Analysing translational survey data through multialignment

Michael Cysouw Jürg Fleischer

November 22, 2017

Abstract

In the analysis of (dialectal) survey data there is a long path of interpretation from the data that is collected to the eventual interpretation. In most past (and current) research there is no paper-trail of all the large and small decisions being taken in the processing of the data. This paper describes as series of methods to document the processing of translational survey data, i.e. data that consists of translational equivalents. As an example we will process 2500 translations of a single sentence from the original Wenker data, transliterated from the original questionnaire from the 19th century. In addition to the well-known geographic distribution of sounds, we will show that it is just as well possible to extract syntactic and lexical variables from this data.

1 Multialigning translations

There are many ways in which comparable data can be collected to compare different language variants. Possibly the most traditional kind of comparable data (and also often critcized, REF?) are translational equivalent utterances. Language consultants are simply asked to produce the closest possible translation of a given utterance in their language. It is this kind of data that will be the focus of the paper, though the techniques proposed have a much wider application. We will analyse Wenker sentence 9: '*Ich bin bei der Frau gewesen und ich habe es ihr gesagt, und sie sagte, sie wolle es auch ihrer Tochter sagen.*' (I have been at the women and I have told it to her, and she said that she would tell it to her daughter). We have transliterated about 2500 translations from the original Wenker questionnaire, extended with translations from Austria, Switzerland, the Netherlands, Belgium and various german-speaking linguistic enclaves. This data was used at the start of the 20th century to produce the infamous dialectmaps.

2 Formats for multialignment

Although one can argue that historical linguistics is from its inception a kind of multialignment, the first explicitly written-down multialignments arose in the context of DNA sequence analysis in biology (cf. Figure 1). In such files each element to be aligned (here amino-acids) are abbreviated with a single character, and the alignment is specified by fixing the position of the characters in a line. Typically, the first ten characters are reserved for a name, and afterwards the characters are just put directly after each other. When an element is missing for one of the species (as so-called 'gap'), then a dash '-' is inserted to keep the sequences aligned.

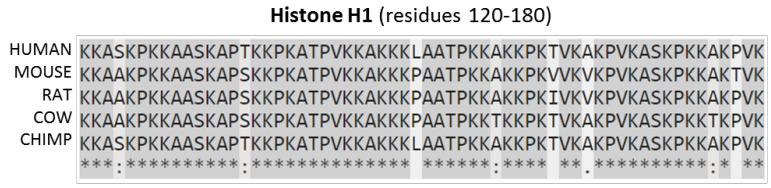

Figure 1: Example of a multiple sequence alignment of amino-acids across biological species.[2]

This approach can rather straighforwardly be adapted for the multiple alignment of sounds in comparative linguistics. An early example can be seen in Figure 2, aligning various words for 'day'. However, one problem with linguistic data is that the assumption of 'one element = one letter' is difficult to maintain without strongly simplifying the data. For example, in phonetic transcriptions a single element often consists of multiple Unicode characters, like /aː/ or /t͡ʃ/. In the benchmark data collected by List & Prokić (2014)[3] they opted to put explicit tab marks between the columns, as illustrated by an example from their database in Figure 3.

References

Bhargava, Aditya & Grzegorz Kondrak. 2009. Multiple word alignment with profile hidden markov models. In *Proceedings of human language technologies (NAACL-HLT 2009)*, 43–48. Boulder, Colorado: Association for Computational Linguistics.

[2] Example from `https://en.wikipedia.org/wiki/Sequence_alignment` by T. Shafee.
[3] available online at `http://alignments.lingpy.org`

```
D--E--N-        D--E--NY
Z--E--N-        DZ-E--N-
DZIE--N-        D--A--N-
DI-E--NA        D--E--IZ
D--I--A-        D--Y--DD
D--I--E-        Z-----U-
Z--U--E-        Z-----I-
J--O--UR        D-----I-
DJ-O--U-        G--IORNO
```

Figure 2: Linguistic multialignment from Bhargava & Kondrak (2009: 47).

```
Germanic
"all"
American English.....................    -    ɑː   ɬ
Canadian English.....................    -    ɒː   ɬ
Central German (Cologne).............    ʔ    a    l
Central German (Honigberg)...........    ʔ    ɒ    ɬ
Central German (Luxembourg)..........    ʔ    a    l
Central German (Murrhardt)...........    ʔ    a    l
Danish...............................    ʔ    ɛ    lˀ
Dutch (Antwerp)......................    -    a    ɬ
Belgian Dutch........................    -    ɑ    ɬ
Dutch (Limburg)......................    -    a    l
Dutch (Ostend).......................    -    ɑo   -
Dutch................................    -    ɑ    ɬ
```

Figure 3: Linguistic multialignment with aligned multigraphs and using tabs as separators (List & Prokić 2014)

List, Johann-Mattis & Jelena Prokić. 2014. A benchmark database of phonetic alignments in historical linguistics and dialectology. In *LREC*, 288–294.