Final version by the author. This paper has been published as:

**List, J.-M. (2012): SCA: Phonetic alignment based on sound classes.
In Slavkovik, M. and Lassiter, D. (eds.):
New directions in logic, language, and computation.
Berlin and Heidelberg: Springer. 32-51.
URL: http://rd.springer.com/chapter/10.1007/978-3-642-31467-4_3**

SCA: Phonetic Alignment Based on Sound Classes

Johann-Mattis List

Heinrich Heine University Düsseldorf
listm@phil.uni-duesseldorf.de

Abstract. In this paper I present the most recent version of the SCA method for pairwise and multiple alignment analyses. In contrast to previously proposed alignment methods, SCA is based on a novel framework of sequence alignment which combines new approaches to sequence modeling in historical linguistics with recent developments in computational biology. In contrast to earlier versions of SCA [1, 2] the new version comes along with a couple of modifications that significantly improve the performance and the application range of the algorithm: A new sound class model was defined which works well on highly divergent sequences, the algorithm for pairwise alignment was modified to be sensitive to secondary sequence structures such as syllable boundaries, and an algorithm for the pre-processing of the data in multiple alignment analyses [3] was included to cope for the bias resulting from progressive alignment analyses. In order to test the method, a new gold standard for pairwise and multiple alignment analyses was created which consists of 45 947 sequences covering a total of 435 different taxa belonging to six different language families.

1 Introduction

During the last two decades there ∼ has been an increasing interest in automatic approaches to historical linguistics which is reflected in a large amount of literature on phylogenetic reconstruction [4, 5], statistical aspects of genetic relationship [6, 7], and automatic approaches to sequence comparison [8–10]. In this context phonetic alignment plays a crucial role since it constitutes the first step of the traditional comparative method which seeks to detect regular sound correspondences in the lexical material of the languages of the world in order to determine cognate words and to prove their genetic relationship.

The SCA (**S**ound-**C**lass-Based Phonetic **A**lignment) method for pairwise and multiple phonetic alignment, whose most recent version shall be presented in the following, differs from previously proposed alignment methods [11, 8, 9] in so far as it is based on a new framework of sequence modeling which closely mimics traditional manual approaches. SCA is implemented as part of the LingPy library[1], a suite of open source Python modules with C++ extensions for time-consuming

[1] Online available under http://lingulist.de/lingpy.

and memory-intensive operations which provides solutions for different quantitative tasks in historical linguistics, and can be invoked from the Python prompt or in Python scripts.

2 Sequence Comparison in Historical Linguistics

Comparison plays a crucial role in historical linguistics. It constitutes the basis of the *comparative method* which takes similarities in the lexical material of different languages as evidence to prove their genetic relationship and to uncover unattested ancestor languages by means of linguistic reconstruction [12]. The basis of the reconstruction of proto-languages within the framework of the comparative method is the identification of *cognates* within languages which are assumed to be genetically related. Cognates are words or morphemes which are descendants of a common ancestor word or morpheme [13, pp. 62f]. The identification of cognates is based on a recursive procedure which starts from a small sample of presumed cognate words taken from different language varieties. These words are then compared for *sound correspondences*, i.e. sound pairs which recurrently occur in similar positions of the presumed cognate words [13, pp. 336f]. Once a preliminary sample of such sound correspondences is identified, the initial samples of presumed cognate words and sound correspondences are repeatedly modified by adding and removing words or correspondence pairs from the samples depending on whether they are consistent with the rest of the data or not.

The specific strength of this procedure is its underlying similarity measure. The similarity between words is determined on the basis of *functionally corresponding phonetic segments* as opposed to similarity based on *surface resemblances*. Comparing, for example, English *token* [təʊkən] and German *Zeichen* [tsaɪçən] "sign", the comparative method may prove that these words are cognates, despite the fact that they do not sound quite similar, since their sound segments can be shown to correspond regularly regarding their distinctive function within other cognates of both languages.[2] In the literature, this notion of similarity has been called *genotypic* as opposed to a *phenotypic* notion of similarity [14, p. 130].

The main comparanda in historical linguistics are *phonetic sequences* (words, morphemes). Generally spoken, sequences are ordered collections of objects whose identity is a product of both their *order* and their *content*. The objects of sequences, the *segments*, receive distinctivity only due to their specific composition. The comparison of sequences requires the comparison of both the structure and the substance of the segments constituting a sequence. This comparison is usually carried out within the framework of *alignment analyses*. In alignment analyses two or more sequences are arranged in a matrix in such a way that all

[2] Compare, for example, English *weak* [ʋiːk] vs. German *weich* [vaɪç] "soft" for the correspondence of [k] with [ç], and English *tongue* [tʌŋ] vs. German *Zunge* [tsʊŋə] "tongue" for the correspondence of [t] with [ts].

corresponding segments appear in the same column, while empty cells of the matrix, resulting from non-corresponding segments, are filled with gap symbols [15, p. 216]. Basically, an alignment analysis consists of two steps. In the first step, corresponding segments are identified, and in the second step, gap symbols (usually a dash -) are introduced as placeholders for non-corresponding segments, as illustrated in Fig. 1 for the sequences German *Tochter* [tɔxtər] "daughter" and English *daughter* [dɔːtər].

Although the term "alignment" has never been explicitly used in historical linguistics, it is obvious that the identification of sound correspondences inevitably relies on sequence alignment, since functionally corresponding sound segments could otherwise not be identified. The traditional analysis in the framework of the comparative method, however, is mostly restricted to a qualitative comparison that leaves it to the researcher to decide which segments are matched and which are not.

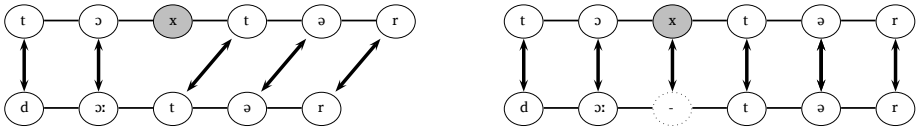

Fig. 1. Alignment analysis of German *Tochter* and English *daughter*

3 Automatic Alignment Analyses

The algorithmic basis of automatic alignment analyses was developed quite early. In the begin of the seventies, independent scholarly teams proposed the first algorithms for pairwise sequence alignment [16, 17]. Although up to today – especially in such disciplines as computational biology – many modifications and refinements to the basic algorithm have been proposed, and new improved methods are being constantly developed, it was only recently that historical linguists became interested in automatizing their traditional manual methods.

3.1 The Basic Algorithm for Pairwise Sequence Alignment

The basic algorithm for pairwise sequence alignment (PSA) belongs to the family of *dynamic programming algorithms* (DPA) [18]. The main idea of dynamic programming is to find an approach for the solution of complicated problems 'that essentially works the problem backwards' [19, p. 4]. Thus, instead of calculating all possible alignments between two sequences, the DPA for pairwise sequence alignment '[builds] up an optimal alignment using previous solutions for optimal alignments of smaller subsequences' [20, p. 19]. This is done by creating a matrix which confronts all segments of the sequences under comparison either

with each other or with alternative null-sequences (gaps). In a further step, the algorithm recursively calculates the total scores for the optimal alignment of all subsequences by filling the matrix from top to bottom and from left to right. Once the score for one subsequence has been determined, the score for a larger subsequence can also be calculated. In each step of the recursion, a specific *scoring function* evaluates whether the segments in the respective cell of the matrix should be matched with themselves or with one of the gap characters. Once the matrix is filled, the value in the last cell of the matrix yields the general score of the alignment of the sequences. The alignment is then obtained by applying a *traceback function* which finds the 'path of choices [...] which led to this final value' [20, p. 19].

The alignment of the strings `"HEART"` and `"HERZ"` is illustrated in Fig. 2. ① shows the completed alignment matrix for a scoring function which penalizes mismatches and gaps with -1 and matches with 1. ② illustrates the traceback procedure.

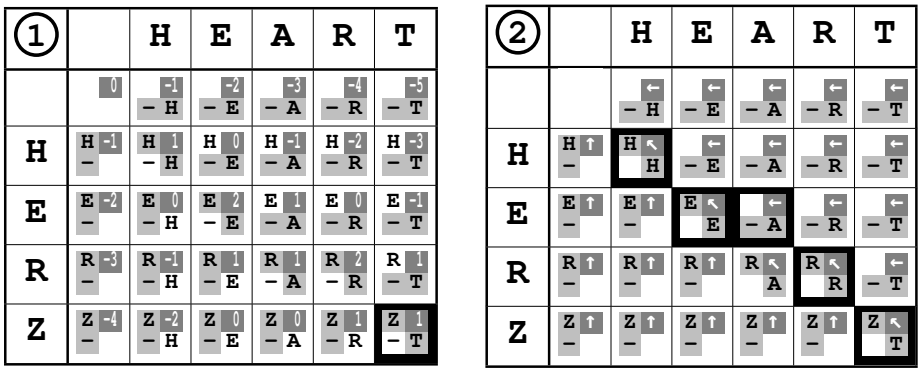

Fig. 2. The DPA matrix for the alignment of the strings `"HEART"` and `"HERZ"`.

3.2 Common Extensions of the Basic Algorithm

Many modifications of the basic algorithm have been proposed in order to address specific alignment problems. Among these modifications one can distinguish those which deal with the structure of sequences (*structural extensions*), and those which deal with their substance (*substantial extensions*). The former are based on the modification of the main recursion part of the algorithm, while the latter deal with the scoring function.

From a structural perspective, the basic algorithm aligns two sequences *globally*. All segments of the sequences are treated equally. Possible prefixes, infixes, and postfixes contribute equally to the alignment score. A *global alignment* may, however, not be what one wants to achieve with an alignment analysis. Often,

only specific domains of two sequences are comparable, while others are not. This problem is addressed in *local alignment analyses* where 'subsections of the sequences are aligned without reference to global patterns' [19, p. 12]. The most common solution for local alignment, the Smith-Waterman algorithm [21], seeks for the best alignment between subsequences of the sequences. While in this approach, only the most similar subsequences are aligned and the rest of the sequences is ignored, the DIALIGN algorithm [22] proceeds globally while at the same time searching for local similarities. The differences between global alignment, local alignment, and DIALIGN are illustrated in Tab. 1, where the strings `"GREEN CATFISH HUNTER"` and `"A FAT CAT HUNTER"` are aligned according to the different modes.

The structural extensions can help to enhance alignment analyses significantly, depending on the respective goal of the alignment analysis. The most important aspect of all alignment analyses, however, is the comparison on the segment level. This comparison is handled by the scoring function which penalizes the matching of segments and the introduction of gaps. A very simple scoring function which penalizes gaps and mismatches with -1 and matches with 1 is employed in the illustration of the basic algorithm in Fig. 2. This scoring function works well in certain applications, such as spelling correction or database searches. In linguistic or biological applications, however, it will often fail to find a satisfying alignment, since the segments usually exhibit different *degrees of similarity*. It is therefore important to modify the scoring function to yield individual scores depending on the segments which are being matched. In biology scoring functions for the alignment of protein alignments are usually derived from an empirical basis [23]. In linguistics it is common to derive scoring functions from phonetic features [8].

Table 1. Comparing the output of different alignment modes.

Mode	Alignment
global	G R E E N C A T F I S H H U N T E R A F A T C A T - - - - H U N T E R
local	GREEN CATFISH ‖ H U N T E R A FAT CAT ‖ H U N T E R
DIALIGN	- - - - - G R E E N C A T F I S H H U N T E R A F A T - - - - - C A T - - - - H U N T E R

3.3 Multiple Sequence Alignment

While it is guaranteed that the basic methods for pairwise alignment find the optimal alignment for two sequences depending on the respective criteria, multiple sequence alignment (MSA) is usually based on 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 analyzed [15, p. 345]. The most common way to address this problem in computational biology is to base the calculation on a *guide-tree* along which the sequences are successively aligned, moving from the leaves of the tree to its root. The guide-tree itself is reconstructed from the pairwise alignment scores, using traditional cluster algorithms such as UPGMA [24] or Neighbor-Joining [25]. This strategy is usually called *progressive alignment* [26].

Having the advantage of being fast to compute, progressive alignment bears, however, certain shortcomings. Due to the greediness of the procedure, 'any mistakes that are made in early steps of the procedure cannot be corrected by later steps' [19, p. 17]. The accuracy of progressive alignment can be enhanced in different ways. One common modification is the application of *profiles*. A profile represents the relative frequency of all segments of an MSA in all its positions and can therefore be seen as a sequence of vectors [20, p. 146f]. In profile-based approaches, sequences which have been joined during the alignment process are further represented as profiles and the traditional DPA is used for the alignment of profiles with profiles and profiles with sequences. The advantage of this approach is that position-specific information of already aligned sequences can be taken into account when joining more sequences along the guide tree. Fig. 3 gives an illustration for the profile-based progressive alignment of the sequences Russian *čelovek* [ʧɪlɐvʲɛk], Czech *člověk* [ʧlovʲɛk], Polish *człowiek* [ʧwɔvʲɛk], and Bulgarian *čovek* [ʧovɛk] "person".

Further enhancements which make progressive alignment less greedy consist of *pre-* and *post-processing* the data before and after the progressive analysis is carried out. *Library*-based alignment methods, such as the T-Coffee algorithm [3], for example, use the information given in sets (*libraries*) of pairwise global and local alignments of the sequences to derive an alignment-specific scoring function which is later used in the progressive phase. In a similar manner, a post-processing of a given alignment can be carried out with the help of *iterative refinement methods*. In these analyses one or more sequences are repeatedly selected from the completed alignment and realigned [20, p. 148f]. If the overall alignment score increases, the new alignment is retained, otherwise it is discarded.

4 Sequence Modeling in SCA

When dealing with automatic alignment analyses in historical linguistics, it is important to be clear about the underlying sequence model. Apparently, phonetic sequences differ crucially from biological sequences in several respects. The segmentation of sequences into phonetic segments, for example, poses a problem of itself which is addressed in the fields of phonology and phonetics. Another specific characteristic of phonetic sequences which is difficult to model is that they exhibit great *substantial differences* in the languages of the world. While all alignment approaches are based on the assumption that the sequences being compared are drawn from the same alphabet, this does not hold for languages

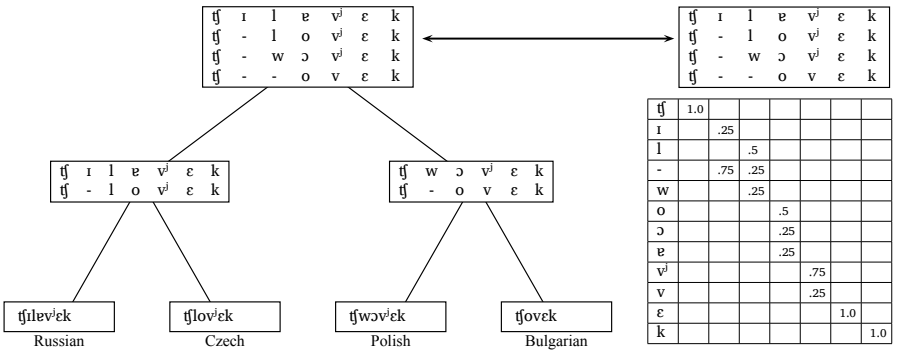

Fig. 3. Multiple sequence alignment based on a guide tree and profiles.

whose sound systems may differ crucially despite the fact that they are genetically related. SCA is based on new approaches to sequence modeling in historical linguistics which shall be presented in the following.

4.1 Paradigmatic Aspects

Sound Classes The concept of *sound classes* in historical linguistics goes back to A. B. Dolgopolsky [27, 28]. His main idea was to group sounds into different types such that 'phonetic correspondences inside a "type" are more regular than those between different "types"' [28, p. 35]. In his original study, Dolgopolsky proposed ten fundamental sound classes, based on partially empirical observations of sound correspondence frequencies in the languages of the world which are – unfortunately – not further specified by the author. [3]

In a recent study, Dolgopolsky's sound class model has been used as a heuristic device for automatic cognate identification [10]. According to this method, semantically identical basic words are judged to be cognate if their first two consonant classes match, otherwise, no cognacy is assumed. The advantage of this approach is that the number of false positives is usually considerably low, the apparent disadvantage lies in the fact that many true positives are missed, since no true alignment analysis is carried out. Thus, the cognate words German *Tochter* [tɔxtər] "daughter" and English *daughter* [dɔːtər] do not match in their first two consonant classes (`"TKTR"` vs. `"TTR"`). An alignment analysis, however, can easily show, that three of four consonant classes match perfectly.

The advantage of sound class representations of phonetic segments compared to pure phonetic representations lies in the specific probabilistic notion of segment similarity inherent in the sound class approach. Since sound classes constitute a model of sound correspondence probabilities, they can be seen as an

[3] The sound classes are: `P` (labial obstruents), `T` (dental obstruents), `S` (sibilants), `K` (velar obstruents and dental affricates), `M` (labial nasal), `N` (remaining nasals), `R` (liquids), `W` (labial approximant), `J` (palatal approximant) and `0` (laryngeals and initial velar nasal).

intermediate solution between the strict language-specific genotypic notion of segment similarity and the language-independent general notion of phenotypic similarity which were discussed in Sec. 2. Another advantage of sound classes is that this meta-phonological way to represent phonetic sequences helps us to get around the specific linguistic problem of aligning sequences drawn from two different alphabets: Ignoring minor phonetic differences enables us to stick to the "one alphabet paradigm" and make use of the traditional alignment algorithms.

Scoring Functions Dolgopolsky's original sound class approach defined sound classes as absolute entities. Transitions between sound classes were not allowed, although they are surely desirable, since transitions between classes are well-known to every historical linguist. Transition probabilities between sound classes can be easily modeled in the scoring functions of alignment algorithms. Scoring functions can be based on an *empirical* or a *theoretical* approach. Within an empirical approach scoring functions can be derived from studies on sound correspondence frequencies in the languages of the world. The SCA approach makes use of the data in [29] to derive such a scoring scheme for the sound class model employed by the ASJP project [30].

When deriving scoring functions from a theoretical basis, it is important to find a way to model the nature of sound change and sound correspondences. One crucial characteristic of certain well-known sound changes is their directionality, i.e. if certain sounds change, this change will go into a certain direction and the reverse change can rarely be attested. Other processes of sound change are bidirectional and it cannot be decided which direction occurs more frequently. Thus, regarding velar plosives ([k, g]), we know that they easily can be palatalized, and that the palatalization consists of certain steps, where the velares first become affricates and then turn into sibilants (e.g. [k, g] > [ʧ, ʦ, ʤ, ʣ] > [ʃ, ʒ, z, s]). The same process of palatalization may happen with dental plosives (e.g. [t, d] > [ʧ, ʦ, ʤ, ʣ] > [ʃ, ʒ, z, s]). The opposite direction of these changes, however, is rarely attested, and this is the reason, why we often find velar plosives and sibilants or dental plosives and sibilants corresponding regularly in genetically related languages, but rarely velar and dental plosives.

In order to reflect the directionality of certain sound changes, the SCA method applies the following approach: The scoring function is derived from a directed weighted graph. All sound classes which are known to be in very close connection to each other are connected by directed edges which reflect the direction of the respective sound changes. The assumed probability of the sound changes is defined by the edge weights. The higher the assumed probability of sound change, the smaller the weight. If sound change processes are not directional, both directions are reflected in the graph. The similarity score for two segments in the directed graph is calculated by subtracting the similarity score of one segment to itself from the length of the shortest path connecting two segments. Fig. 4 gives an example on how the similarity scores can be calculated for the above-mentioned cases of palatalization of dentals and velars: The resulting similarity score for dentals and fricatives is calculated by subtracting the length

of the shortest path (4) from the similarity score for a segment to itself (10). If no shortest path can be found, the similarity score is automatically set to 0.

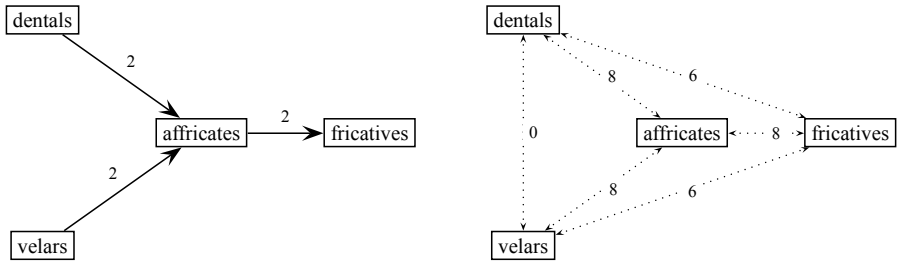

Fig. 4. Modelling the directionality of sound change patterns in scoring schemes.

4.2 Syntagmatic Aspects

Prosodic Profiles In biological alignment algorithms it is common to treat specific positions of certain sequences differently by modifying the penalties for the introduction of gaps in the alignment [31]. That certain types of sound change (including the loss of segments) are more likely to occur in specific environments is also a well-known fact in historical linguistics. In alignment analyses this can be modeled by constructing *prosodic profiles* from phonetic sequences which were first introduced in [2]. A prosodic profile is hereby understood as a vector representation of sequences which assigns a specific score to each segment depending on its sonority. The sonority score is derived from the sonority hierarchy of [32, p. 30].[4] Once a prosodic profile is constructed, the sound segments can be assigned to different *prosodic environments*. The SCA approach currently distinguishes seven different prosodic environments: `#` (word-initial consonant), `V` (word-initial vowel), `C` (ascending sonority), `v` (sonority peak), `c` (descending sonority), `$` (word-final consonant), and `w` (word-final vowel). Following [32, p. 31-34] these environments are ordered in a *hierarchy of strength*.[5] Relative weights for the modification of gap penalties and substitution scores are derived from this hierarchy in such way that it is easier to introduce gaps in weak positions than in strong ones, and that the score for the matching of segments is increased when they belong to the same prosodic environment. How relative weights are derived is illustrated in Tab. 2 for Bulgarian *jabălko* [jabəlka] "apple".

[4] The hierarchy along with relative scores for sonority, is: *plosives* (1), *affricates* (2), *fricatives* (3), *nasals* (4), *liquids* (5), *glides* (6), and *vowels* (7).

[5] The current hierarchy is: `# > V > C > c > v > $ > w`.

Table 2. Prosodic profiles, prosodic environments, and relative weights.

Phonetic Sequence	j	a	b	ə	l	k	a
Prosodic Profile	6	7	1	7	5	1	7
Prosodic Environments	#	v	C	v	c	C	>
Relative Weight	7	3	5	3	4	5	1

Secondary Sequence Structures A specific characteristic of sequences in general is that they may exhibit a *secondary structure* in addition to their *primary structure*. Primary structure is hereby understood as the order of segments, i.e. the smallest units of sequences. Apart from the primary structure, however, sequences can also have a secondary structure where the primary units are grouped into higher ones. The criteria for the secondary segmentation of sequences may vary, depending on the objects one is dealing with, or the specific goal of a certain analysis. Thus, in linguistic applications it may be reasonable to segment a word not only into sound units but also into syllables. Such a secondary segmentation is especially useful when dealing with South-East Asian tone languages like Chinese, since we know that the morphemes in these languages are almost exclusively monosyllabic, while the words usually are not. An alignment analysis of these languages should be able to keep track of the syllable boundaries and avoid matching the sounds of one syllable in one word with sounds in two syllables of the other. Traditional alignment analyses will usually fail to do so. A comparison of Haikou Chinese [zit^{3}] "sun" with Beijing Chinese [zη^{51}t^{h}ou^{1}] "sun" usually wrongly matches the dental plosives of both words, ignoring that one word has only one morpheme while the other one has two (see Tab. 3).[6]

Fortunately, it is possible to modify the traditional DPA algorithm to be sensitive to secondary sequence structures. I shall call such alignment analyses *secondary alignments* as opposed to traditional *primary alignments*. The SCA approach for secondary sequence alignment (SSA) employs the following strategy: Given a specific boundary marker which marks the end of a secondary segment (such as a tone letter in the phonetic transcription of Sinitic languages, or a whitespace in sentences), additional penalties are introduced into the main recursion. Whenever the recursion comes to a point where the boundary marker of one sequence could be matched with a character that is no boundary marker in the other sequence, or with a gap which is introduced *inside* a secondary segment, this matching is prohibited.

Table 3. Primary vs. Secondary Alignment analyses.

Primary Alignment							
Haikou	z	i	-	t	-	^{3}	
Beijing	z$_{\circlearrowleft}$	ɻ	^{51}	t^{h}	ou	^{1}	

Secondary Alignment							
Haikou	z	i	t	^{3}	-	-	-
Beijing	z$_{\circlearrowleft}$	ɻ	-	^{51}	t^{h}	ou	^{1}

[6] The data on the Chinese dialects is based on [33].

5 Phonetic Alignment in SCA

5.1 Working Procedure

The basic working procedure of SCA consists of four stages: (1) tokenization, (2) class conversion, (3) alignment analysis, and (4) IPA conversion. In stage (1) the input sequences (which should be given in IPA) are tokenized into phonetic segments. In stage (2) the segments are converted into their internal representation format, whereas each sequence is further represented by its corresponding sound class sequence and its prosodic profile. The pairwise or multiple alignment analysis is carried out in stage (3). After the alignment analysis has been carried out, the aligned sequences are converted back to original format in stage (4). This procedure is illustrated in Fig. 5.1 for the sequences German *Tochter* [tɔxtər] "daughter" and English *daughter* [dɔːtər].

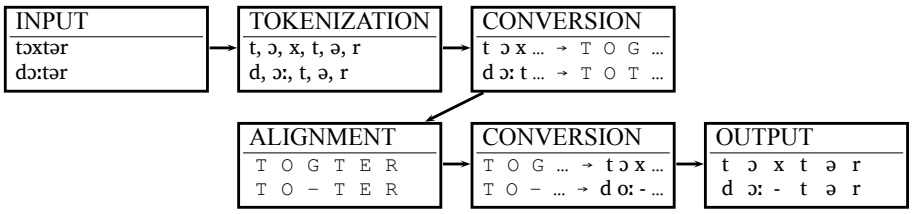

Fig. 5. The working procedure of SCA.

5.2 Sequence Models

LingPy comes along with three predefined sequence models and offers support for the import of user-defined models via specific input files. All models consist of two parts, one *class model* which handles the conversion of phonetic transcriptions into sound classes, and one *scoring scheme* which handles the transition probabilities between the sound classes. The three basic models are: (1) The DOLGO model which is based on Dolgopolsky's original proposal extended by a specific class for all vowels along with a simplified scoring function which prohibits the matching of vowels with consonants. (2) The SCA model which is based on extension of the DOLGO model, consisting of 28 sound classes and a refined scoring scheme that reflects common sound change processes which are often discussed in the literature. The scoring scheme was created by means of the procedure described in Sec. 4.1. (3) The ASJP model which is based on the sound class model used by the ASJP project [30], and a scoring scheme which was derived from a study on sound correspondence frequencies in the languages of the world as they are reflected in the ASJP database [29].

Table 4. The SCA sound class model.

No.	Cl.	Description	Examples
1	A	unrounded back vowels	a, ɑ
2	B	labial fricatives	f, β
3	C	dental / alveolar affricates	ʦ, ʣ, ʧ, ʤ
4	D	dental fricatives	θ , ð
5	E	unrounded mid vowels	e, ɛ
6	G	velar and uvual fricatives	ɣ , x
7	H	laryngeals	h, ʔ
8	I	unrounded close vowels	i, ɪ
9	J	palatal approxoimant	j
10	K	velare and uvular plosives	k, g
11	L	lateral approximants	l
12	M	labial nasal	m
13	N	nasals	n, ŋ
14	O	rounded back vowels	œ, ɒ
15	P	labial plosives	p, b
16	R	trills, taps, flaps	r
17	S	sibilant fricatives	s, z, ʃ, ʒ
18	T	dental / alveolar plosives	t, d
19	U	rounded mid vowels	ɔ , o
20	W	labial approx. / fricative	v, w
21	Y	rounded front vowels	u, ʊ, y
22	0	low even tones	11, 22
23	1	rising tones	13, 35
24	2	falling tones	51, 53
25	3	mid even tones	33
26	4	high even tones	44, 55
27	5	short tones	1, 2
28	6	complex tones	214

5.3 Pairwise Sequence Alignment

SCA supports all extensions to the basic algorithm for pairwise sequence alignment which are mentioned in Sec. 3.2. Additionally, all alignment modes are sensitive to secondary sequence structures, following the extension of the main recursion of the basic algorithm described in Sec. 4.2. When carrying out secondary alignment analyses, the boundary marker has to be defined by the user. By default the boundary marker is a tone letter as they are used in phonetic transcriptions of Sinitic languages. Gap penalties and substitution scores are modified with respect to prosodic context as described above. The relative weights of the different prosodic environments can be defined by the user.

5.4 Multiple Sequence Alignment

SCA's algorithm for multiple sequence alignment is based on the progressive alignment paradigm. Based on pairwise alignment analyses of all sequence pairs, a distance matrix is computed using the formula of [34] for the conversion of similarity into distance scores. With help of the Neighbor-Joining algorithm [25], a guide tree is reconstructed from the distance matrix and the sequences are successively aligned. In order to cope for the shortcomings of progressive alignment analyses, SCA employs profiles and offers the possibility to carry out a pre- and post-processing of the data. Furthermore, SCA includes a method for the detection of swapped sites in multiple alignments which is described in detail in [2].

The pre-processing in SCA follows the T-Coffee algorithm for multiple sequence alignment in biology [3]. The basic idea of the algorithm is to use the information given in pairwise alignment analyses of the data to derive an alignment-specific scoring matrix. The algorithm starts by computing a set of pairwise alignments of all input sequences, using different alignment approaches, such as, e.g., global and local alignments, or DIALIGN. Based on this set of pairwise alignments (the *primary library*) an alignment-specific scoring matrix is created.

In the initial stage, the substitution scores for all residue pairs in the scoring matrix are set to 0. After the pairwise alignments have been carried out, the scoring matrix is extended by increasing the score for each residue pair which occurs in the primary library by a certain weight. While the original T-Coffee algorithm derives the weight from sequence identity, SCA employs a different strategy. Given the sequences A and B, the weight W_{xy} for the residues x and y being matched in an alignment of A and B is derived by the formula

$$\frac{1}{2}\left(\frac{S_{AB}}{L_{AB}} + M_{xy}\right), \tag{1}$$

where S_{AB} is the similarity score of the alignment of A and B, L_{AB} is the length of the alignment, and M_{xy} is the original substitution score for the residues. If a given residue pair occurs more than once in the primary library, the sum of all weights is taken. In a second stage, the primary library is extended by means of a *triplet approach* where two sequences are aligned via a third sequence (see [3, p. 208f] for details), and the resulting weights are again added to the new scoring matrix. The specific strength of this approach is its sensitivity to both global as well as local similarities between sequences. This is illustrated in Fig. 6 where the words for "ashes" in three different dialects of Bai are aligned using simple (left) and library-based (right) progressive alignment.[7]

The post-processing is based on iterative refinement analyses, where a given MSA is split into two parts, one containing a set of probably misaligned sequences, and one containing the rest of all sequences. If realigning the two parts yields a new alignment with an improved score, the new alignment is retained, if not, it is discarded. In order to find probably misaligned sequences in a given MSA, SCA employs three different heuristics: (1) the *similar-gap-sites* heuristic which splits an MSA into sequences in which the same gaps have been introduced in the same positions; (2) the *flat-cluster* heuristic which splits an MSA into sets of sequences whose average distance is beyond a certain threshold, and (3) the *orphan*[8] heuristic which extracts those sequences from an MSA whose distance to all other sequences is greater than the average distance between all sequences.

Gongxing	x	w	ɛ	22	ɕ	y	55	ʁ	w	a	12
Jinman	kʰ	w	a	55	l	a	21	-	ɕ	u	55
Enqi	x	w	i	22	-	-	-	-	ɕ	u	24

Gongxing	x	w	ɛ	22	-	-	-	ɕ	y	55	ʁ	w	a	12
Jinman	kʰ	w	a	55	l	a	21	ɕ	u	55	-	-	-	-
Enqi	x	w	i	22	-	-	-	ɕ	u	24	-	-	-	-

Fig. 6. Simple (left) and library-based (right) progressive alignment.

[7] The dialect data is taken from [35].

[8] In evolutionary biology, the term 'orphan' is commonly used to refer to 'distant members of a family' [36, p. 2684].

6 Evaluation

6.1 Gold Standard

In previous analyses, it could be shown that sound class based alignment analyses perform equally well or even better than alternative proposals for pairwise [8] and multiple sequence alignment [9]. A drawback of these analyses was, however, that the testsets underlying the studies were either considerably small [1], or only covered a low range of genetically very close language varieties [2]. In order to test the method more closely, two new gold standards were compiled, one for MSA and one for PSA analyses. The sources of the data and further information are given in Tab. 5.

The MSA gold standard (see Supp. Mat. B) was designed to reflect a large range of different language varieties taken from different language families which show quite different kinds and degrees of variation. It consists of 600 manually edited multiple alignments, covering six different language families, 435 different taxa, and a total of 45 947 sequences. A large part of the gold standard (the subset of Bulgarian dialects) was compiled for the study of [9] and kindly provided by the authors. The rest of the gold standard was edited by the author himself.

The PSA gold standard was created by automatically extracting up to ten of the most divergent unique sequence pairs from each file of the MSA gold standard. Following the practice in computational biology, the diversity of the sequences was measured in terms of the Percentage Identity (PID) of the aligned sequences. The PID is calculated by dividing the number of identical positions in an alignment by the sum of aligned positions and internal gap positions [37].[9] This procedure yielded a set of 5 506 sequence pairs with an average PID of 17.14 % (see Supp. Mat. A).

Table 5. The gold standard for pairwise and multiple sequence alignments.

Dataset	**Languages**	**PSA**	**MSA**	**Words**	**Taxa**	**Source**
Sindial	Chinese dialects	200	20	341	40	[33]
Andean	Andean dialects (Aymara, Quechua)	597	94	983	20	[38]
BulDial	Bulgarian dialects	1504	152	32418	198	[9]
BaiDial	Bai dialects	892	90	1416	17	[35, 39]
NorDial	Norwegian dialects	496	51	2183	53	[40]
TPPSR	French dialects	707	82	3830	62	[41]
GerDial	Germanic languages and dialects	1110	111	4776	45	[42]

[9] Although this score has been criticized for certain obvious shortcomings [37], it provides an easy way to check the diversity of a given alignment independent of any further assumptions.

6.2 Evaluation Measures

The simplest way to test how well an alignment algorithm performs is to calculate the perfect-alignments score (PAS), i.e. the proportion of alignments which are identical with the gold standard. Since this score only reveals very strong tendencies, a couple of different methods have been proposed to test how well an alignment algorithm performs in comparison with a benchmark dataset [9, 36]. In computational biology, the most common evaluation measures are the column score (CS) and the sum-of-pairs score (SPS)[36]. The column score is calculated by dividing the number of identical columns in test and reference alignment by the total number of columns in the reference alignment. The sum-of-pairs score is defined as the size of the intersection of aligned segment pairs in test and reference alignment divided by the number of segment pairs in the reference alignment.

In [2], the advantages and disadvantages of different evaluation scores were discussed in detail, and certain shortcomings of both the CS and the SPS were pointed out. In practice, however, these problems rarely show up, and all evaluation measures discussed in [2] reflected the same tendencies in this study. Therefore, only the PAS, the CS and the SPS will be reported in the following.

6.3 Results

Pairwise Sequence Alignment The PSA gold standard was analyzed using the three different sound-class models provided by LingPy. For each model, analyses in four different modes were carried out: (1) a simple global alignment analysis (BASIC), (2) a global alignment analysis in which gap costs and substitution scores were scaled in dependence of prosodic environments (SCALE), (3) a simple global alignment analysis which was sensitive to secondary sequence structures (SEC), and (4) a global alignment analysis which was sensitive to both prosodic environments and secondary sequence structures (SEC/SCALE). These different modes were chosen in order to test to what degree the syntagmatic modifications described in Sec. 4.2 might influence the performance of SCA.

As can be seen from the results shown in Tab. 6, the SCA sound-class model performs best throughout all modes, while DOLGO performs worst. Furthermore, the accuracy of the scores increases as the modes become more complex, with SEC/SCALE showing the best performance. Given that the differences in CS and SPS between BASIC and SEC/SCALE are significant for all sound-class models,[10] a clear improvement of secondary alignment analyses in combination with prosodic profiles can be attested. The benefits of secondary alignment analyses become even more evident when considering only the 1 092 sequence pairs

[10] Assuming equal variances, a two sample t-test yielded: t=-2.20, p=0.03 for SCA (CS); t=-2.28, p=0.02 for SCA (SPS); t=-3.93, p=0.00 for ASJP (CS); t=-2.94, p=0.00 for ASJP (SPS); t=-3.91, p=0.00 for DOLGO (CS); t=-3.48, p=0.00 for DOLGO (SPS).

belonging to dialects of the tonal languages Bai and Chinese, where a drastic increase in alignment accuracy can be reported for all models (see Fig. 7).

Table 6. Comparing the results for the four different modes on the PSA gold standard.

	BASIC			**SCALE**			**SEC**			**SEC/SCALE**		
Model	**PAS**	**CS**	**SPS**	**PAS**	**CS**	**SPS**	**PAS**	**CS**	**SPS**	**PAS**	**CS**	**SPS**
SCA	85.72	92.32	95.93	86.12	92.38	96.01	87.21	93.37	96.57	87.67	93.47	96.67
ASJP	83.62	91.17	95.43	85.05	91.91	95.76	84.87	92.00	95.90	86.63	92.89	96.34
DOLGO	82.02	89.47	94.29	83.73	90.65	94.96	83.33	90.42	94.86	85.09	91.63	95.53

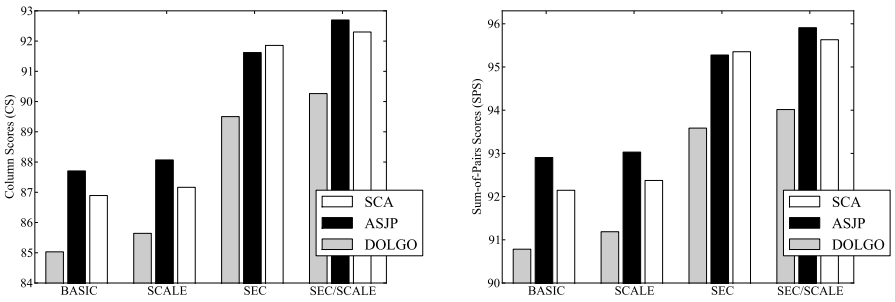

Fig. 7. CS and SPS in the tonal partition of the PSA gold standard.

Multiple Sequence Alignment In a way similar to the analysis of the PSA gold standard, the MSA gold standard was analyzed using the three different sound-class models provided by LingPy. Again, analyses in four different modes were carried out for each model: (1) a progressive alignment analysis (PROG), (2) a progressive alignment analysis with iterative refinementx (ITER), (3) a library-based alignment analysis (LIB), and (4) a library-based alignment analysis in combination with iterative refinementx (LIB/ITER). The iterative refinement analysis was based on the three heuristics described in Sec. 5.4. The library was created from pairwise global, local, and DIALIGN alignment analyses of all sequence pairs. All alignment analyses were based on the extensions for secondary alignment and prosodic profiles described in Sec. 4.2.

The results for the analyses are given in Tab. 7. As can be seen from the table, the analyses using the SCA model again outperformed the other models, while the analyses using the DOLGO model again performed worst. The pre- and post-processing of the data also results in clear improvements of the analyses regardless of the underlying sound-class model, whereas the combination of

library-based alignment and iterative refinement seems to be the best approach, showing significant improvements for CS and SPS in almost all models.[11]

In order to check where the specific strengths of the different sound-class models lie, the results for the analyses were divided into four partitions based on the PID of the gold standard alignments: PID-100 (100 – 75), PID-75 (75 – 50), PID-50 (50 – 25), and PID-25 (25 – 0). The results for LIB/ITER on these partitions are plotted in Fig. 8. As one may expect, the accuracy of all methods decreases the more divergent the sequences become. Yet while there are only minor differences in the accuracy of the SCA model compared to the ASJP model in the first three partitions, the SCA model performs considerably better in the PID-25 partition of highly divergent MSAs. This shows that the SCA model is specifically apt for the task of sequence comparison in greater time depths.

Table 7. Comparing the results for the four different modes on the MSA gold standard.

	PROG			**ITER**			**LIB**			**LIB/ITER**		
Model	**PAS**	**CS**	**SPS**	**PAS**	**CS**	**SPS**	**PAS**	**CS**	**SPS**	**PAS**	**CS**	**SPS**
SCA	79.83	89.74	98.00	83.83	91.98	98.72	81.50	90.42	98.64	86.17	93.19	99.26
ASJP	78.00	88.73	97.88	84.17	92.26	98.70	80.83	89.99	98.76	84.83	92.25	99.17
DOLGO	76.33	87.73	97.68	78.67	89.10	98.13	79.67	89.62	98.33	80.83	90.05	98.42

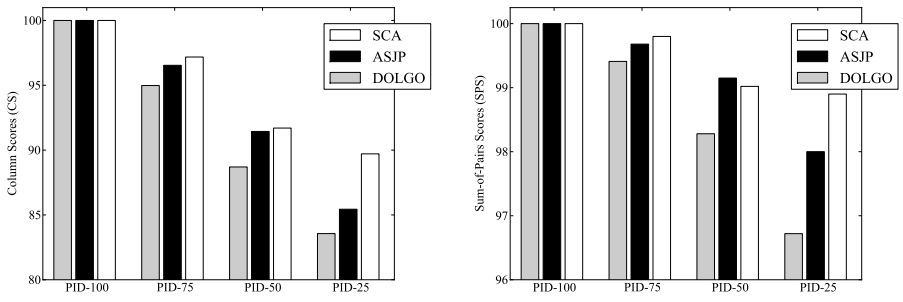

Fig. 8. CS and SPS in the different partitions of the MSA gold standard.

Identification of Swapped Sites Of the 600 files in the MSA gold standard, 50 MSAs were identified to contain swapped sites. Along with the LIB/ITER

[11] Assuming equal variances, a two sample t-test yielded: t=-2.91, p=0.00 for SCA (CS); t=-4.28, p=0.00 for SCA (SPS), t=-2.81, p=0.00 for ASJP (CS); t=-4.41, p=0.00 for ASJP (SPS), t=-1.75, p=0.08 for DOLGO (CS); t=-2.15, p=0.03 for DOLGO (SPS).

analyses, the algorithm for swap detection in multiple sequence alignments described in [2] was tested. The results are given in Tab. 8. As can be seen from the table, the SCA model again performed best. It correctly identified 46 of 50 swaps while at the same time only 2 swaps were wrongly proposed.

Table 8. Identification of swaps.

	SCA	**ASJP**	**DOLGO**
True Positives	46	41	44
False Positives	2	1	9
False Negatives	4	9	6

7 Conclusion

In this paper, I presented the most recent version of the SCA method for pairwise and multiple phonetic alignments. As could be shown in extensive tests, the current state of the method has many advantages compared to both alternative approaches and older versions of sound-class-based alignments. The specific strength of the SCA method lies in the specific way traditional linguistic concepts are modeled and combined with recent computational approaches. The research on automatic methods for phonetic alignment is still in its infancy, and the results of pairwise and multiple alignment analyses lack far behind the intuitive judgments of trained historical linguists. This should, however, not discourage us from trying to improve our automatic methods, but rather motivate us to put more effort into the development of new models which bring traditional and automatic approaches closer to each other.

References

1. List, J.M.: Phonetic alignment based on sound classes. In Slavkovik, M., ed.: Proceedings of the 15th Student Session of the European Summer School for Logic, Language and Information, Kopenhagen (2010) 192–202
2. List, J.M.: Multiple sequence alignment in historical linguistics. A sound class based approach. Proceedings of ConSOLE XIX (2011) (forthcoming)
3. Notredame, C., Higgins, D.G., Heringa, J.: T-Coffee. A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology **302** (2000) 205–217
4. Gray, R.D., Atkinson, Q.D.: Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature **426**(6965) (2003) 435–439
5. 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., Egorov, D.: Automated dating of the world's language families based on lexical similarity. Current Anthropology **52**(6) (2011) 841–875

6. 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. McDonald Institute for Archaeological Research, Cambridge (2000) 167–188
7. Kessler, B.: The significance of word lists. Statistical tests for investigating historical connections between languages. CSLI Publications, Stanford (2001)
8. Kondrak, G.: Algorithms for language reconstruction. Dissertation, University of Toronto, Toronto (2002)
9. Prokić, J., Wieling, M., Nerbonne, J.: Multiple sequence alignments in linguistics. In: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education, Stroudsburg, PA, Association for Computational Linguistics (2009) 18–25
10. Turchin, P., Peiros, I., Gell-Mann, M.: Analyzing genetic connections between languages by matching consonant classes. Journal of Language Relationship **3** (2010) 117–126
11. Covington, M.A.: An algorithm to align words for historical comparison. Computational Linguistics **22**(4) (1996) 481–496
12. Ross, M., Durie, M.: Introduction. In Durie, M., ed.: The comparative method reviewed. Regularity and irregularity in language change. Oxford University Press, New York (1996) 3–38
13. Trask, R.L., ed.: The dictionary of historical and comparative linguistics. Edinburgh University Press, Edinburgh (2000)
14. Lass, R.: Historical linguistics and language change. Cambridge University Press, Cambridge (1997)
15. Gusfield, D.: Algorithms on strings, trees and sequences. Cambridge University Press, Cambridge (1997)
16. 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
17. Wagner, R.A., Fischer, M.J.: The string-to-string correction problem. Journal of the Association for Computing Machinery **21**(1) (1974) 168–173
18. Eddy, S.R.: Where did the BLOSUM62 alignment score matrix come from? Nature Biotechnology **22**(8) (2004) 1035–1036
19. Rosenberg, M.S.: Sequence alignment. Concepts and history. In Rosenberg, M.S., ed.: Sequence alignment. Methods, models, concepts, and strategies. University of California Press, Berkeley and Los Angeles and London (2009) 1–22
20. Durbin, R., Eddy, S.R., Krogh, A., Mitchinson, G.: Biological sequence analysis. Probabilistic models of proteins and nucleic acids. 7 edn. Cambridge University Press, Cambridge (2002)
21. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology **1** (1981) 195–197
22. Morgenstern, B., Dress, A., Werner, Thomas, D.: Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proceedings of the National Acadamy of Science, USA **93** (October 1996) 12098–12103
23. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. PNAS **89**(22) (1992) 10915–10919
24. Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin **28** (1958) 1409–1438
25. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**(4) (1987) 406–425

26. 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
27. Dolgopolsky, A.B.: 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** (1964) 53–63
28. Dolgopolsky, A.B.: A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. In Shevoroshkin, V.V., ed.: Typology, Relationship and Time. Karoma Publisher, Ann Arbor (1986) 27–50
29. Brown, C.H., Holman, E.W., Wichmann, S.: Sound correspondences in the world's languages. Online manuscript, PDF: http://wwwstaff.eva.mpg.de/∼wichmann/wwcPaper23.pdf (2011)
30. Brown, C.H., Holman, E.W., Wichmann, S., Velupillai, V., Cysouw, M.: Automated classification of the world's languages. Sprachtypologie und Universalienforschung **61**(4) (2008) 285–308
31. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W. Nucleic Acids Research **22**(22) (1994) 4673–4680
32. Geisler, H.: Akzent und Lautwandel in der Romania. Narr, Tübingen (1992)
33. Hóu, J., ed.: Xiàndài Hànyǔ fāngyán yīnkù [Phonological database of Chinese dialects]. Shànghǎi Jiàoyǔ, Shanghai (2004)
34. Downey, S.S., Hallmark, B., Cox, M.P., Norquest, P., Lansing, S.: Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction. Journal of Quantitative Linguistics **15**(4) (2008) 340–369
35. Wang, F.: Comparison of languages in contact. Institute of Linguistics Academia Sinica, Taipei (2006)
36. Thompson, J.D., Plewniak, F., Poch, O.: A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Research **27**(13) (1999) 2682–2690
37. Raghava, G.P.S., Barton, G.J.: Quantification of the variation in percentage identity for protein sequence alignments. BMC Bioinformatics **7**(415) (2006)
38. Heggarty, P.: Sounds of the Andean languages. Online ressource http://www.quechua.org.uk/.
39. Allen, B.: Bai Dialect Survey. SIL International (2007)
40. Almberg, J., Skarbø, K.: Nordavinden og sola. En norsk dialektprøvedatabase på nettet [The North Wind and the Sun. A Norwegian dialect database on the web]. Online ressource (2011) http://www.ling.hf.ntnu.no/nos/.
41. Gauchat, L., Jeanjaquet, J., Tappolet, E.: Tableaux phonétiques des patois suisses romands. Attinger, Neuchâtel (1925)
42. Renfrew, C., Heggarty, P.: Languages and origins in europe. Online resource, URL: http://www.languagesandpeoples.com/