A Database of Phonetic Alignments in Historical Linguistics and Dialectology

Johann-Mattis List and Jelena Prokić
mattis.list@uni-marburg.de, prokic@uni-marburg.de

w	a	l	-	d	e	m	a	r	-
w	o	l	-	d	e	m	o	r	t
v	-	l	a	d	i	m	i	r	-
v	o	l	o	d	y	m	y	r	-

Phonetic Alignment Analyses

Alignment analyses are the most common way to represent differences between sequences. Since the formal aspect of the linguistic sign can be easily represented as a sequence of sounds, it is straightforward to use alignment analyses for the task of automatic sequence comparison in the historical disciplines of linguistics. The new methods for *phonetic alignment* do not only speed up the process, but also provide an explicit quantification of similarities and distances between words and morphemes.

	1	2	3	4	5	6	7	8	9	10
1	W	O	L	–	D	E	M	O	R	T
2	W	A	L	–	D	E	M	A	R	–
3	V	O	L	O	D	Y	M	Y	R	–
4	V	–	L	A	D	I	M	I	R	–

Figure 1: Alignment analysis of four sequences: Corresponding elements occur in the same column, while empty cells in the matrix, resulting from symbols which do not correspond with other symbols, are filled with a gap symbol.

Phonetic Alignment Modes

There are different modes for pairwise alignment analyses. *Global alignment* compares sequences as a whole. *Local alignment* compares the most similar subsequences. *Semi-global alignment* allows the stripping of pre- or postfixes in one of the sequences.

Global	–	C	A	T	E	R	I	N	G	
	S	K	A	T	E	R	–	–	–	
Local	C		A	T	E	R		ING		
	SK		A	T	E	R				
Semi-Global		C	A	T	E	R		ING		
	S	K	A	T	E	R				

Figure 2: Basic modes for pairwise alignment analyses.

Phonetic Alignment Formats

We represent phonetic alignments in text-files using very simple format prescriptions.

NR	<harry_potter.msa>
1	Harry Potter Testset
2	"WOLDEMORT"
3	English V O L - D E M O R T
4	German W A L - D E M A R -
5	Russian V - L A D I M I R -
6	SWAPS . + - + ...
7	LOCAL * * * . * * * * * .
8	# Alignments are charming ;-)

Figure 3: Alignment format for multiple alignments.

NR	FILE <harry_potter.psa>
1	Harry Potter Testset
2	"WOLDEMORT" (German, Russian)
3	German w a l - d e m a r
4	Russian v - l a d i m i r
5	
6	"WOLDEMORT" (English, Russian)
7	English w o l - d e m o r t
8	Russian v - l a d i m i r -
9	
10	"WOLDEMORT" (English, German)
11	English w o l d e m o r t
12	German w a l d e m a r -

Figure 4: Alignment format for pairwise alignments.

A Benchmark Database of Phonetic Alignments

The *Benchmark Database for Phonetic Alignments* (BDPA, http://alignments.lingpy.org) is a publicly available benchmark database of manually edited phonetic alignments, designed as a platform to test and improve the performance of automatic alignment algorithms. The BDPA is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License. The data was collected from various publicly available sources which cover both language families of different time depths (Germanic, Slavic, Romance, etc.), as well as dialects of single language varieties (Norwegian, Bulgarian, Dutch, etc.). All data is given in IPA transcription, but the detail of the transcriptions may vary from source to source.

Subset	Description	Alignments	Words	Varieties	Diversity (PID)	Sources
Andean	Andean language varieties (Aymara, Quechua)	76	883	20	55	Heggarty 2006
Bai	dialects of Bai (a Sino-Tibetan language)	90	1416	17	32	Wang 2006, BDS
Bulgarian	Bulgarian dialects	152	32418	197	48	Prokić et al. 2009
Dutch	Dutch dialects	50	3024	62	44	MAND
French	dialect varieties of Swiss French	76	3797	62	41	Gauchat et al. 1925
Germanic	Germanic languages and dialects	111	4775	45	32	Renfrew and Heggarty 2009
Japanese	Japanese dialects	26	224	10	40	Shirō 1973
Norwegian	Norwegian dialects	51	2183	51	46	Almberg and Skarbø 2011
Ob-Ugrian	Uralic languages	48	689	21	45	GLD
Romance	Romance languages	30	240	8	37	Renfrew and Heggarty 2009
Sinitic	Chinese dialects	20	346	40	35	Hóu (2004)
Slavic	Slavic languages	20	100	5	38	DERKSEN
TOTAL	–	750	50095	538	42	–

Table 1: Major sources, major subsets, and basic statistics of the BDPA.

The BDPA Webinterface

The BDPA website (http://alignments.lingpy.org) offers all data for download. With help of the BDPA web interface, the data can be browsed in different ways.

Browse the data:

Subset (Family)	Slavic
Concept	
Percentage Identity	Less than 40 and more than 0
Variety (Language)	
Metathesis:	☐ SUBMIT

A: Searching for alignments

Found 10 files matching your query:

ID	FILE	Subset	Label	PID	HTML	MSA
655	phonalign_655	Slavic	Proto-Slavic *pepelə	27	HTML	MSA
657	phonalign_657	Slavic	Proto-Slavic *piti	28	HTML	MSA
658	phonalign_658	Slavic	Proto-Slavic *ognɪ	25	HTML	MSA
659	phonalign_659	Slavic	Proto-Slavic *azə	28	HTML	MSA
661	phonalign_661	Slavic	Proto-Slavic *edinə	30	HTML	MSA
662	phonalign_662	Slavic	Proto-Slavic *zelenə	40	HTML	MSA
663	phonalign_663	Slavic	Proto-Slavic *ajɪce	33	HTML	MSA
665	phonalign_665	Slavic	Proto-Slavic *sɐdeti	22	HTML	MSA
666	phonalign_666	Slavic	Proto-Slavic *zemlja	33	HTML	MSA
668	phonalign_668	Slavic	Proto-Slavic *jɪmẽ	26	HTML	MSA

B: Selecting alignments

Plot of file "phonalign_661.msa" [BACK]

File:	phonalign_661.msa	Number of Words (all):	5
Dataset:	Slavic [?]	Number of Words (unique):	4
Label:	Proto-Slavic *edinə	Percentage Identity:	55

Sort alphabetic | Show all sequences

Variety	Alignment			
Russian	-	ɐ	dʲ	i	n
Bulgarian	-	ɛ	d	i	n
Serbian	j	e	d	a	n
Czech	j	ɛ	d	ɛ	n

C: Inspecting alignments (HTML plot)

Source of file "phonalign_661.msa": [BACK]

```
Slavic
Proto-Slavic **edinə*
Russian.. - ɐ dʲ i n
Czech.... j ɛ d ɛ n
Polish... j ɛ d ɛ n
Bulgarian - ɛ d i n
Serbian.. j e d a n
LOCAL.... * * * * *
```

D: Inspecting alignments (raw text)

Figure 5: The Webinterface of the BDPA.

References

Allen, B. (2007). *Bai Dialect Survey*. SIL International. PDF: http://www.sil.org/silesr/2007/silesr2007-012.pdf.

Almberg, J. and K. Skarbø (2011). *Nordavinden og sola. En norsk dialektprøvedatabase på nettet [The North Wind and the Sun. A Norwegian dialect database on the web]* [The North Wind and the Sun. A Norwegian dialect database on the web]. Recordings and transcriptons by J. Almberg. Technical implementation by K. Skarbø. URL: http://www.ling.hf.ntnu.no/nos/.

Derksen, R. (2008). *Etymological dictionary of the Slavic inherited lexicon*. Leiden and Boston: Brill.

Gauchat, L., J. Jeanjaquet, and E. Tappolet (1925). *Tableaux phonétiques des patois suisses romands*. Neuchâtel: Attinger.

Heggarty, P. (2006). *Sounds of the Andean languages*. URL: http://www.quechua.org.uk/.

Prokić, J., J. Nerbonne, V. Zhobov, P. Osenova, K. Simov, T. Zastrow, and E. Hinrichs (2009). "The computational analysis of Bulgarian dialect pronunciation". *Serdica Journal of Computing* 3.3, 269–298.

Renfrew, C. and P. Heggarty (2009). *Languages and origins in Europe*. URL: http://www.languagesandpeoples.com/.

Schutter, G. de, B. van den Berg, T. Goeman, and T. de Jong, eds. (2007). *MAND. Morfologische Atlas van de Nederlandse Dialecten [Morphological atlas of Dutch dialects]. Morfologische Atlas van de Nederlandse Dialecten*. URL: http://www.meertens.knaw.nl/mand/database/. Amsterdam: Meertens Instituut. URL: http://www.meertens.knaw.nl/mand/database/.

Shirō, H. (1973). "Japanese dialects". In: *Diachronic, areal and typological linguistics*. Ed. by H. M. Hoenigswald and R. H. Langacre. The Hague and Paris: Mouton, 368–400.

Starostin, G. and P. Krylov (2011). *The Global Lexicostatistical Database. Compiling, clarifying, connecting basic vocabulary around the world: From free-form to tree-form*. URL: http://starling.rinet.ru/new100/main.htm.

Wang, F. (2006). *Comparison of languages in contact. The distillation method and the case of Bai*. Taipei: Institute of Linguistics Academia Sinica.

Acknowledgements: This work was supported by the ERC starting grant 240816 and the German Federal Ministry of Education and Research. We are thankful to H. Geisler, M. Dickmanns, S. M. Oetzel, and K. Vogt, V. Persien, and Wáng Fēng sharing data and providing technical help.