Analysing translational survey data through multialignment

Michael Cysouw Jürg Fleischer

December 1, 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

2.1 Multialignment of sounds in words

Although one can argue that historical linguistics is from its inception based on 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.

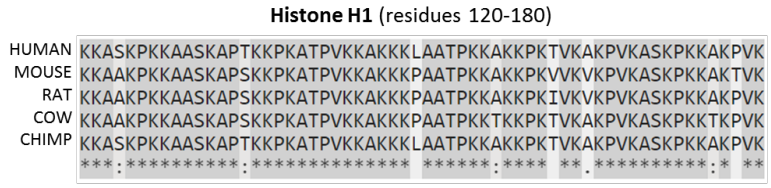

Figure 1: Example of a multiple sequence alignment of amino-acids across biological species.[2]

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) they opted to put explicit tab marks between the columns, as illustrated by an example from their database in Figure 3.[3] A recent development is to allow more columns with different kind of additional information in such files (e.g. data references or expert status). A multialignment then should be just a single column in such a file. Therefor, the column separator for multialignment has become a single space. This is found for example in the data underlying the article by Hill & List (2017), an excerpt of which is shown in Table 1.[4] This approach is codified

[2] Example from `https://en.wikipedia.org/wiki/Sequence_alignment` by T. Shafee.

[3] The full data is available online at `http://alignments.lingpy.org`

[4] The full data is available at `https://zenodo.org/record/886179`.

in the recent *Cross-Linguistic Data Formats* (CLDF) standard.[5]

```
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).............    ?    a     l
Central German (Honigberg)...........    ?    ɒ     ɬ
Central German (Luxembourg)..........    ?    a     l
Central German (Murrhardt)...........    ?    a     l
Danish...............................    ?    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).

There are two beta-stage software projects that try to develop a user interface to make it easier to manually edit linguistic multialignment: the Edictor and the MSA-Editor.[6] For the current example we have used the MSA-Editor to prepare the data.

[5] The CLDF specification and documentation is available online at `http://cldf.clld.org`.

[6] Both project have a common origin. The Edictor is available at `https://github.com/digling/edictor`, described in List (2017). The MSA-Editor is available at `https://github.com/cysouw/msa-editor`.

Word_ID	Form	Cognate_set_ID	Alignment	Cognate_source
4097	a^{31}tsi^{31}	501	a ^{31} + ts i - - ^{31}	Hill2017
2	pz̨a^{55}pz̨a^{55}	33	p z̨ a ^{55} + p z̨ a ^{55}	Hill2017
5462	pjã^{22}	518	p j ã ~ ^{22}	Hill2017
9	pan^{31}ʃɵʔ^{55}	567	p a n ^{31} + ʃ ɵ ʔ ^{55}	Hill2017
10	pəŋ^{31}ʃɔʔ^{55}	567	p ə ŋ ^{31} + ʃ ɔ ʔ ^{55}	Hill2017

Table 1: Excerpt of data from Hill & List (2017)

2.2 Multialignment of words in sentences

The concept of alignment between languages is also an important notion in the field of machine translation. Parallel translations between languages (or 'bitexts' as they are known in computational linguistics, see Tiedemann 2011) are the basic starting point for any machine translation, as this is the principal route to learn equivalent expressions between languages. The classical approach to learn equivalent structure between languages is to align bitexts, i.e. assign a linking between some parts of one sentence ('words') to parts of the translated sentence. The most influential software to perform such bitext alignment has been Giza++ (Och & Ney 2003). It used a specific output format for aligned sentences, which has become known as the 'Pharaoh' format.[7] The Pharaoh-format counts the words, starting at zero, and then links these indices by a dash. For example, the bitext alignment shown in Figure 4 will be summarised in Pharaoh-format as '0-3 1-1 2-1 4-2'.

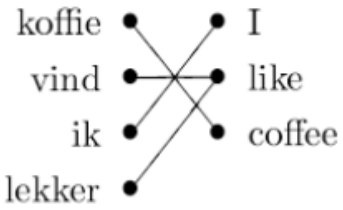

Figure 4: Bitext alignment (Tiedemann 2011: 4).

The obvious reason for this format is the crossing lines that occur regularly in word alignment. In sound alignment this also occurs (viz. in metathesis

[7] Many names in the statistical machine translation have some kind of reference to Egypt, which seems to be a reference to one of the early highly influential machine translation toolkits, called 'EGYPT' http://web.archive.org/web/20060826032742/http://www.clsp.jhu.edu:80/ws99/projects/mt/.

as shown in Figure 5), but it is not extremely common. For that reason, it is possible to use the 'trick' shown in Figure 5, namely to use two columns for the representation of one alignment.

Language	**Alignment**							
Russian	s	ɔ	-	-	n	ts	ə	
Polish	s	-	w	ɔ	n^j	ts	ɛ	
French	s	ɔ	l	-	-	-	-	ɛj
Italian	s	o	l	-	-	-		e
German	s	ɔ	-	-	-	-		nə
Swedish	s	uː	l	-	-	-	-	

Figure 5: Example of multialignment of metathesis (List 2014).

3 Workflow

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.

Hill, Nathan W & Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: a case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting* 3(1). 47–76.

List, Johann-Mattis. 2014. *Sequence comparison in historical linguistics* (Dissertations in Language and Cognition). Düsseldorf: Düsseldorf University Press. http://dup.oa.hhu.de/244/.

List, Johann-Mattis. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the EACL 2017 software demonstrations*, 9–12.

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

Och, Franz Josef & Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1). 19–51.

Tiedemann, Jörg. 2011. *Bitext alignment*. Morgan & Claypool.