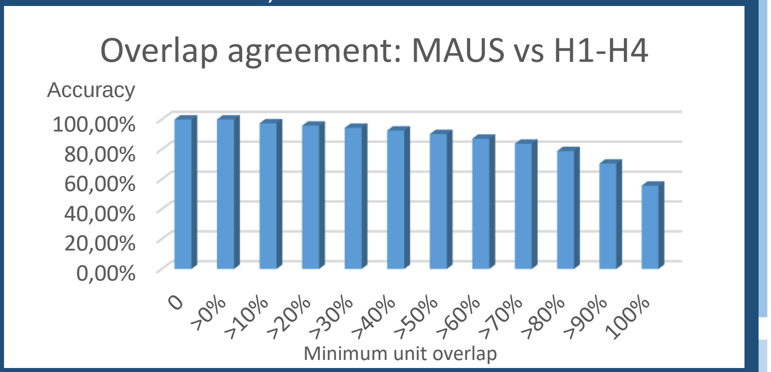

Fig.3: Proportions of units for a given ratio of overlap

Other uses and perspectives

This algorithm can be used to align any sequences of behavioral data with different boundaries or mismatched segmentation (such as annotations of speech or gesture), to prepare sequences for measurement of inter-annotation agreement. It is also useful for aligning sequences that are coded at different levels, such as morphemes and words or words and utterances.

References: **Cohen, J.** (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4), 213-220. **Holle, H. & Rein, R.** (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior research methods* 47(3), 837-847. **Lücking A., Ptock S., Bergmann K.** (2012) Assessing Agreement on Segmentations by Means of *Staccato*, the *Segmentation Agreement Calculator according to Thomann*. In: Efthimiou E., Kouroupetroglou G., Fotinea S. E. (eds) Gesture and Sign Language in Human-Computer Interaction and Embodied Communication. GW 2011. Lecture Notes in Computer Science, vol 7206. Springer, Berlin, Heidelberg. **Mathet, Y., Widlöcher, A. & Métivier, J.-P.** (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational linguistics* 41(3), 437-479. **Needleman, S.B. & Wunsch, C.D.** (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48(3), 443-453. **Ozerov, P.** (2018). *A community-driven documentation of natural discourse in Anal, an endangered Tibeto-Burman language*. London: SOAS, Endangered Languages Archive. <elar.soas.ac.uk/Collection/MPI1035093>**. Kisler, T., Reichel, U.D. & Schiel, F.** (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326-347. **Krippendorff, K.** (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks: Sage. **Schnell**, S.(2015). Multi-CAST Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts.* <multicast.aspra.uni-bamberg.de/#veraa>. Seifart, F. (2009). Resígaro corpus. Nijmegen: TLA. <hdl.handle.net/1839/00-0000-0000-0008-38F4-0>.