Phonetic Alignment Based on Sound Classes

A New Method for Sequence Comparison in Historical Linguistics

Johann-Mattis List *
Institute for Romance Languages and Literature
Heinrich Heine University Düsseldorf

ESSLLI 2010: Students' Session

In this talk, 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 Introduction

1.1 Sequence Comparison in Historical Linguistics

- Basic of the comparative method
- Basic of the detection of regular sound correspondences
- Basic of the proof of genetic relationship
- Basic of genetic language classification

1.2 Alignment Analyses in Historical Linguistics

- Sequences – in contrast to sets – consist of non-unique elements which retrieve distinctive function only because of their order.
- In alignment analyses, the corresponding elements of two or more sequences are ordered in such a way that they are set against each other.
- Sequence comparison in historical linguistics is always based on phonetic alignment.

2 Basic Procedures for Automatic Alignment Analyses

2.1 The Dynamic Programming Algorithm

- Create a matrix which confronts all segments of the sequences under comparison, either with each other, or with alternative null-sequences (fills).
- Seek the path through the matrix which is of the lowest general costs.
- Calculate the costs cumulatively by means of a specific scoring function that penalizes the matching of segments with each other and likewise the insertion and deletion of segments in any of the sequences[1] .

*The research leading to this talk was carried out in the research project "Evolution and Classification in Biology, Linguistics and the History of Science (EvoClass: http://www.evoclass.de)" funded by the German Federal Ministry of Education and Research (BMBF). I would like to thank my biological colleagues Tal Dagan, Shiju Lal NS and Ovidiu Popa for helpful discussions, and my supervisor Hans Geisler for inspiring critic and advice. If You have any further questions, feel free to contact me under: listm@phil-fak.uni-duesseldorf.de

[1] Cf. Wagner & Fischer (1974), Needleman & Wunsch (1970), Gotoh (1982), Oommen (1995), Smith & Waterman (1981)

2.2 Multiple Sequence Alignment

Guide-Tree Heuristics:

- Due to computational restrictions, multiple sequence alignment (MSA) is based on heuristics (Gusfield 1997:345).
- Heuristics based on guide-trees are the most common ones used in computational biology (Durbin 2002:143).
- Based on pairwise alignment scores, a guide-tree is reconstructed, and the sequences are stepwise added to the MSA along it (Feng & Doolittle 1987).

Profiles:

- The guide-tree heuristic can be enhanced by the application of profiles (Durbin 2002:146f).
- A profile consists of the relative frequency of all segments of a MSA in all its positions, thus, a profile represents a MSA as a sequence of vectors (Gusfield 1997:337).
- Aligning profiles to profiles instead of aligning two representative sequences of two given MSA yields better results, since more information can be taken into account.

3 Sound Classes in Historical Linguistics

3.1 Two Perspectives on Similarity in Linguistis

Synchronic Similarity: Sounds in different languages are judged to be similar, if they show resemblences regarding the way they are produced or perceived.

Diachronic Similarity: Sounds in different languages are judged to be similar, if they go back to a common ancestor.

3.2 The Conception of Sound Classes

Key Assumption of the Sound Class Approach:

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" (Burlak & Starostin 2005, 272, my translation).

A Diachronic Definition of Similarity:

Similarity is not based on synchronic resemblances of sounds but on on class-membership: two sounds, how dissimilar they may be from a synchronic perspective, may still belong to the same class.

Sound Classes proposed by Dolgopolsky (1986)

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

4 The Python Library for Sound-Class-Based Alignments

4.1 General Working Principle

- Input IPA-transcribed strings which shall be aligned.
- Tokenize the input sequences.

- Convert the tokenized sequences into strings of capitals representing the sound classes.
- Perform alignment analyses on the sound class sequences.
- Convert the aligned strings back to their original IPA transcription[2].

4.2 Pairwise and Multiple Alignments

Pairwise Alignments

- Based on pairwise2 of BioPython (Cock *et al.* 2009)
- Scoring functions adapted for Dolgopolsky sound classes
- Performs Global and local alignment analyses

Multiple Alignments

- MSA based on guide-trees (Feng & Doolittle 1987)
- MSA based on profiles (Thompson *et al.* 1994)
- Guide-trees calculated with PyCogent (Knight *et al.* 2007)
- Scoring function based on sum of pairs (Durbin 2002:139f)

5 Performance of the Method

5.1 Pairwise Alignments

Sound Classes vs. ALINE (Kondrak 2002) on Covington's (1996) Testset

Identical results:	71 / 82 cases
Double outputs where ALINE has one output:	6 cases
Double outputs matching ALINE's single output:	4 cases
Double outputs superior to ALINE:	1 case
Double outputs both fail:	1 case
ALINE superior to Sound Classes:	3 cases
Sound Classes superior to ALINE:	2 cases

Pairwise Alignments: Examples

	Sound-Class-Approach	ALINE
1	Engl. daughter / Old Grk. θυγατήρ "daughter"	
	d o - - t ə r t^h u g a t eː r	d o t ə r t^hu g a t eː r
2	Engl. this / Grm. dieses "this"	
	ð i s d iː z əs di:	ð i z z ə s
3	Engl. tooth / Lat. dentis "tooth"	
	t u - θ d e n t is den	t u θ t i s

[2] A preliminary version of the modules, including two testsets, will soon be online available under `http://www-public.rz.uni-duesseldorf.de/~jorom002/sca.zip`.

5.2 Multiple Alignments

Multiple Alignments: First Tests on Small Samples

Simple Guide-Tree-Based MSA						
ʧ	-	l	o	vʲ	ɛ	k
ʧ	-	-	o	v	ɛ	k
ʧʲ	ɪ	l	ɐ	vʲ	ɛ	k
ʧ	w	-	ɔ	vʲ	ɛ	k
Profile-based MSA						
ʧ	-	l	o	vʲ	ɛ	k
ʧ	-	-	o	v	ɛ	k
ʧʲ	ɪ	l	ɐ	vʲ	ɛ	k
ʧ	-	w	ɔ	vʲ	ɛ	k
Czech člověk / Bulgarian човек / Russian человек / Polish człowiek "human"						

Simple Guide-Tree-Based MSA						
tʰ	u	g	a	t	eː	r
t	o	x	-	t	ə	r
d	u	-	ʃ	t	i	-
d	u	h	i	t	aː	r
Profile-based MSA						
tʰ	u	g	a	t	eː	r
t	o	x	-	t	ə	r
d	u	ʃ	-	t	i	-
d	u	h	i	t	aː	r
Old Grk. θυγατήρ / Grm. Tochter / OCS дъщи / Skr. duhitār "daughter"						

References

Burlak, Svetlana Anatol'evna, & Sergej Anatol'evic Starostin. 2005. *Sravnitel'no-istoričeskoe jazykoznanie (Comparative-historical linguistics)*. Moskva: Akademia.

Cock, P J, T Antao, J T Chang, B A Chapman, C J Cox, A Dalke, I Friedberg, T Hamelryck, F Kauff, B Wilczynski, & M J de Hoon. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25.1422–1423.

Covington, Michael A. 1996. An algorithm to align words for historical comparison. *Computational Linguistics* 22.481–496.

Dolgopolsky, A. B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. In *Typology Relationship and Time*, ed. by T. L. Shevoroshkin, Vitaly V.; Markey, Notes on Linguistics, 27–50. Karoma Publisher, Inc. Originally published in Russian as "Gipoteza drevnejščego rodstva jazykov Severnoj Evrazii (problemy fonetičeskich sootvetstvij)" in 1964.

Durbin, Richard. 2002. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press, 7th print edition.

Feng, D. F., & R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* 25.351–360.

Gotoh, Osamu. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology* 162.705 – 708.

Gusfield, Dan. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge: Cambridge University Press.

Knight, Rob, Peter Maxwell, Amanda Birmingham, Jason Carnes, J Gregory Caporaso, Brett Easton, Michael Eaton, Micah Hamady, Helen Lindsay, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Michael Robeson, Raymond Sammut, Sandra Smit, Matthew Wakefield, Jeremy Widmann, Shandy Wikman, Stephanie Wilson, Hua Ying, & Gavin Huttley. 2007. Pycogent: a toolkit for making sense from sequence. *Genome Biology* 8.R171.

Kondrak, Grzegorz, 2002. *Algorithms for language reconstruction*. Toronto: University of Toronto dissertation.

Needleman, Saul B., & Christan D. Wunsch. 1970. A gene method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48.443–453.

Oommen, B. John. 1995. String alignment with substitution, insertion, deletion, squashing, and expansion operations. *Inf. Sci. Inf. Comput. Sci.* 83.89–107.

Smith, T. F., & M. S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 1.195–197.

Thompson, J. D., D. G. Higgins, & T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22.4673–4680.

Wagner, Robert A., & Michael J. Fischer. 1974. The string-to-string correction problem. *J. ACM* 21.168–173.