Southern African Linguistics and Applied Language Studies 2009, 27(4): 381–389
Printed in South Africa — All rights reserved

SOUTHERN AFRICAN LINGUISTICS
AND APPLIED LANGUAGE STUDIES
ISSN 1607–3614 EISSN 1727–9461
DOI: 10.2989/SALALS.2009.27.4.2.1022

Modelling distances between genetically related languages using an extended weighted Levenshtein distance

Filip Paluncić*, Hendrik C Ferreira, Theo G Swart and Willem A Clarke
Department of Electrical and Electronic Engineering Science, University of Johannesburg, Auckland Park, 2006, South Africa
**Corresponding author, e-mail: fpaluncic@uj.ac.za*

Abstract: This article proposes the use of an extended weighted Levenshtein distance to model the time depth between parent and direct descendant languages and also the dialectal separation between sibling languages. The parent language is usually a proto-language, a hypothetical reconstructed language, whose precise date is usually conjectural. Phonology is used as an indicator of language difference, which is modelled by means of an extended weighted Levenshtein distance. This idea is applied specifically to the Iranian language family.

Introduction

Levenshtein (1965) originally defined the Levenshtein distance as a means to determine the insertion/deletion error correcting capability of a codebook. However, insertion/deletion correcting codes have not found a wide application, partly due to the fact that many problems associated with these codes remain unsolved. Nevertheless, the Levenshtein distance has been applied to many applications outside error control coding, particularly in computer science, biology and linguistics. Kruskal (1983) presents an overview of the various applications of the Levenshtein distance. This article presents an application of an extension of the Levenshtein distance in linguistics from the perspective of information theory.

Extensions of the Levenshtein distance have been proposed for correction of spelling errors. Okuda, Tanaka and Kasai (1976) proposed a weighted Levenshtein distance to correct garbled words. A non-negative weight is assigned to substitutions, insertions and deletions. Damerau (1964) also extended on the Levenshtein distance by introducing transpositions in addition to substitutions, insertions and deletions. Oommen and Loke (1997) considered the problem of when a substitution occurs within the transposed symbols. These metrics are not limited to spelling error detection and correction, but have also been applied to string searches (Pfeifer, Poersch & Fuhr, 1996) and other similarity measures.

Within linguistics, extensions of the Levenshtein distance have been used to measure distances between dialects. Heeringa and Nerbonne (2002) used the Levenshtein distance to measure distances between Dutch dialects and correlate these distances to geographical distances between the respective dialects. The Levenshtein distance has also been applied to Norwegian (Gooskens & Heeringa, 2004) and Irish Gaelic (Kessler, 1995) dialects. Gooskens and Heeringa (2004) compared the Levenshtein distances to the dialect speaker's perception of the similarity or dissimilarity between the respective dialects. In all these cases, the Levenshtein distance is calculated between the phonetic transcriptions (pronunciations) of words in the respective dialects.

This article proposes to use an extension of the weighted Levenshtein distance in a much broader context than simply to measure dialect distances. Comparative linguistics is concerned with the genetic relationships between languages. Similar languages are systematically compared to reconstruct the original language from which the attested languages stem. The original language is referred to as a proto-language. Usually a proto-language is not directly attested, therefore the time depth between proto-language (parent language) and descendant language are largely hypothetical, based on the estimates of linguists.

Southern African Linguistics and Applied Language Studies is co-published by NISC (Pty) Ltd and Routledge, Taylor & Francis Group

An early example of a mathematical approach to quantify language distance is Kroeber and Chrétien (1937). They use a formula for the degree of association to measure language distance. This formula is a function of number of features exhibited by both languages, number of features exhibited by neither, etc. However, this method relies on accurate, unambiguous reconstructions.

Another method to model language distance was proposed by Swadesh (1951). Swadesh postulated that languages lose their core vocabulary at a constant rate. By assuming a loss rate of 85% per millennium, it is possible to estimate time differences between related languages. This method is termed glottochronology.

An information theoretical approach to this problem has also been proposed (Raman & Patrick, 1997). It is based on the phonological evolution of languages. It models historical phonological processes by means of a Probabilistic Finite State Automaton (PFSA). The PFSA is used to model an inductive hypothesis. A Minimum Message Length (MML) principle is to find the best hypothesis that fits the available data. This method requires a detailed reconstruction of the chronology of the phonological changes that occurred from parent to descendant language.

In this article, the extended weighted Levenshtein distance is also applied to phonology as a measure of language distance. An inherent assumption is that phonology evolves at a constant rate. This assumption is tested by applying the extended weighted Levenshtein distance to the Iranian languages. The next section gives a short overview of the Iranian languages considered in this article and some relevant linguistic notation. The subsequent section defines the Levenshtein distance used to measure language distance. The section thereafter applies the extended weighted Levenshtein distance to four Iranian languages and interprets the results. Finally, the penultimate section explores further work that can be used to test this method.

Iranian languages

The Iranian language family is a sub-branch of Indo-Iranian, which in turn is a sub-branch of Indo-European. Apart from Iranian, Proto-Indo-Iranian is the parent of also the Indo-Aryan languages. On linguistic and historical grounds, a late 3rd millennium BC date is usually assigned to Proto-Indo-Iranian. Around 1400–1300 BC, the Iranian tribes began their migrations southward, occupying modern Iran from about 1300 BC (Young, 1967).

In this article, Old Iranian reconstructed forms are used. This is the standard form that is used in the literature. However, Old Iranian is not Proto-Iranian. In literature, Common Iranian is in fact Proto-Iranian. Reconstructed Old Iranian can be approximately dated to 1400 BC. Common Iranian can be dated to around 1800 BC. Note that these dates are rough approximants, with a possible error of a few hundred years. For a detailed description of this difference and of the position of Iranian within Indo-Iranian and Indo-European, refer to Sims-Williams (1998).

In this article, we consider four Iranian languages: Bactrian, Sogdian, Khotanese, which are Middle Iranian languages, and Pashto, a New Iranian language.

Bactrian was the official language of the Kushan Empire, 1st–3rd century AD. It was spoken mainly in northern Afghanistan, north of the Hindu Kush. Bactrian was written in the Greek script. The most important documents, linguistically speaking, come from after the collapse of the Kushan Empire. These are some 150 documents ranging from 4th–8th centuries AD (Sims-Williams, 2000). For Bactrian orthography and phonology, refer to Sims-Williams (1989a).

Sogdian was the language spoken in ancient Sogdia, situated within modern Uzbekistan and Tajikistan. Its importance is confirmed by the fact that Sogdian became the lingua franca of the Silk Route. Sogdian is attested in three different scripts, Sogdian, Manichean and Christian. The vast majority of the Sogdian documents come from Xinjiang province, China, dating to 7th–10th century AD. For a description of Sogdian orthography and phonology, refer to Sims-Williams (1989b).

Khotanese was spoken in the Chinese province of Xinjiang, around the oasis of Khotan. Khotanese was written in the Brāhmī script. Khotanese documents date to the 7th–10th centuries AD. For Khotanese orthography and phonology, refer to Emmerick (1989).

Pashto is a modern Iranian language spoken in southern Afghanistan. It is one of the official languages of Afghanistan.

Finally, a note on some linguistic notational conventions used in this article. An asterisk (*) before a word is used to indicate a reconstructed, unattested form. Transliterations are shown in italics. Transcriptions are placed between slanted brackets //. Individual phonemes are placed between square brackets [ ].

Extended weighted Levenshtein distance

Since the distance used in this article is but a slight modification of the weighted Levenshtein distance (WLD) proposed by Okuda *et al.* (1976), we will use their notation. Let *X* and *Y* be two strings of equal or unequal length. Then |*X*| and |*Y*| represent the lengths of the respective strings. The strings *X* and *Y* are a concatenation of symbols, where the symbols are derived from an alphabet *A*. The WLD is the minimum number of substitutions, insertions and deletions necessary to transform *X* into *Y*, where there is a weight assigned to each type of transformation. Formally, the WLD from *X* to *Y* is

$$\mathrm{WLD}(X \rightarrow Y) = \min_{i} (pk_i + qn_i + rm_i)$$

where k_i is the number of symbol substitutions, m_i the number of symbol insertions and n_i the number of symbol deletions. The values *p*, *q* and *r* are the weights of substitutions, insertions and deletions respectively. If $q = r$, then $\mathrm{WLD}(Y \rightarrow X) = \mathrm{WLD}(X \rightarrow Y)$.

The extended weighted Levenshtein distance (EWLD) used in this article is but a slight modification of WLD. For EWLD, *p* is not constant, but depends on the pair of symbols being compared. This adjustment to the WLD is due to phonological considerations. Phonologically, [p] and [b] are closer to each other than [p] and [k]. Therefore, the substitution distance should reflect this. The easiest manner to define the weight *p* for substitutions is to base it on the phonological features which define a phoneme. Thus [p] is voiced bilabial stop and [b] is voiceless bilabial stop. Of the three features, one differs, therefore $p = 1$ for the phoneme pair [p] and [b]. Table 1 shows the place and manner of articulation of various consonants. This table is taken from Odden (2005).

Place of articulation refers to the place where the tip or blade of the tongue or the lips create a constriction in the production of the consonant. Manner of articulation refers to the degree to which air flow occurs during consonant production. Manner of articulation is independent of place of articulation. The first row in the table gives the manner of articulation in order. Thus, an affricate is in between a stop and a fricative. In the table, the phonemes are mostly given in pairs. The phoneme on the left is voiced, while the voiceless counterpart is on the right. The substitution distance between phonemes is then the number of blocks in-between the phonemes, where one can only move up, down, left or right. The change from voiced to voiceless and vice versa is a distance of one. Thus EWLD([b],[β]) = 2, EWLD([t],[θ]) = 3 and EWLD([č],[ž]) = 2. From the perspective of historical phonology, nasals do not alternate with stops, affricates, etc. and so nasals are kept separate from the other manners of articulation. The only exception is [m] and [b], which is

Table 1: Phonetic features of various consonants

	Stop	Affricate	Fricative	Lateral fricative	Lateral	Nasal
Bilabial	b,p	(b^{β}),(p^{ϕ})	β,ɸ			m
Labiodental		b^{v},p^{f}	v,f			ɱ
Dental	d̪,t̪	$\text{d}^{\text{ð}}$,t^{θ}	ð,θ			n̪
Alveolar	d,t	d^{z},t^{s}	z,s	ɮ,ɬ	l	n
Alveopalatal		ǰ,č	ž,š			ñ
Retroflex	ḍ,ṭ	$\text{ḍ}^{\text{ẓ}}$,$\text{ṭ}^{\text{ṣ}}$	ẓ,ṣ		ḷ	ṇ
Palatal	ɟ,c	($\text{ɟ}^{\text{ʝ}}$),($\text{c}^{\text{ç}}$)	ʝ,ç		ʎ	ñ
Velar	g,k	g^{γ},k^{x}	ɣ,x			ŋ
Uvular	q,G	G^{γ},q^{χ}	ɣ,χ			ŋ̣
Pharyngeal			ʕ,ħ			
Laryngeal	ʔ		ɦ,h			

assigned a substitution distance of one, because the interchange between these phonemes is common in various languages.

A similar scheme can also be applied for vowels. Table 2 shows the features associated with vowels. This table is from Odden (2005). Where vowels appear in pairs, the left is unrounded, while the right is rounded. Old Iranian possessed a very simple vowel system (see Sims-Williams, 1998), therefore distinctions between tense and lax, round and unrounded vowels will not be used when determining the substitution distance. Only high, mid, low and front, central, back distinctions are considered. As with the consonants, the distance is given by the number of blocks between the respective phonemes. Thus EWLD([a],[e]) = 2 and EWLD([i],[ə]) = 2. Furthermore, the distinction between short versus long vowels and additional vowel features, such as nasalisation is given a substitution weight of 0.5.

A weight of one is assigned between syllabic and non-syllabic variants. Thus, EWLD([i],[y]) = 1, EWLD([u],[w]) and EWLD([r],[r̥]) = 1. The phoneme [w], voiced rounded labiovelar approximant often alternates with [v], voiced labiodental fricative, therefore EWLD([w],[v]) = 1. Finally, EWLD([l],[r]) = 1.[1] This covers the attested phonological changes in the Iranian languages considered in this article.

The weights assigned to insertions and deletions are equal and $q = r = 2.5$. We can argue this selection based on syncopated vowels. A short vowel is first reduced and then lost, [a] > [ə] > φ, where φ is a null string. A weight of 2.5 is selected because the maximum value of p is $2q = 2r$. This is because any substitution can be achieved by a deletion followed by an insertion. The value of 2.5 is selected to allow a maximum p to be 5. Note that the extended weighted Levenshtein distance as defined here is similar to the distance used by Heeringa and Nerbonne (2002). However, they assigned different weights to insertions and deletions and used more phoneme features to define the substitution weights.

Since Old Iranian forms are transcribed in APA (American Standard), all phonetic transcriptions in this article will follow APA. Since the acronym EWLD is somewhat cumbersome, it will be replaced by D in formulas in the rest of the article. To clarify these ideas, consider the following example. Let A_1 represent the phonological inventory of Sogdian, and A_2 of Old Iranian. Then, let $A = A_1 \cup A_2$. Consider the Sogdian word *βyjy* /βeži/ < *bazdyah. Let X = bazdyah and Y = βeži. Then $X \in A^{|X|}$ and $Y \in A^{|Y|}$. The transformation of X into Y is depicted below, where s stands for a substitution and d for a deletion:

b	a	z	d	y	a	h
β	e	ž		i		
s	*s*	*s*	*d*	*s*	*d*	*d*

Therefore, D(bazdyah, βeži) = 13.5.

The distance between languages can be quantified as follows. Let $L = (|X|+|Y|)/2$. $D(X,Y)$ is the minimum Extended Weighted Levenshtein Distance between the strings X and Y. The normalised EWLD, $|D(X,Y)|$ is defined as$|D(X,Y)| = D(X,Y)/L$. A set of pairs of etymologically related words is selected. The normalised EWLD is calculated for all pairs in the set. The distance between two languages is then the average normalised EWLD. The EWLD is calculated using an adaptation of the dynamic programming algorithm presented by Okuda *et al.* (1976).

Languages evolve their phonology at different rates at different times. It is reasonable to assume that over long periods of time, these variations even out. Is it a correct assumption that the evolution

Table 2: Phonetic features of various vowels

tense	i,ü	ɨ,ʉ	ɯ,u	high
lax	ɩ,ö̈		ʊ	
tense	e,ö	ə,ɵ	ɤ,o	mid
lax	ɛ,ɔ̈		ɔ	
	æ,œ	a	ɑ,ɒ	low
	front	central	back	

of phonology in languages takes place at a constant rate over longer periods of time? Using the ideas presented above, it is possible to test the correctness of the above assumption. The Iranian languages present cases to which to apply these tests. Four general tests can be applied:

1. The attested Middle Iranian languages were all spoken at roughly the same time, around 700 AD. They all descend from a single proto-language. If their phonology developed at the same constant rate, they should all have approximately the same distance to Old Iranian.
2. The approximate date of Old Iranian is known. Then the ratio of the time between Old Iranian and the attested Middle Iranian language and the time between Old Iranian and a modern Iranian language should be approximately equal to the ratio of the respective distances.
3. Within Middle Iranian, languages which are geographically the closest to each other show the most similarity. By comparing sibling languages of approximately the same date, do the phonological distances reveal dialectal similarities as proposed by linguists?
4. Consider, for example, the modern Iranian language Pashto. Its Middle Iranian ancestor is unknown. Of the attested Middle Iranian languages, Bactrian has been described as the closest to Pashto. Therefore, one would expect that the distance from Old Iranian to Bactrian plus the distance from Bactrian to Pashto would be greater than the distance from Old Iranian to Pashto.
5. The first two tests are applied partially in the next section.[2]

Results

In this section we test the feasibility of applying the EWLD to model phonological change by applying the first two tests outlined in the previous section. In selecting the sample vocabulary, there are certain conditions that should be satisfied:

1. Only pairs of words that are etymologically related should be selected.
2. Only inherited vocabulary should be considered. This, therefore, excludes loan words, even those from related languages.
3. The selected vocabulary should reflect a diversity of the various phonological developments in the language.
4. Words of varying length should be selected. This is because words of varying syllable lengths tend to have different developments.

Furthermore, one may suppose that the larger the sample set, the finer the accuracy obtained when calculating the distance between languages. For the first test, three Middle Iranian languages are used, Bactrian, Sogdian and Khotanese. The sample set consists of 20 randomly selected words. Table 3 shows the results for Bactrian. The words and their etymologies are obtained from Sims-Williams (2000). Unfortunately, phonetic transcriptions for Bactrian are not yet attested in literature, since Bactrian was deciphered only about a decade ago. Based on Bactrian orthography and etymological considerations, it is possible to give relatively accurate reconstructions. Refer to Sims-Williams (1989a) for Bactrian orthography and historical phonology. Table 4 shows the results for Sogdian. Sogdian etymologies are obtained from Sims-Williams (1989b, 2000) and transcriptions from Skjærvø (2007).

Table 5 shows the results for Khotanese. Khotanese etymologies and transcriptions are obtained from Emmerick (1989). The second test is applied by using the modern Iranian language Pashto. Table 6 shows the results for Pashto. Pashto etymologies are obtained from Skjærvø (1989).

For the first test, the following are the results: |D(Bac.,OIr.)| = 1.46, |D(Sog.,OIr.)| = 1.55 and |D(Khot.,OIr.)| = 1.59. As expected, the obtained results are approximately equal. To expect perfectly equal distances would be erroneous. There are many factors which would prevent this, which may aptly be termed noise:

1. It would be difficult to believe that phonological evolution takes place at a perfectly constant rate through all time. One may rather expect that the rate sometimes increases and decreases, but that over many centuries, these variations average out.
2. The attested Middle Iranian languages cannot all be assigned the same date. They are all attested within a certain time frame of a couple of hundred years.
3. The sample data set used is extremely small. Larger sets would be expected to result in a more accurate distance.

Table 3: Bactrian sample set

Bactrian < Old Iranian	**D**	***L***	**|D|**
1. *αβζαο-* /əβd^{z}əw/ < *abiǰawa	9	6	1.50
2. *αβιστaοoαγo* /əbistāwəg/ < *apastāwākā	10.5	9.5	1.11
3. *αγαλγo* /āγālg/ < *āgādaka	12	6	2.00
4. *αζo* /az/ < *azam	5	3	1.67
5. *βαγo* /βəγ/ < *baga	6.5	3.5	1.86
6. *βανζo* /βənd^{z}/ < *bandačī	10.5	5.5	1.91
7. *γαo* /γāw/ < *gāw	2	3	0.67
8. *ζαμιγo* /zəmīg/ < *zamīkā	3.5	5.5	0.64
9. *ιωγo* /yōg/ < *aiwaka	12	4.5	2.67
10. *καμιρδo* /kəmird/ < *kamr̥da	6	6	1.00
11. *λαδo* /lād/ < *dāta	7.5	3.5	2.14
12. *μιλανo* /milān/ < *madyānā	11	6	1.83
13. *μιυγαρo* /mihgār/ < *miθahkāra	8.5	7.5	1.13
14. *ναμαγo* /nāməg/ < *nāmaka	4.5	5.5	0.82
15. *νιþαλμo* /nəšalm/ < *nišadman	11	7	1.57
16. *οαρσoχoανδo* /wərt^{s}əxwənd/ < *warčahwant	7	10	0.70
17. *οαχo* /wəx/ < *waxša	5	4	1.25
18. *οιγαλφo* /wigālf/ < *wikāθwan	11	7	1.57
19. *πιδoοασ-* /pidwāt^{s}/ < *patiwāča	10	7	1.43
20. *χοαδo* /xwəd/ < *hwatah	9	5	1.80
Average		5.75	1.46

Table 4: Sogdian sample set

Sogdian < Old Iranian	**D**	***L***	**|D|**
1. *zng* /zəng/ < *zanaka	7	5	1.40
2. *pnj* /panǰ/ < *panča	3.5	4.5	0.78
3. *’zw* /əzu/ < *azam	6.5	3.5	1.86
4. *z’ṭyy* /zātē/ < *zātakah	10	5.5	1.82
5. *βγy* /βəγi/ < *bagah	10.5	4.5	2.33
6. *fnyš-* /fnēš/ < *franasya	13.5	6	2.25
7. *jmnw* /žəmnu/ < *ǰamanam	10	6	1.67
8. *zyrn* /zern/ < *zaranya	9.5	5.5	1.73
9. *zrync* /zərēnǰ/ < *uzrinčaya	15	7.5	2.00
10. *γzn* /γəzn/ < *gazna	5.5	4.5	1.22
11. *”m’ţyy* /āmātē/ < *āmātakah	10	6.5	1.54
12. *‘wst’t* /ōstāt/ < *awastāta	10	6.5	1.54
13. *βrt* /βart/ < *barati	7	5	1.40
14. *ps’k* /psāk/ < *pusākā	5	5	1.00
15. *βyjy* /βeži/ < *bazdyah	13.5	5.5	2.45
16. *w’crn* /wāčərən/ < *wahāčarana	9.5	8.5	1.12
17. *šyr* /šir/ < *srīra	6.5	4	1.63
18. *mrtxmy* /mərtəxmē/ < *martatauxmaka	17	10.5	1.62
19. *γ’δwk* /γāθuk/ < *gāθuka	4.5	5.5	0.82
20. *stryc* /strīč/ < *strīčīā	5	6	0.83
Average		5.78	1.55

4. The attested Middle Iranian languages all descendent directly from Common Iranian and not (strictly speaking) Old Iranian. However, in literature, Old Iranian reconstructions are given, not Common Iranian.[3]

For Pashto, the following distance is obtained: |D(Pash.,OIr.)| = 2.31. The approximate time span from Old Iranian to the considered Middle Iranian languages is 2 100 years, and from Old Iranian

Table 5: Khotanese sample set

Khotanese < Old Iranian	**D**	***L***	**|D|**
1. *biśśa* /βiša/ < *vispa	4.5	4.5	1.00
2. *hambūva* /hãbūwa/ < *hampūta	9	6.5	1.38
3. *bera* /βεra/ < *bārya	7	4.5	1.56
4. *ttāra* /tāra/ < *tanθra	5.5	5	1.10
5. *ṣvīda-* /šwīda/ < *xšvifta	7.5	6	1.25
6. *patält-* /padẹlʸd/ < *patikr̥ta	10.5	7	1.50
7. *birgga* /βirga/ < *vr̥ka	5.5	4.5	1.22
8. *yäda-* /yẹda/ < *kr̥ta	11	4	2.75
9. *hor-* /hor/ < *frabara	17	5	3.40
10. *jasta* /ǰasta/ < *yazata	8.5	5.5	1.55
11. *ttuvāy-* /tuwāy/ < *ativādaya	13	7	1.86
12. *bihan-* /βihan/ < *vixanda	9	6	1.50
13. *ysāra* /zāra/ < *hazahra	8	5.5	1.45
14. *kṣundaa* /tṣundaa/ < *fšuyantaka	13.5	8.5	1.59
15. *hambruīttä* /hãbruītẹ/ < *hamraudati	11	9	1.22
16. *būnaa-* /βūnaa/ < *bagnaka	10.5	6	1.75
17. *ṣṣavā* /ṣawā/ < *xšapā	8.5	4.5	1.89
18. *nyūs-* /nyūs/ < *niyuxsa	8	5.5	1.45
19. *āṣṣeiṇa-* /āṣεiṇa/ < *axšaina	8	6.5	1.23
20. *pahaiga* /pahaiγa/ < *apahaxta	8.5	7.5	1.13
Average		5.93	1.59

Table 6: Pashto sample set

Pashto < Old Iranian	**D**	***L***	**|D|**
1. ās < *aspah	8	3.5	2.29
2. plār < *pitarah	12	5.5	2.18
3. špaẓ̌ < *xšwašam	14.5	5.5	2.64
4. dyārlas < *θrayahdasa	21	8.5	2.47
5. zoy < *zahakah	16	5	3.20
6. zṛə < *zr̥dayah	13	5	2.60
7. γwā < *gawā	4.5	3.5	1.29
8. žay < *īziyakah	16.5	5.5	3.00
9. sra < *suxrā	5.5	4	1.38
10. šna < *axšainā	10.5	5	2.10
11. xpəl < *xwaipaθiyah	21.5	7.5	2.87
12. rwaj < *raučah	11	5	2.20
13. čina < *kaniyā	12.5	5	2.50
14. āxšay < *āxwasrukah	18.5	7.5	2.47
15. lwaš- < *dauxšaya	19	6	3.17
16. žwand < *ǰīwantah	9.5	6.5	1.46
17. war < *dwara	5	4	1.25
18. bən < *hapaθniy	14.5	5.5	2.64
19. xob < *xwāpah	10.5	4.5	2.33
20. psarlay < *upasaradakah	20.5	9.5	2.16
Average		5.6	2.31

to Pashto, 3 400 years. Then, 3 400/2 100 = 1.62 and 2.31/1.53 = 1.51, where 1.53 is the average distance for the three Middle Iranian languages. Therefore, the ratio of the time gap in years closely corresponds to the ratio of EWLD for the respective languages. Keeping in mind the presence of noise, this is the expected result.

Although the above two tests are applied to a limited data set, they clearly portray the potential of applying the extended weighted Levenshtein distance to model the time difference between related languages.

Future work

Within the Iranian language family, there are more Middle Iranian languages to which the first test can be applied. In particular, Manichean Middle Persian and Manichean Parthian are attested from 8th century AD. The application of the third and fourth test as outlined in the third section would be interesting.

Conclusion

It is proposed that language distances may be modelled using an extension of the weighted Levenshtein distance. This metric is applied to pairs of etymologically related words from the languages under consideration. An inherent assumption is that the phonology of languages evolves at an approximately constant rate. This assumption is tested by using data from the Iranian language family.

It is possible that an adjustment of EWLD or of the weights assigned to various types of transformations may result in better distance measures. The selection of the weights is to a degree a subjective matter. Nevertheless, the results using the Iranian language family does indicate the potential of this approach. It is important to note that the calculation of EWLD is independent of the historical phonological processes. Only by applying this method to a larger variety of languages will it be possible to fully evaluate the validity of using the EWLD to model language distance.

Notes

[1] This distance is selected because this phoneme interchange is very common in various Iranian languages. For other language groups, a different value may be more appropriate.

[2] The other two tests will receive attention in future research.

[3] It would have been preferable to use Common Iranian reconstructions. However, these are rare, with the standard reconstructions being Old Iranian. One can consider Common Iranian as Early Proto-Iranian and Old Iranian as Late Proto-Iranian, therefore, the difference is small. Refer to Sims-Williams (1998) for a description of this difference.

References

Damerau F. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM* **7**: 171–176.

Emmerick RE. 1989. Khotanese and Tumshuqese. In Schmitt R (ed.) *Compendium Linguarum Iranicarum*. Weisbaden: Dr. Ludwig Reichert Verlag, pp 204–229.

Gooskens C & Heeringa W. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect area. *Language Variation and Change* **16**: 189–207.

Heeringa W & Nerbonne J. 2002. Dialect areas and dialect continua. *Language Variation and Change* **13**: 375–400.

Kessler B. 1995. Computational dialectology in Irish Gaelic. Proceedings of the European ACL 1995, Dublin, Ireland, pp 60–67.

Kroeber AL & Chrétien CD. 1937. Quantitative classification of Indo-European languages. *Language* **13**: 83–103.

Kruskal JB. 1983. An overview of sequence comparison: time warps, string edits, and macromolecules. *SIAM Review* **25**(2): 201–237.

Levenshtein VI. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademmi Nauk SSSR* **162**(4): 845–848.

Odden D. 2005. *Introducing phonology*. Cambridge: Cambridge University Press.

Okuda T, Tanaka E & Kasai T. 1976. A method for the correction of garbled words based on the Levenshtein metric. *IEEE Transactions on Computers* **25**(2): 172–178.

Oommen BJ & Loke RKS. 1997. Pattern recognition of strings with substitutions, insertions, deletions and generalized transpositions. *Pattern Recognition* **30**(5): 789–800.

Pfeifer U, Poersch T & Fuhr N. 1996. Retrieval effectiveness of proper name search methods. *Information Processing & Management* **32**(6): 667–679.

Raman A & Patrick J. 1997. Linguistic similarity measures using the minimum message length principle. In Blench R & Spriggs M (eds) *Archaeology and language I: theoretical and methodological orientations.* New York: Routledge, pp 262–279.

Sims-Williams N. 1989a. Bactrian. In Schmitt R (ed.) *Compendium Linguarum Iranicarum*. Weisbaden: Dr. Ludwig Reichert Verlag, pp 230–235.

Sims-Williams N. 1989b. Sogdian. In Schmitt R (ed.) *Compendium Linguarum Iranicarum*. Weisbaden: Dr. Ludwig Reichert Verlag, pp 173–192.

Sims-Williams N. 1998. The Iranian languages. In Ramat AG and Ramat P (eds) *The Indo-European Languages, Routledge Language Family Series.* New York: Routledge, pp 125–152.

Sims-Williams N. 2000. *Bactrian Documents from Northern Afghanistan I: legal and economic documents*. Oxford: The Nour Foundation.

Skjærvø PO. 1989. Pashto. In Schmitt R (ed.) *Compendium Linguarum Iranicarum.* Weisbaden: Dr. Ludwig Reichert Verlag, pp 384–410.

Skjærvø PO. 2007. An introduction to Manichean Sogdian. Available at: http://www.fas.harvard.edu/~iranian/ [accessed 29 October 2009].

Swadesh M. 1951. Diffusional cumulation and archaic residue as historical explanations. *South-western Journal of Anthropology* **7**: 1–21.

Young Jr. TC. 1967. The Iranian migration into the Zagros. *Iran* **5**: 11–34.