A Benchmark Database of Phonetic Alignments in Historical Linguistics and Dialectology

Johann-Mattis List[1], **Jelena Prokić**[1]
[1]Research Center Deutscher Sprachatlas, Philipps-Universität Marburg,
mattis.list@uni-marburg.de, prokic@uni-marburg.de

Abstract

In the last two decades, alignment analyses have become an important technique in quantitative historical linguistics and dialectology. Phonetic alignment plays a crucial role in the identification of regular sound correspondences and deeper genealogical relations between and within languages and language families. Surprisingly, up to today, there are no easily accessible benchmark data sets for phonetic alignment analyses. Here we present a publicly available database of manually edited phonetic alignments which can serve as a platform for testing and improving the performance of automatic alignment algorithms. The database consists of a great variety of alignments drawn from a large number of different sources. The data is arranged in a such way that typical problems encountered in phonetic alignment analyses (metathesis, diversity of phonetic sequences) are represented and can be directly tested.

Keywords: phonetic alignment, benchmark, computational historical linguistics

1. Phonetic Alignment Analyses

In the past two decades, quantitative approaches have been repeatedly applied in historical linguistics and dialectology. They are dealing with problems of deep genetic relations between languages (Holman et al., 2011), but also with the classification of dialects and individual language systems (Nerbonne et al., 1996). Quantitative research in historical linguistics and dialectology relies on algorithms which quantify the similarities and distances between speech sounds, sound sequences, and – consequently – between words and entire languages (Kondrak, 2000; Heeringa, 2004; Holman et al., 2011). While in historical linguistics and dialectology the identification of regularly corresponding sounds is traditionally carried out manually, more recent approaches make use of *alignment analyses*, employing algorithms originally designed for applications in computer science and bioinformatics.

Alignment analyses are the most common way to represent differences between sequences. In an alignment analysis, two or more sequences are arranged in a matrix in such a way that corresponding segments occur in the same column. Segments which do not have corresponding segments in the other sequences are represented with help of *gap symbols* (usually a dash `"-"`, see the example in Figure 1). 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 quantitfication of similarities and distances between words and morphemes.

Sequence alignment techniques are regularly used for different tasks in computational linguistics, such as spell checking (Oflazer, 1996), information retrieval (Lambert, 1997), or plagiarism detection (Horton et al., 2010). Since the 1970s, automatic alignment methods have been successfully applied in bioinformatics in order to compare DNA or protein sequences (Durbin et al., 2002). Sequence alignment algorithms are commonly divided into those which align two sequences (*pairwise sequence alignment*, PSA), and those which align more than two sequences at the same time (*multiple sequence alignment*, MSA, Durbin, 2002). Inspired by the great progress made in bioinformatics in the past three decades, both PSA and MSA techniques have now also made their way into historical linguistics and dialectology (Kondrak, 2000; Prokić et al., 2009b; List, 2012a).

	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.

When dealing with *pairwise phonetic alignment*, different *alignment modes* can be distinguished (List, 2012b, 35f). The most important ones are *global*, *local*, and *semi-global* alignment analyses. *Global alignment analyses* compare two sequences as a whole. *Local alignment analyses* compare only the most similar parts of two sequences. *Semi-global alignment analyses* compare two sequences as a whole, but they allow the stripping of prefixes or postfixes in one of the sequences. Figure 2 illustrates the differences between the three alignment modes by comparing global, local, and semi-global alignment analyses of the sequences `"CATERING"` and `"SKATER"`.

While multiple alignment analyses are very common in bioinformatics, their application in historical linguistics and dialectology is still in its infancy. The main problem of multiple alignment analyses is their complexity. While algorithms for pairwise alignment analyses are guaranteed to find an optimal solution, exhaustive search is not feasible in multiple alignment analyses. Therefore, different heuristic strategies, such as iterative procedures (Prokić et al., 2009b), profile hidden Markov models (Bhargava and Kondrak, 2009), or libraries of pairwise alignments (List, 2012b) need to be employed. The advantage of multiple alignment analyses, on the other hand, is that they can take much more information into account, thus allowing researchers to look into much more fine-graded variation patterns. Figure 3 gives an example for multiple alignment analyses in linguistics and shows how 20 Chinese dialects words for 'tomato' can be aligned (original data taken from Hóu, 2004).

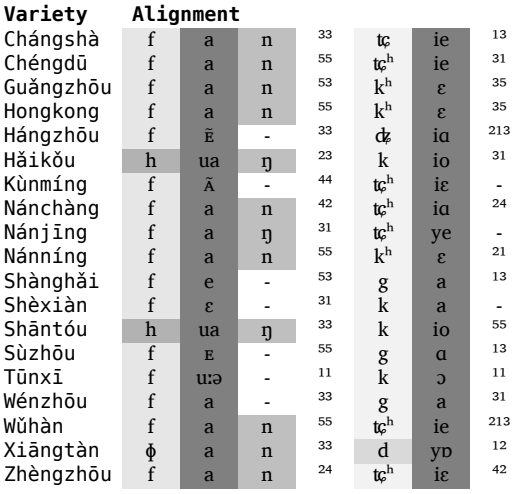

Figure 3: Multiple alignment of Chinese dialect words for 'tomato'. The coloring of the sound segments roughly reflects to which sound class (in the sense of Dolgopolsky, 1964) they belong.

2. BDPA: A Benchmark Database for Phonetic Alignments

Despite an increasing number of studies that rely on pairwise and multiple alignment analyses in historical linguistics and dialectology, a systematic evaluation of the performance of those algorithms has only rarely been carried out so far. In many studies the evaluation is done in an indirect way by looking at the number of correctly identified cognates or language families. Only in a few studies the performance of the algorithms is evaluated directly by comparing it to a manually corrected expert alignments and examining the percentage of correctly aligned sequences and the types of errors made by the algorithms (Prokić et al., 2009b; Wieling et al., 2009; List, 2012a; List, 2012b). The reason for this is that the number of publicly available benchmark datasets that allow the direct evaluation of phonetic alignment algorithms is rather limited.

In Covington (1996), a small dataset consisting of 82 word pairs was used to test a new pairwise phonetic alignment algorithm. Unfortunately, only the results of the algorithm were presented. The correct solutions were not provided. In later studies, the test set was nevertheless frequently used as a benchmark for pairwise alignment algorithms (Kondrak, 2000; Somers, 1999; Oakes, 2000). For multiple phonetic alignments, Prokić et al. (2009b) used manually aligned Bulgarian dialect data, consisting of 152 multiple alignments covering 197 dialect locations and more than 30 000 words, as a benchmark for testing the multiple alignment algorithm by Alonso et al. (2004). In List (2012b) a new benchmark dataset was presented. Including the data of Prokić et al. (2009b), it consisted of 600 multiple alignments covering six different language families, 435 different language varieties, and a total of 45 947 words. The extended benchmark database also included a pairwise partition which was directly extracted from the multiple alignments by taking the 5 506 most divergent, unique word pairs.

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: 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.

Here we present the *Benchmark Database for Phonetic Alignments* (BDPA, `http://alignments.lingpy.org`), 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 database is an extended version of the benchmark presented by List (2012b). A prototype was first presented and tested in the study of List (forthcoming). 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.

3. Structure of the BDPA

The BDPA contains benchmark data for pairwise and for multiple alignment analyses. For pairwise align-

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), Allen (2007)
Bulgarian	Bulgarian dialects	152	32418	197	48	Prokić et al. (2009a)
Dutch	Dutch dialects	50	3024	62	44	de Schutter et al. (2007)
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	Starostin and Krylov (2011)
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 (2008)
TOTAL	–	750	50095	538	42	–

Table 1: Major sources, major subsets, and basic statistics of the BDPA. **Diversity** refers to the diversity of the alignments measured in terms of *percentage identity* (see text).

ment analyses, three different benchmarks are currently offered, namely

a) an early benchmark dataset proposed by Covington (1996), consisting of 82 alignments in slightly adjusted phonetic transcription,

b) a master dataset of 7 197 alignments, and

c) a specific dataset consisting of 1 088 alignments which are exclusively taken from tone languages.

Covington's benchmark (a) is included in the BDPA for historical reasons, since it has been used by quite a few authors in the past. The master dataset (b) and the specific dataset (c) were automatically extracted from our masterset of multiple alignments by selecting the most diverse, unique sequence pairs.

The benchmark for multiple alignment analyses contains a total of 750 sets of manually corrected multiple alignments, covering 8 different language families, 538 different taxonomic units (language and dialect varieties), and a total of 50 095 word forms of which 14 185 are unique. The 750 multiple alignments can be further divided into 12 different *subsets*. With the exception of the data for the Bai dialects (a Sino-Tibetan language spoken in South-West China), which was taken from two different sources, all subsets were taken from one specific source in order to guarantee the homogeneity of the phonetic transcriptions. The size of the 12 subsets varies. The smallest data set is the data set for Slavic languages, consisting of only 20 multiple alignments covering 5 language varieties and 100 words. The biggest subset is the data on Bulgarian dialects, containing 152 multiple alignments consisting of phonetic transcriptions for 32 418 words collected at 197 different sites.

Table 1 lists the sources of the subsets along with further statistical information (number of multiple alignments, number of words, number of linguistic varieties, average diversity of the sequences). The column **Diversity** gives a rough estimate regarding the diversity of the subsets by listing the average *percentage identity* of the sequences in the multiple alignments.[1] As can be seen from the table, the subsets of the BDPA cover both highly diverse alignments, such as the subsets for Germanic and Bai with an average percentage identity of 32, and highly homogeneous ones, such as the subsets for Andean and Bulgarian, with average percentage identities of 55 and 48, respectively. The different degrees of phonetic diversity, which also reflect the different time depths of the subsets in the BDPA, allow for a great flexibility in testing and evaluating phonetic alignment algorithms.

4. Alignment formats in the BDPA

When dealing with pairwise and multiple alignment analyses, it is important to define clear-cut formats for the handling of alignments in order to guarantee that the data can be easily handled, maintained, and shared. For practical reasons, the BDPA uses the alignment formats of LingPy, a Python library for quantitative tasks in historical linguistics (`http://lingpy.org`, see also List and Moran, 2013). LingPy defines different formats for pairwise and multiple alignment analyses. All formats are text-based and can be edited with help of simple text editors.

4.1. Multiple Alignments: MSA Format

The basic format for the representation of multiple alignment analyses is the MSA format. Files in this format have the extension `"msa"`. Table 2 illustrates the structure of the format. The first line of an MSA file serves as an identifier for the dataset from which the alignment was taken. There are no further format restrictions and the user can freely decide what to use as an identifier, as long as it does not exceed the first line. In the BDPA, we use the names of our subsets (see the first column of Table 1) as dataset identifiers. The second line is reserved as an identifier for the set of aligned sound sequences. The identifier can again be freely chosen by the user. In the BDPA, we generally use the meaning of the sound sequences as identifier,

[1] Percentage Identity is a standard measure for the diversity of a given multiple alignment in bioinformatics. There are different ways to measure the percentage identity. In the BDPA, we use the most common approach which divides the number of identical positions in an alignment by the sum of the number of aligned positions and the number of internal gap positions, as described in Raghava and Barton (2006).

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 ;-)

Table 2: The MSA format for the representation of multiple alignment analyses. Line 1 is reserved as an identifier for the dataset. Line 2 serves as an identifier for the cognate set. Lines 3, 4, and 5 give the phonetic sequences in aligned form, separated by a tab-stop, and preceded by the language identifiers. Specific headwords can be used to indicate further characteristics of the MSA, such as metathesis in line 6, or highly consistent sites in line 7. Comments are preceded by a hash (#) symbol, and ignored when parsing the file, as show in line 8.

but we also add additional information, such as the ancestral form (in language families) or the orthography of the corresponding word in the standard variety (in dialect datasets). The following lines give the phonetic sequences in aligned form, separated by a tab-stop, and preceded by language identifiers (ISO-code, language name, or dialect location) in the first column of the alignment matrix. The hash symbol (`"#"`) is used as a comment character. When placed in the beginning of a line, it indicates that the line should be ignored when parsing the file. Line 8 in Table 2 gives an example for the use of the comment characters.
Inspired by alignment formats in bioinformatics, such as the Stockholm format used in the Pfam database for protein families (Finn et al., 2008), LingPy allows for specific additional lines which can be used to annotate the alignments. Instances of metathesis, for example, may be represented by adding a line which starts with the keyword `"SWAPS"`, with a plus character (`"+"`) marking the beginning of a swapped region, the dash character (`"-"`) its center and another plus character the end. All sites which are not affected by swaps contain a dot (`"."`, see line 6 in Table 2). In the BDPA, 66 out of 750 multiple alignments contain instances of metathesis and are regularly annotated in the way just described. Highly consistent sites of a multiple alignment (local peaks) can be annotated by adding an extra line which starts with the keyword `"LOCAL"`. Consistent columns (with a low amount of gaps) are marked with an asterisk (`"*"`). All other columns are marked with a dot (`"."`) as shown in line 7 in Table 2.

4.2. Pairwise Alignments: PSA Format

Generally, the MSA format can also be used to represent pairwise alignment analyses. However, since each MSA file, is a single text file, we would need 7 197 different text files to represent all sequence pairs of our master benchmark for pairwise alignment analyses. Using such a large amount of text files to represent the rather small amount of information available in pairwise alignments is not only impractical as a shared digital resource, but also very inefficient for computation.
In order to deal with large amounts of pairwise alignments in one and the same text file, LingPy offers an additional format for pairwise alignment analyses. This format is called PSA format, and files in the format have the extension `"psa"`. Table 3 gives an example for the PSA format. As for the MSA format, the first line of a PSA file is reserved for an identifier that refers to the dataset from which the data was taken. The sequence pairs themselves are given in triplets, with a sequence identifier in the first line of a triplet (containing the meaning, or orthographical information, as shown in lines 2, 6, and 10 in Table 3) and the two sequences in the second and third line (lines 3/4, 7/8, and 11/12 in the table) contain the alignment matrix with the language identifiers being placed in the first column. All triplets (sequence pair identifier and two sequences) are separated by one empty line as illustrated in lines 5 and 9 in Table 3. In the BDPA, the pairwise benchmarks, as described above, are provided in PSA format.

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 -

Table 3: Example for the PSA format. Line 1 is reserved as a dataset identifier. The pairwise alignments are given in triples, with a sequence identifier in the first line, and the aligned sequence pairs in the following lines. Triples are separated by one empty line.

5. The BDPA Website

The BDPA website (`http://alignments.lingpy.org`) offers all data for download. The pairwise alignment benchmark is provided in PSA format and can be downloaded as a whole. The multiple alignment benchmark is provided in MSA format and can be downloaded either as a whole, or in separate parts for each of the 12 subsets (as shown in Table 1). Additionally, all pairwise alignments which can be extracted from a given multiple alignment are offered along with the MSA files in PSA format.
With help of the BDPA web interface, the data can be browsed in different ways. With help of the query

interface for multiple alignments, one can search for specific datasets, specific glosses ('concepts'), different ranges of alignment diversity, specific language varieties, or alignments containing metathesis. In the example in Figure 4A, all alignments from the Slavic dataset with a percentage identity lower than 40 are selected, and the 10 alignments for which this condition holds are shown in Figure 4B.

Once an alignment is selected, the user can decide between different 'views' of the data. The first view plots the alignment in HTML markup. Sound segments are colored differently, depending on their *sound classes*. In the BDPA, we use the simple sound-class schema of Dolgopolsky (1964) in which ten different consonant classes and one vowel class are distinguished. Additional meta data, such as the number of words, the number of unique words, and the average percentage identity of the alignment, are also displayed. Specific JavaScript functions allow the user to change the sorting order of the sound sequences (*alphabetic* sorting applied to the names of the language varieties or *phonetic* sorting applied to the sound sequences), or to hide all but one of a set of unique sequences. The last option has been selected in Figure 4C, where only four out of five reflexes of Proto-Slavic *edinə 'one' are displayed. Since the reflexes in Czech and Polish are identical (at least in their given phonetic transcription [jɛdɛn]), the Polish reflex was removed from the view. The second view, which is illustrated in Figure 4D, shows the alignment in plain MSA format, as outlined in the previous section.

6. Open Questions and Future Challenges

6.1. Coverage

The data in the BDPA comes from various publicly available sources. On the project website, all sources are described in detail. If available, links to the original data are given. In most cases, only a small subset of the original data was selected for inclusion in the BDPA. The main goal of the database in its current form was not to describe one language resource exhaustively, but rather to provide as many different alignments as possible in order to allow the user to grasp a fair amount of true linguistic variation in the world's languages. Although the BDPA is much larger than previous alignment benchmarks, we are still far away from this goal. We hope that we find time to expand the database in the future.

6.2. Alignment Accuracy

Although we are pretty confident that the alignments in BDPA are of a generally good quality, it is hard to tell how accurate our alignments really are. One reason for this lies in the nature of alignments themselves: Alignments are hypotheses about the genetic relatedness of sound segments in genetically related words. Carrying out alignment analyses requires a sound knowledge of the history of the languages being compared. In most cases our knowledge of language history is based on the application of the comparative method. The comparative method, however, is based on sequence comparison itself, and the general uncertainty we may have regarding proto-forms resulting from linguistic reconstruction, holds also for the uncertainty regarding alignment analyses.

But this is not the only form of uncertainty one encounters when dealing with phonetic alignment. A more important problem relates to the practical aspects of alignment analyses themselves. Manually aligning multiple words is a tedious business, and it requires not only a good intuition regarding sound change processes in general, but also expert knowledge on common sound change patterns in the language varieties under investigation. In shallow time depths, alignment analyses are usually uncontroversial, and inter-annotator agreement is often very high.[2] In deeper time depths, however, specific knowledge in the individual language varieties under investigation becomes more and more important, and disagreement among scholars, even specialists in the same field, may grow drastically. Due to the lack of resources, the alignments in the BDPA were carried out by the authors of this paper themselves. We are generally confident that the number of errors is low, but we cannot give a guarantee that all individual decisions in the data are correct, especially in cases where we lack specialist knowledge of the respective language families. In the future, we hope to convince more scholars to join the BDPA project by providing new or correcting current alignments.

7. Conclusion

Thompson (2009, 154f) lists four requirements for benchmark databases in biology: (1) relevance, (2) solvability, (3) accessibility, and (4) evolution. *Relevance* refers to the tests in the benchmark which should be 'representative of the problems that the system is reasonably expected to handle in a natural [...] setting' (Thompson, 2009, 154). *Solvability* refers to the tasks presented by the benchmark. They should not be 'too difficult for all or most tools' (ibid., 154f), in order to allow for comparisons between different algorithms and methods. *Accessibility* refers to both the easiness to obtain and to use the data. *Evolution* refers to the requirement that benchmarks change constantly in order to avoid that programs are being optimized with respect to the benchmark instead of the general task the benchmark was designed to represent. When designing the Benchmark Database for Phonetic Alignments we tried to address these requirements as closely as possible. In order to guarantee relevance, the number of datasets from different languages and language families was drastically increased, in comparison to early

[2] A test on 36 cognate sets covering 6 593 words taken from the 'Phonetischer Atlas Deutschlands' (not included in the BDPA, for a description of the data, see Nerbonne and Siedle, 2005) showed that the authors of this paper agree in 95.79% of all columns and in 99.68% of all rows.

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 4: Browsing the BDPA with help of the web interface.

benchmarks for pairwise multiple alignments. In order to guarantee solvability, only cognate sets with little morphological variation were included. In order to guarantee consistency and applicability, only words transcribed in IPA transcription. The last of the four requirements, evolution, cannot be addressed at the moment. Since – in contrast to biology – automatic phonetic alignment still plays a minor role in historical linguistics, it is not clear whether it will be possible to find the resources to change the current benchmark regularly. We hope, however, that our initial efforts will eventually encourage scholars from historical linguistics and dialectology to join the BDPA project by providing helpful critics, fresh data, and new ideas.

8. Acknowledgements

This work was supported by the ERC starting grant 240816 ('Quantitative modelling of historical-comparative linguistics') and the German Federal Ministry of Education and Research. We are thankful to H. Geisler for providing us with his digitized version of Gauchat et al. (1925), to M. Dickmanns, S. M. Oetzel, and K. Vogt for digitizing the data of Shirō (1973), to V. Persien, for checking all alignments for formal errors, and to Wang Feng for providing us with a digital version of his Bai dialect data.

9. References

Allen, B. (2007). *Bai Dialect Survey*. SIL International.

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

Alonso, L., Castellon, I., Escribano, J., Xavier, M., and Padro, L. (2004). Multiple sequence alignment for characterizing the linear structure of revision. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 403–406.

Bhargava, A. and Kondrak, G. (2009). Multiple word alignment with profile hidden Markov models. In *Proceedings of the NAACL Conference 2009*, pages 43–48.

Covington, M. A. (1996). An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496.

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

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

Dolgopolsky, A. B. (1964). Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]. *Voprosy Jazykoznanija*, 2:53–63.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchinson, G. (2002). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, 7 edition.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Res.*, 36(Database issue):D281–288.

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

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, Rijksuniversiteit Groningen, Groningen.

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

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., Sauppe, S., Jung, H., Bakker, D., Brown, P., Belyaev, O., Urban, M., Mailhammer, R., List, J.-M., and Egorov, D. (2011). Automated dating of the world's language families based on lexical similarity. *Curr. Anthropol.*, 52(6):841–875.

Horton, R., Olsen, M., and Roe, G. (2010). Something borrowed: Sequence alignment and the identification of similar passages in large text collections. *Digital Studies*, 2(1).

Hóu, J., editor. (2004). *Xiàndài Hànyǔ fāngyán yīnkù [Phonological database of Chinese dialects]*. Shànghǎi Jiàoyù, Shànghǎi.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings of the NAACL Conference 2000*, pages 288–295.

Lambert, B. L. (1997). Predicting look-alike and sound-alike medication errors. *American Journal of Health-System Pharmacy*, 54(10):1161–1171.

List, J.-M. and Moran, S. (2013). An open source toolkit for quantitative historical linguistics. In *Proceedings of the ACL 2013 System Demonstrations*, pages 13–18.

List, J.-M. (2012a). Multiple sequence alignment in historical linguistics. In Boone, E., Linke, K., and Schulpen, M., editors, *Proceedings of ConSOLE XIX*, pages 241–260.

List, J.-M. (2012b). SCA. phonetic alignment based on sound classes. In Slavkovik, M. and Lassiter, D., editors, *New directions in logic, language, and computation*, pages 32–51. Springer, Berlin and Heidelberg.

List, J.-M. (forthcoming). *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Nerbonne, J. and Siedle, C. (2005). Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik*, 72(2):129–147.

Nerbonne, J., Heeringa, W., van den Hout, E., van de Kooi, P., Otten, S., and van de Vis, W. (1996). Phonetic distance between Dutch dialects. In *Proceedings of the CLIN '95 meeting*, pages 185–202.

Oakes, M. P. (2000). Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *J. Quant. Linguist.*, 7(3):233–243.

Oflazer, K. (1996). Error-tolerant finite-state recognition with applications to morphological analysis and spelling. *Computational Linguistics*, 22(1):73–89.

Prokić, J., Nerbonne, J., Zhobov, V., Osenova, P., Simov, K., Zastrow, T., and Hinrichs, E. (2009a). The computational analysis of Bulgarian dialect pronunciation. *Serdica Journal of Computing*, 3(3):269–298.

Prokić, J., Wieling, M., and Nerbonne, J. (2009b). Multiple sequence alignments in linguistics. In *Proceedings of the LaTeCH-SHELT&R Workshop 2009*, pages 18–25.

Raghava, G. P. S. and Barton, G. J. (2006). Quantification of the variation in percentage identity for protein sequence alignments. *BMC Bioinformatics*, 7(415).

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

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

Somers, H. L. (1999). Aligning phonetic segments for children's articulation assessment. *Computational Linguistics*, 25(2):267–275.

Starostin, G. and Krylov, P. (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`.

Thompson, J. D. (2009). Constructing alignment benchmarks. In Rosenberg, M. S., editor, *Sequence alignment. Methods, models, concepts, and strategies*, pages 151–177. University of California Press, Berkeley, Los Angeles, and London.

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

Wieling, M., Prokić, J., and Nerbonne, J. (2009). Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the LaTeCH-SHELT&R Workshop 2009*, pages 26–34.