Phonetic Alignment Based on Sound Classes

A New Method for Sequence Comparison in Historical Linguistics

Johann-Mattis List (July 23, 2010)

Heinrich Heine University Düsseldorf

Abstract. In this paper, I present a new method for the automatic implementation of pairwise and multiple alignment analyses in historical linguistics which is based on sound classes and implemented as a Python library. While sound classes are usually employed in historical linguistics as a stochastic device for detecting possible sound correspondences among languages and the proof of genetic relationship among languages, it shall be shown that they are equally well apt for phonetic alignment tasks. Moreover, they have two further advantages: Firstly, due to the fact that sound classes constitute a rather small alphabet, they are perfectly apt for subsequent use in biological software tools for sequence alignment, which makes it possible to carry out quick pairwise and multiple alignment analyses. Secondly, since sound classes can be based on explicit historical considerations regarding phonetic similarity, the alignments are capable of yielding certain outputs which cannot be retrieved by applying similarity metrics which are solely based on synchronic phonetic resemblences.

1 Sequence Comparison and Alignment Analyses in Historical Linguistics

Among the different aspects of comparative-historical linguistics, sequence comparison plays a crucial role. It constitutes the basis of the comparative method which seeks to detect regular sound correspondences in lexical material of different languages in order to prove their genetic relationship and to uncover the unattested ancestor language by means of linguistic reconstruction [1]. Since sequences – in contrast to sets – consist of non-unique elements which retrieve their distinctive function only because of their order, sequence comparison is always based on phonetic alignment, i.e. the corresponding phonetic segments of two or more sequences are ordered in such a way that they are set against each other.

In the following, I shall present a new method for phonetic alignment, which is not only easy to implement and to modify but also explicitly historically oriented. The paper is structured as follows: After giving a short introduction into the basic algorithms which are usually employed when carrying out pairwise and multiple sequence alignments, I shall describe the method by presenting the basic idea behind the sound classes employed and their implementation in the Python library. In a further step, I shall discuss the performance of the method in contrast to an alternative proposal by G. Kondrak [2].

2 Basic Procedures for Alignment Analyses

2.1 Pairwise Alignments

Almost all procedures for alignment analyses in historical linguistics which have been proposed so far are based on the Dynamic Programing Algorithm (DPA) which was independently proposed by different scholars as a way to carry out sequence alignments in such different disciplines as biology, linguistics, and gas chromatography [3][1].

The basic idea of the DPA is to create a matrix which confronts all segments of the sequences under comparison either with each other or with alternative null-sequences (fills). In a further step, the algorithm seeks the path through the matrix which is of the lowest general cost. The general cost is cumulatively calculated by means of a specific scoring function that penalizes the matching of segments with each other and the insertion or deletion of segments in the two sequences. Figure 1 illustrates this process for the alignment of Engl. "heart" vs. Germ. "herz" based on the Levenshtein scoring function which penalizes both fills and mismatches with 1. The left matrix shows the shortest alignment path chosen by the algorithm, the right matrix reflects the cumulatively calculated costs for each cell (for a more detailed description of the algorithm, cf. [6]).

Fig. 1. DPA-Matrix for Engl. "heart" vs. Germ. "herz"

Regarding alignment analyses in historical linguistics, three aspects of the DPA are of crucial importance: (1) the basic part of the algorithm which iterates over the matrix, (2) the scoring function for segment-to-segment comparison, and (3) the pre-segmentation of the linguistic units, i.e. the transcription of the phonetic characters, carried out when compiling the datasets for sequence comparison. All these aspects have a crucial impact on the quality of the alignments created by the procedure.

While earlier approaches mainly concentrated on the scoring function [7][2] or additional edit operations [2] the method proposed here divides sounds into specific classes whose members show a high probability of interchange during language evolution.

---

[1] As examples for different proposals regarding the DPA, cf. e.g. [4] and [5]

2.2 Multiple Alignments

While pairwise alignment analyses can be carried out without problems using the above-mentioned dynamic programming algorithm or certain of its extensions [8][9][10], multiple sequence alignments (MSA) have to make use of certain heuristics which do not guarantee that the optimal alignment for a set of sequences has been found, since the computational effort increases enormously with the number of sequences being analysed [6, 345]. The most common heuristics which is applied in computational biology are the so-called progressive algorithms which are based on a guide-tree that is reconstructed from the pairwise alignment scores of all sequences and along which the sequences are stepwise added to the multiple alignment [11, 143f] (see Figure 2).

Fig. 2. MSA Based on a Guide-Tree
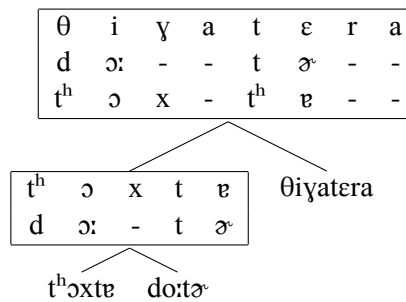
While the original approach by Feng & Doolittle[12] for progressive MSA compares sequences only pairwise, thus taking a pair of sequences as representative for a whole multiple alignment, profile-based approaches allow for a more refined approach to align multiple sequence alignments to each other [11, 146f]. A profile consists of the relative frequency of all segments of a multiple alignment in all its positions [6, 337] (see Figure 3). Thus, a profile represents a multiple alignment as a sequence of vectors. Aligning profiles to profiles instead of aligning two represantative sequences of two given multiple alignments usually yields better results in MSA, since more information can be taken into account which would otherwise be ignored.

3 Sound Classes in Historical Linguistics

The main idea behind sound classes in historical linguistics is the assumption that it is possible "to divide sounds into such groups, that changes within the boundary of the groups are more probable than transitions from one group into another" [13, 272][2]. Thus, when comparing the dental consonants t, d, tʰ, θ with the velars k, g, kʰ, ɣ one can assume that

---

[2] My translation, original text: «[...] выделить такие группы звуков, что изменения в пределах группы более вероятны, чем переводы из одной группы в другую.»

Fig. 3. MSA and Profiles

Multiple Alignment: Traditional Format						
ʧ	-	l	o	vʲ	ε	k
ʧ	-	-	o	v	ε	k
ʧʲ	ɪ	l	ɐ	vʲ	ε	k
ʧ	-	w	ɔ	vʲ	ε	k
Multiple Alignment: Profile Representation						
ʧ	.75					
ʧʲ	.25					
l			.50			
o				.50		
v					.25	
vʲ					.75	
ɐ				.25		
ε						1.0
ɪ		.25				
k						1.0
w			.25			
ɔ				.25		
-		.75	.25			

it is more probable that any of the dentals may change to a dental than to a velar sound, and vice-versa. This does, of course, not mean that a sound change from one class into another is impossible, yet most linguists would certainly agree that such a sound change would be rather unexpected and strange. Starting from this general assumption, A. B. Dolgopolsky [14] was the first to carry out empirical studies of the most typical sound changes in a large sample of languages. He proposed ten fundamental sound classes, which are given in Table 1.

Table 1. Dolgopolsky's Sound Classes

No.	Class	Description	Example
1	P	labial obstruents	p,b,f
2	T	dental obstruents	d,t,θ,ð
3	S	sibilants	s,z,ʃ,ʒ
4	K	velar obstruents, dental and alveolar affricates	k,g,ts,ʧ
5	M	labial nasal	m
6	N	remaining nasals	n,ɲ,ŋ
7	R	liquids	r,l
8	W	voiced labial fricative and initial rounded vowels	v,u
9	J	palatal approximant	j
10	ø	laryngeals and initial velar nasal	h,ɦ,ŋ

Sound classes have been employed in a couple of recent studies which largely deal with stochastical aspects of the prove of genetic relationship among languages [15] [16] or as a heuristic for the automatical implementation of cognate judgments [17] [3]. In contrary to the approach presented here, these studies are not based on sequence alignment, but rather check whether the first or the first two consonants of basic words match regarding their respective sound classes.

4 The Python Library for Sound-Class-Based Alignment

4.1 General Working Procedure

The method for sound-class-based alignment has been implemented as a Python library and can be invoked from the Python prompt or within Python scripts[4]. The core function of the library, the alignment function, executes the following operations: After tokenizing the input sequences (which should be in IPA-transcription), it first converts the input sequences into strings of capitals which represent the 11 sound classes employed by the method. These strings are then passed to a function that carries out an alignment analysis of the class-strings. The aligned strings are then converted back to their original IPA-transcription (see Figure 4).

The sound classes employed in the library are mainly based on the suggestions of Dolgopolsky [14], but they are extended to cover the full range of IPA, including the most common diacritics, and vowels (simple vowels and diphtongs), which are ignored in Dolgopolsky's original system, are included as an eleventh class of sounds.

4.2 Pairwise Alignments

Pairwise alignments are implemented by the pairwise2-module of the BioPython library [19], which allows one to carry out both local and global alignment analyses. While global alignment analyses carry out alignments for two entire strings, local alignment analyses, which are based on an extension of the DPA (the Smith-Waterman algorithm [8]), seek the two substrings which show the highest similarity and eventually leave prefixes and postfixes unaligned [11, 22-24].

In order to enhance the alignment analysis, a special matching dictionary has been prepared as an input for the scoring function (see Table 2). Note that the scoring function of pairwise2 is based on similarity of segments as apposed to distance. Segments which should be matched by the algorithm are therefore given higher scores than segments whose matching should be avoided.

---

[3] As a matter of fact, nearly all alignment analyses in historical linguistics, such as the ones carried out by the ASJP project [18] or the approach proposed by G. Kondrak [2], are based on "sound classes" in a broader sense, since they usually abstract from a strict phonetic notion to a broader phonemic one. Yet, appart from the algorithm proposed by Covington [7], none of these approaches makes use of historical knowledge regarding the probability of sound change processes when carrying out the similarity judgments.

[4] A preliminary version of the modules, including two testsets, is online available under http://www-public.rz.uni-duesseldorf.de/ jorom002/sca.zip.

Fig. 4. The Alignment Analysis of the Sound-Class-Approach
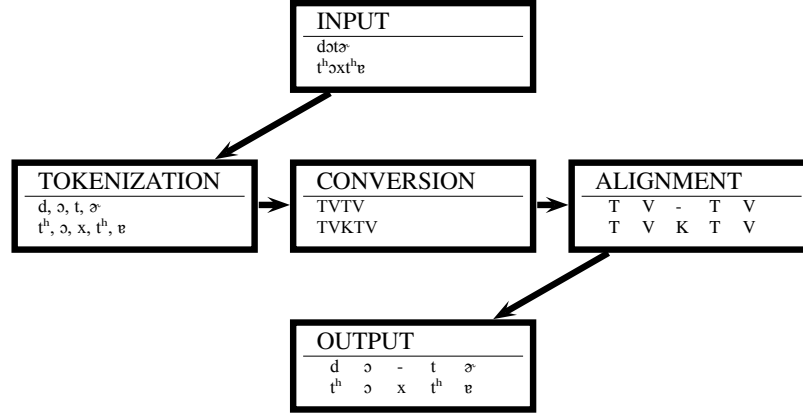
Table 2. Matching Dictionary for the Scoring Function of BioPython.pairwise2

Score	Condition	Example
5	Consonant-Class-Identity	K + K
4	Vowel-Class-Identity	V + V
-10	Vowel-Non-Vowel	V + K
-4	Non-Identity of Consonants	K + M
2	Specific Combinations	K + S, K + T, T + S
-1	Gaps	- + K

4.3 Multiple Alignments

In the current implementation of the library, both traditional MSA, based on the Feng-Doolittle-algorithm [12], as well as profile-based MSA roughly based on the CLUSTALW-implementation [20] for MSA in evolutionary biology is possible. The code for MSA analyses has been written by the author. The calculation of the guide-tree, which can be carried out either by the UPGMA clustering algorithm [21] or the Neighbor-joining method [22], makes use of the cogent.phylo-module of the PyCogent library [23]. The scoring function for profile-sequence and profile-profile alignments is based on the *sum of pairs score*, a standard way to score multiple alignments in evolutionary biology [11, 139f], which consists of the mean of the sums of all pairwise segment scores of two MSAs.

5 Performance of the Method

5.1 Pairwise Alignments

The method was tested (using local alignment) on a testset of 82 cognate pairs proposed by M. A. Covington [7], which was slightly modified in order to be appropriate for the input

Table 3. Comparison of the Sound-Class-Approach with ALINE's Alignments

Sound-Class-Approach	ALINE
1 Engl. daughter / Old Grk. θυγατήρ "daughter"	
d o - - t ə r / tʰ u g a t e: r · d - - o t ə r / tʰ u g a t e: r	d o t ə r / tʰu g a te: r
2 Spa. decir / Fre. dire "say"	
de θ i r / d i r	d e θ i r / d - - i r
3 Engl. this / Grm. dieses "this"	
ð i s / d i: z əs	ð i z / di: z ə s
4 Fox kiinwaawa / Menomini kenuaq "you (Pl.)"	
k i: n w a: w a / k e n - u - a ʔ	k i: n w a: wa / k e n u a ʔ
5 Old Grk. δίδωμι / Lat. do "I give"	
di d o: mi / d o: · d i do:mi / d o:	di do: mi / do:
6 Engl. tooth / Lat. dentis "tooth"	
t u - θ / d e n t is	t u θ / den t i s
7 Engl. I / Lat. ego "I"	
ai / e go · ai / eg o	ai / e go
8 Engl. eye / Grm. Auge "eye"	
ai / au gə · ai / aug ə	a i / a u ge
9 Spa. todos / Fre. tous "all"	
t o dos / t u · to d o s / t u	t o dos / t u
10 Engl. one / Lat. unus "one"	
w ə n / u: n us	wə n / u: n us
11 Engl. round / Lat. rotundus "round"	
r - - au n d / r o t u n d us · r au - - n d / r o t u n d us	r a - u n d / r o t u n d us

format required for the sound class approach[5]. The results of the alignment analyses were compared to the ALINE algorithm of G. Kondrak [2] which shows the best performance of recently proposed alignment algorithms for linguistic purposes. Comparing the output of the sound-class approach to ALINE's alignments for the Covington testset, there are 71 cases, where both methods yield exactly the same results. Of the 11 ones which are aligned differently (see Table 3) there are six cases where the sound-class approach gives

[5] The non-IPA-characters of Covington's testset were converted to IPA-symbols and the halfvowels of the diphtongs, which were originally coded as glides were converted to the respective full vowels.

two equivalent outputs. In four out of these six cases, one of these double-outputs matches with ALINE's single-output (Nos. 5, 7, 8, and 9 in Table 3), in one case both outputs produced by the sound-class-approach are superior to ALINE's output (No. 1) and in the last one (No. 11), it cannot be decided, which of the outputs given by either of the approaches is better. In the remaining 5 cases of different output, there are three cases where ALINE performs better (Nos. 2, 4, and 10) and two cases where the sound-class-approach gives the better alignment (Nos. 3, and 6).

It becomes obvious that, apart from similar results in the majority of the cases, there are a couple of significant differences between the alignments of the sound-class approach and ALINE's alignments. Firstly, there are cases, where the sound-class approach yields multiple outputs while ALINE has a single one. These results are mostly due to the information loss accompanying the conversion of IPA-strings into sound-class-strings. This, however, does not constitute a general problem for the method, since the double-outputs occur mostly in cases which are problematic for alignments in general: No historical linguist would dare to align words such as Engl. 'I' and Lat. 'ego', lacking the relevant facts from other related languages. Secondly, there are cases which show a particular benefit of the sound-class approach: Since this approach is not based on a synchronic idea of phonetic similarity, but on a 'historical' notion of phonetic similarity, it yields convincing outputs in such challenging cases as Engl. 'daughter' vs. Old Grk. 'thugatēr', where it matches all consonants correctly, while ALINE opposes *d* and *g*. Table 4 gives some representative examples for cognate pairs (transcribed with more phonetic detail) where the alignments of the sound-class-approach are superior to those of ALINE.

Table 4. Additional Examples for Different Alignments of the Sound-Class-Approach and ALINE

Sound-Class-Approach	ALINE
1 Old Grk. καρδιά / Skr. hr̥t "heart"	
k a r d | ia / h - r t	ka | r d | ia / h | r d
2 Grm. Herz / Lat. cor "heart"	
h ɛɐ | tsʰ / k o | r	hɛɐ | tsʰ / k | or
3 Mod. Grk. νέος / Rus. новый "new"	
n e - o s / n o v ɨ j	n e o s / n - o v | ɨj
4 Mod. Grk. καρδιά/ Grm. Herz "heart"	
k a r ð | ia / h ɛɐ - tsʰ	kar | ð i a / h ɛ ɐ | tsʰ

5.2 Multiple Alignments

Apart from the fact that the sound-class-approach is easy to implement and to modify, while yielding satisfying results comparable to that of more refined algorithms for sequence comparison, a major advantage of the approach lies in its flexibility to be adapted

for more complex approaches. Thus, the implementation of more complex algorithms is far less complicated, since many functions available in biological software modules for python can be easily included. This makes it even possible to carry out multiple alignment analyses which are rarely implemented in the current algorithms for sequence comparison in historical linguistics, the only exception known to the author being a proposal by Prokič et al. (2009) [24].

Since MSA has only recently been added to the library, no full-size tests runs can be reported at the moment, yet the first tests on small samples of dialectal data and cognate sets of Indo-European languages yield quite promising results. Furthermore, it can be easily demonstrated that the profile-based approach yields better results than other progressive approaches, as the comparison on profile-based MSAs and MSAs based on the traditional Feng-Doolittle algorithm in Table 5 for Indo-European and Slavic cognate pairs demonstrates, where Old Church Slavonic дъщи "daughter" and Polish człowiek "human" are incorrectly aligned in the non-profile-based approach.

Table 5. Multiple Alignments Yielded by the Sound-Class Approach

No.	Traditional MSA based on the Feng-Doolittle algorithm						
1	Old Grk. θυγατήρ / Grm. Tochter / Engl. daughter / OCS дъщи / Skr. duhitār "daughter"						
	tʰ	u	g	a	t	eː	r
	t	o	x	-	t	ə	r
	d	o	-	-	t	ə	r
	d	u	-	ʃ	t	i	-
	d	u	h	i	t	aː	r
2	Czech člověk / Bulgarian чoвeк / Russian чeлoвeк / Polish człowiek "human"						
	tʃ	-	l	o	vʲ	ɛ	k
	tʃ	-	-	o	v	ɛ	k
	tʃʲ	ɪ	l	ɐ	vʲ	ɛ	k
	tʃ	w	-	ɔ	vʲ	ɛ	k
No.	Profile-based MSA						
1	Old Grk. θυγατήρ / Grm. Tochter / Engl. daughter / OCS дъщи / Skr. duhitār "daughter"						
	tʰ	u	g	a	t	eː	r
	t	o	x	-	t	ə	r
	d	o	-	-	t	ə	r
	d	u	ʃ	-	t	i	-
	d	u	h	i	t	aː	r
2	Czech člověk / Bulgarian чoвeк / Russian чeлoвeк / Polish człowiek "human"						
	tʃ	-	l	o	vʲ	ɛ	k
	tʃ	-	-	o	v	ɛ	k
	tʃʲ	ɪ	l	ɐ	vʲ	ɛ	k
	tʃ	-	w	ɔ	vʲ	ɛ	k

6 Conclusion

In this paper, I presented a new approach for pairwise and multiple sequence alignments in historical linguistics. Although the method is quite simple regarding its basic assumptions and its implementation as a Python library, the performance of the approach is not only comparable to that of previously proposed ones, but it even shows a better performance in very challenging alignment tasks, the reason being its explicit historical orientation regarding phonetic similarity.

References

1. Lass, R.: Historical linguistics and language change. Cambridge University Press, Cambridge (1997)
2. Kondrak, G.: Algorithms for language reconstruction. PhD thesis, University of Toronto, Toronto (2002)
3. Kruskal, J.B.: An overview of sequence comparison: Time warps, string edits, and macromolecules. SIAM Review **25**(2) (April 1983) 201–237
4. Wagner, R.A., Fischer, M.J.: The string-to-string correction problem. J. ACM **21**(1) (1974) 168–173
5. Needleman, S.B., Wunsch, C.D.: A gene method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology **48** (July 1970) 443–453
6. Gusfield, D.: Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge (1997)
7. Covington, M.A.: An algorithm to align words for historical comparison. Computational Linguistics **22**(4) (1996) 481–496
8. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology **1** (1981) 195–197
9. Gotoh, O.: An improved algorithm for matching biological sequences. Journal of Molecular Biology **162**(3) (1982) 705 – 708
10. Oommen, B.J.: String alignment with substitution, insertion, deletion, squashing, and expansion operations. Inf. Sci. Inf. Comput. Sci. **83**(1-2) (1995) 89–107
11. Durbin, R.: Biological sequence analysis. Probabilistic models of proteins and nucleic acids. 7th print edn. Cambridge University Press, Cambridge (2002)
12. Feng, D.F., Doolittle, R.F.: Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of Molecular Evolution **25**(4) (1987) 351–360
13. Burlak, S.A., Starostin, S.A.: Sravnitel'no-istoričeskoe jazykoznanie (Comparative-historical linguistics). Akademia, Moskva (2005)
14. Dolgopolsky, A.B.: A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. In Shevoroshkin, Vitaly V.; Markey, T.L., ed.: Typology Relationship and Time. Notes on Linguistics. Karoma Publisher, Inc. (1986) 27–50 Originally published in Russian as "Gipoteza drevnejščego rodstva jazykov Severnoj Evrazii (problemy fonetičeskich sootvetstvij)" in 1964.
15. Baxter, W.H., Manaster Ramer, A.: Beyond lumping and splitting: Probabilistic issues in historical linguistics. In Renfrew, C., McMahon, A., Trask, L., eds.: Time depth in historical linguistics. Papers in the prehistory of languages. The McDonald Institute for Archaeological Research, Cambridge (2000) 167–188
16. Mortarino, C.: An improved statistical test for historical linguistics. Statistical Methods and Applications **18**(2) (2009) 193–204

17. Turchin, P., Peiros, I., Gell-Mann, M.: Analyzing genetic connections between languages by matching consonant classes. Journal of Language Relationship **1** (2010) 117–126

18. Brown, C.H., Holman, E.W., Wichmann, S., Velupillai, V.: Automated classification of the world's languages: a description of the method and preliminary results. STUF, Berlin **61**(4) (2008) 285–308

19. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.: Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11) (Jun 2009) 1422–1423

20. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**(22) (November 1994) 4673–4680

21. Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin **28** (1958) 1409–1438

22. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**(4) (1987) 406–425

23. Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J.G., Easton, B., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., Lozupone, C., McDonald, D., Robeson, M., Sammut, R., Smit, S., Wakefield, M., Widmann, J., Wikman, S., Wilson, S., Ying, H., Huttley, G.: Pycogent: a toolkit for making sense from sequence. Genome Biology **8**(8) (2007) R171

24. Prokić, J., Wieling, M., Nerbonne, J.: Multiple sequence alignments in linguistics. In: LaTeCH-SHELT&R '09: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R 2009), Morristown, NJ, USA, Association for Computational Linguistics (2009) 18–25