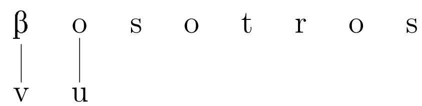
Phonological Correspondence as a Tool for Historical Analysis: An Algorithm for Word Alignment

Daniel M. Albro

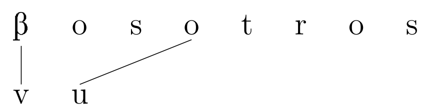
March 11, 1997

1 Introduction

In historical linguistics research it is often necessary to compare words from different languages or different stages of a single language in order to reconstruct sound changes, or simply in order to determine whether the words are cognate. One stage of this process is aligning the words with one another. For example, Spanish *vosotros* ([βosotros]) and French *vous* ([vu]) should align as follows:



as opposed to, for example:



The algorithm to be discussed here attempts to automate this process and find a *best phonological correspondence* between words.

2 The Program

The underpinnings of this algorithm were originally written as part of a word-comparison algorithm that would ignore various types of phonological changes in order to split words up into morphemes (Albro 1997). Covington's (1996) article led to the realization that the algorithm might be used as an improvement (or at least a different take) on his method of aligning words. In its original form, the algorithm allowed the user to specify three different types of phonological laxity:

1. *Featural laxity*—the user could set the maximum number of features that could differ from one phoneme to another. An SPE-style feature-matrix representation was (and is) used (Chomsky and Halle 1968).

2. *Segmental laxity*—the user could set the maximum number of segments that could be deleted or inserted to arrive at one word from another.

3. *Order laxity*—the user could set the maximum number of allowable segment precedence violations per word.

In this incarnation, the feature system given in Halle and Clements (1983) is used, and each feature is given a distinct penalty, according to how likely the feature value is to change, historically. The values used were guesses only, and may be changed by the user. Vowel features were assigned the least penalty — 0.7; SONORANT and CONS were assigned the highest penalty — 3; NASAL, VOICE, and STRIDENT were assigned a low penalty — 1; and the rest were assigned a penalty of 2. In addition to these parameters, the user may also specify a maximum number of order violations in order to keep metathesis down to plausible levels (the default is two), and weights for feature violations, segment violations, and order violations. The weights are used to compute a *score* for each hypothesized alignment between words. The score for an alignment is just the sum of the feature violations, segment violations (number of deletions or insertions), and order violations (number of metatheses) incurred in transforming from one word to the other. By default, feature violations are weighted by one, segment violations by four, and order violations by three. In addition, the algorithm penalizes "gaps"—alignments of segments to non-contiguous segments in the associated form. This is how the proper alignment for *vosotros*/*vous* was determined above — both alignments have the same number of feature changes, segment deletions and insertions, and metatheses, but the first has no gaps and the second does, so the first wins. By default we weight gaps by 1.5, and the number of gaps formed is added to the score for a particular alignment. As an example, in translating from "tat" to "st" we incur one feature violation (changing of voice in the initial segment) and one segment violation. No order or gap violations are incurred. Thus the score is $1 \cdot 1 + 1 \cdot 4 + 0 \cdot 3 + 0 \cdot 1.5 = 5$. The basic function of the algorithm is to find the alignment or alignments with the lowest associated score.

3 Algorithm Overview

In order to conserve space and time, I will not go into the details of the algorithm here, but the general overview is as follows. We slowly increase the number of allowed feature violations per segment from 0 up to the number of features (eighteen). At each stage we produce a chart of correspondences between segments in the two words, using segment equality with allowance for the current number of feature violations. As an example, comparing 'rats' with 'tart' and allowing one feature violation per phoneme would give the following chart:

	t	a	r	t
r			✓	
a		✓		
t	✓			✓
s	✓			✓

The algorithm then attempts to find the paths through the chart that have the lowest associated score, using (more or less) a bounded depth-first search plus some additional backtracking.

4 Efficiency

The algorithm used is not at all efficient. In fact, it is exponential. However, I believe that it would be fairly simple to transform it into a polynomial algorithm using memoization and dynamic programming techniques along the lines of Masek and Paterson (1980), that is, by storing the intermediate results of calculations and attempting to look up information rather than recomputing it.

5 Examples

The words used in this section are taken from Covington (1996). Here is some raw program output wherein we attempt to align French and Spanish words:

```
Alignment of y o and zh E:
Best score: 8.1
y (0) corresponds to zh (0)
o (1) corresponds to E (1)

Alignment of t u and t ue:
Best score: 0.7
t (0) corresponds to t (0)
u (1) corresponds to ue (1)

Alignment of n o s o t r o s and n u:
Best score: 24.7
n (0) corresponds to n (0)
o (1) corresponds to u (1)

Alignment of k y e n and k i:
Best score: 10.0
k (0) corresponds to k (0)
y (1) corresponds to i (1)

Alignment of k e and k w a:
Best score: 7.6
```

```
k (0) corresponds to k (0)
e (1) corresponds to a (2)

Alignment of t o d o s and t u:
Best score: 12.7
t (0) corresponds to t (0)
o (1) corresponds to u (1)

Alignment of u n a and ue n:
Best score: 4.7
u (0) corresponds to ue (0)
n (1) corresponds to n (1)

Alignment of d o s and d oe:
Best score: 4.7
d (0) corresponds to d (0)
o (1) corresponds to oe (1)

Alignment of t r e s and t r w a:
Best score: 11.6
t (0) corresponds to t (0)
r (1) corresponds to r (1)
e (2) corresponds to a (3)

Alignment of o m b r e and o m:
Best score: 12.0
o (0) corresponds to o (0)
m (1) corresponds to m (1)

Alignment of a r b o l and a r b r E:
Best score: 9.1
a (0) corresponds to a (0)
r (1) corresponds to r (1)
b (2) corresponds to b (2)
o (3) corresponds to E (4)
l (4) corresponds to r (3)

Alignment of p l u m a and p l ue m:
Best score: 4.7
p (0) corresponds to p (0)
l (1) corresponds to l (1)
u (2) corresponds to ue (2)
m (3) corresponds to m (3)

Alignment of k a b e th a and k a p:
Best score: 13.0
k (0) corresponds to k (0)
a (1) corresponds to a (1)
b (2) corresponds to p (2)

Alignment of b o k a and b u sh:
Best score: 8.7
b (0) corresponds to b (0)
o (1) corresponds to u (1)
k (2) corresponds to sh (2)

Alignment of p y e and p y e:
Best score: 0.0
p (0) corresponds to p (0)
y (1) corresponds to y (1)
e (2) corresponds to e (2)

Alignment of k o r a th o n and k oe r:
Best score: 16.7
k (0) corresponds to k (0)
o (1) corresponds to oe (1)
r (2) corresponds to r (2)

Alignment of b e r and v w a r:
Best score: 9.6
b (0) corresponds to v (0)
e (1) corresponds to a (2)
r (2) corresponds to r (3)

Alignment of b e n i r and v E n i r:
Best score: 2.7
b (0) corresponds to v (0)
e (1) corresponds to E (1)
n (2) corresponds to n (2)
i (3) corresponds to i (3)
r (4) corresponds to r (4)

Alignment of d e th i r and d i r:
Best score: 9.5
d (0) corresponds to d (0)
i (3) corresponds to i (1)
r (4) corresponds to r (2)

Alignment of p o b r e and p o v r E:
Best score: 2.7
p (0) corresponds to p (0)
o (1) corresponds to o (1)
b (2) corresponds to v (2)
r (3) corresponds to r (3)
e (4) corresponds to E (4)
```

These languages are fairly similar, and as a result the algorithm works fairly well. In fact, for *arbol*/*arbre* it works somewhat better than Covington's (1996) algorithm, due mostly to the incorporation of metathesis. Now, we will show English and German:

```
Alignment of dh I s and d i z E s:
Best score: 9.7
dh (0) corresponds to z (2)
I (1) corresponds to E (3)
s (2) corresponds to s (4)

Alignment of dh ae t and d a s:
Best score: 4.7
dh (0) corresponds to d (0)
ae (1) corresponds to a (1)
t (2) corresponds to s (2)

Alignment of w a t and v a s:
Best score: 10.0
a (1) corresponds to a (1)
t (2) corresponds to s (2)

-or-

w (0) corresponds to v (0)
a (1) corresponds to a (1)
t (2) corresponds to s (2)

Alignment of n a t and n i x t:
Best score: 8.3
n (0) corresponds to n (0)
a (1) corresponds to i (1)
t (2) corresponds to t (3)

Alignment of l o ng and l a ng:
Best score: 2.1
l (0) corresponds to l (0)
o (1) corresponds to a (1)
ng (2) corresponds to ng (2)

Alignment of m ae n and m a n:
Best score: 0.7
m (0) corresponds to m (0)
ae (1) corresponds to a (1)
n (2) corresponds to n (2)

Alignment of f l e sh and f l a y sh:
Best score: 7.6
f (0) corresponds to f (0)
l (1) corresponds to l (1)
e (2) corresponds to a (2)
sh (3) corresponds to sh (4)

Alignment of b l E d and b l u t:
Best score: 3.8
b (0) corresponds to b (0)
l (1) corresponds to l (1)
E (2) corresponds to u (2)
d (3) corresponds to t (3)

Alignment of f e dh E r and f e d E r:
Best score: 2.0
f (0) corresponds to f (0)
e (1) corresponds to e (1)
dh (2) corresponds to d (2)
E (3) corresponds to E (3)
r (4) corresponds to r (4)

Alignment of h ae r and h a r:
Best score: 0.7
h (0) corresponds to h (0)
ae (1) corresponds to a (1)
r (2) corresponds to r (2)

Alignment of i r and o r:
Best score: 2.1
i (0) corresponds to o (0)
r (1) corresponds to r (1)

Alignment of a y and a w g E:
Best score: 11.4
a (0) corresponds to a (0)
y (1) corresponds to w (1)

Alignment of n o w z and n a z E:
Best score: 9.9
n (0) corresponds to n (0)
o (1) corresponds to a (1)
w (2) corresponds to E (3)
z (3) corresponds to z (2)

Alignment of m a w th and m u n t:
Best score: 12.0
m (0) corresponds to m (0)
a (1) corresponds to n (2)
w (2) corresponds to u (1)
th (3) corresponds to t (3)

Alignment of t E ng and t s u ng E:
Best score: 12.3
t (0) corresponds to t (0)
E (1) corresponds to u (2)
ng (2) corresponds to ng (3)

Alignment of f u t and f u s:
Best score: 2.0
f (0) corresponds to f (0)
u (1) corresponds to u (1)
t (2) corresponds to s (2)

Alignment of n i y and k n i:
Best score: 8.0
n (0) corresponds to n (1)
i (1) corresponds to i (2)
```

```
Alignment of h ae n d and h a n t:
Best score: 1.7
h (0) corresponds to h (0)
ae (1) corresponds to a (1)
n (2) corresponds to n (2)
d (3) corresponds to t (3)

Alignment of h a r t and h e r t s:
Best score: 6.1
h (0) corresponds to h (0)
a (1) corresponds to e (1)
r (2) corresponds to r (2)
t (3) corresponds to t (3)

Alignment of l i v E r and l e b E r:
Best score: 2.7
l (0) corresponds to l (0)
i (1) corresponds to e (1)
v (2) corresponds to b (2)
E (3) corresponds to E (3)
r (4) corresponds to r (4)
```

These languages are also fairly close, and the algorithm only messes up on a few words, most notably the first — "this." Note that "what" shows two possible alignments, of which the second is better. Looking at the English data suggests that some knowledge of diphthongs as possible correspondents to single vowels might be a good addition to the algorithm.

Finally, we will look at English versus Latin, in which I have given both the ecclesiastical and classical pronunciations:

```
Alignment of ae n d and a n t e:
Best score: 5.7
ae (0) corresponds to a (0)
n (1) corresponds to n (1)
d (2) corresponds to t (2)

Alignment of ae t and a d:
Best score: 1.7
ae (0) corresponds to a (0)
t (1) corresponds to d (1)

Alignment of b l o w and f l a r e:
Best score: 14.7
b (0) corresponds to f (0)
l (1) corresponds to l (1)
o (2) corresponds to a (2)
w (3) corresponds to e (4)

Alignment of i r and a w r i s:
Best score: 16.0
i (0) corresponds to i (3)
r (1) corresponds to r (2)

-or-

i (0) corresponds to i (3)
r (1) corresponds to s (4)

Alignment of i y t and e d e r e:
Best score: 16.4
i (0) corresponds to e (0)
y (1) corresponds to e (2)
t (2) corresponds to d (1)

Alignment of f i sh and p i s k i s:
Best score: 16.0
f (0) corresponds to p (0)
i (1) corresponds to i (1)
sh (2) corresponds to s (2)

Alignment of f i sh and p i sh i s:
Best score: 10.0
f (0) corresponds to p (0)
i (1) corresponds to i (1)
sh (2) corresponds to sh (2)

Alignment of f l o w and f l u e r e:
Best score: 12.8
f (0) corresponds to f (0)
l (1) corresponds to l (1)
o (2) corresponds to u (2)
w (3) corresponds to e (3)

Alignment of s t a r and s t e l l a:
Best score: 14.1
s (0) corresponds to s (0)
t (1) corresponds to t (1)
a (2) corresponds to e (2)
r (3) corresponds to l (3)

Alignment of f ^ l and p l e n u s:
Best score: 19.4
f (0) corresponds to p (0)
^ (1) corresponds to e (2)
```

```
l (2) corresponds to l (1)

Alignment of g r ae s and g r a m e n:
Best score: 16.7
g (0) corresponds to g (0)
r (1) corresponds to r (1)
ae (2) corresponds to a (2)

Alignment of h a r t and k o r d i s:
Best score: 19.1
a (1) corresponds to o (1)
r (2) corresponds to r (2)
t (3) corresponds to d (3)

Alignment of h o r n and k o r n u:
Best score: 12.0
o (1) corresponds to o (1)
r (2) corresponds to r (2)
n (3) corresponds to n (3)

Alignment of a y and e g o:
Best score: 11.7
a (0) corresponds to e (0)
y (1) corresponds to o (2)

Alignment of n i y and g e n u:
Best score: 14.1
n (0) corresponds to n (2)
i (1) corresponds to e (1)
y (2) corresponds to u (3)

Alignment of m E dh E r and m a t e r:
Best score: 5.1
m (0) corresponds to m (0)
E (1) corresponds to a (1)
dh (2) corresponds to t (2)
E (3) corresponds to e (3)
r (4) corresponds to r (4)

Alignment of m a w n t E n and m o n s:
Best score: 17.6
m (0) corresponds to m (0)
a (1) corresponds to o (1)
n (3) corresponds to n (2)
t (4) corresponds to s (3)

Alignment of n e y m and n o m e n:
Best score: 12.1
n (0) corresponds to n (0)
e (1) corresponds to o (1)
y (2) corresponds to e (3)
m (3) corresponds to m (2)

Alignment of n u w and n o w u s:
Best score: 8.7
n (0) corresponds to n (0)
u (1) corresponds to o (1)
w (2) corresponds to w (2)

Alignment of n u w and n o v u s:
Best score: 12.2
n (0) corresponds to n (0)
u (1) corresponds to o (1)
w (2) corresponds to u (3)

Alignment of w E n and u n u s:
Best score: 12.8
w (0) corresponds to u (0)
E (1) corresponds to u (2)
n (2) corresponds to n (1)

Alignment of r a w n d and r o t u n d u s:
Best score: 17.6
r (0) corresponds to r (0)
a (1) corresponds to o (1)
w (2) corresponds to u (3)
n (3) corresponds to n (4)
d (4) corresponds to d (5)

Alignment of s o w and s u e r e:
Best score: 12.8
s (0) corresponds to s (0)
o (1) corresponds to u (1)
w (2) corresponds to e (2)

Alignment of th r i y and t r e s:
Best score: 10.7
th (0) corresponds to t (0)
r (1) corresponds to r (1)
i (2) corresponds to e (2)

Alignment of t u w th and d e n t i s:
Best score: 19.0
t (0) corresponds to d (0)
u (1) corresponds to e (1)
w (2) corresponds to i (4)
th (3) corresponds to s (5)

Alignment of th i n and t e n u i s:
Best score: 14.7
th (0) corresponds to t (0)
i (1) corresponds to e (1)
n (2) corresponds to n (2)
```

Here the languages are more spread out in time, and the algorithm occasionally gives incorrect results, mostly by over-enthusiastically skipping segments and engaging in metathesis, for example in "eat." Perhaps a parameter change might alleviate this problem in the future.

6 Future Directions

The most pressing need is for improvement in efficiency. The algorithm takes approximately half an hour on a fast machine to analyze the English-Latin data (although it takes about five minutes for the German, for some reason), and this is too slow for the algorithm to be used with large data sets. I have also done some work on using autosegmental representations based on precedence, dominance, and overlap relations. I believe these will allow a more sophisticated feature system, although their use will probably lead to a decline in efficiency.

7 Discussion

There are two possible conclusions that could be derived from this work. One is that it might be possible to use this tool or others like it to automate the comparative method for language reconstruction by taking lists of word-gloss pairs from different languages and automatically computing the phonological distance and correspondence between the words whose glosses match. This task might be aided also by the program described in Albro (1997), which automatically finds the morpheme divisions in words given a gloss and a surface form. Thus one could take lists of words from different languages, select the words whose glosses correspond to the basic vocabulary list, split those words into morphemes, use the algorithm to compare the corresponding morphemes between languages, and use the resulting alignments and correspondence distances to reconstruct the sound changes from the proto-language, if there is one. Of course, in order to be used in this way, the program would have to be made to be much faster than it currently is, but doing this would be relatively simple. The other possible conclusion, however, is that programs doing this sort of thing make errors that a human would not make, and therefore either humans are doing the right thing and machines should be left out of the comparative reconstruction business, or humans are making unwarranted assumptions and the reliability of the comparative method should be reëxamined.

References

Albro, Daniel M., 1997. A morpho-phonological learner. Unpublished manuscript.

Chomsky, Noam A., and Morris Halle. 1968. *The Sound Pattern of English*. New York, NY: Harper and Row.

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

Halle, Morris, and G. N. Clements. 1983. *Problem Book in Phonology*. Cambridge, MA: MIT Press.

Masek, William J., and Michael S. Paterson. 1980. A faster algorithm computing string edit distances. *Journal of Computer and System Sciences* 20.18–31.