Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

ROBUSTNESS AND STATISTICAL SIGNIFICANCE OF PAM-LIKE MATRICES FOR COGNATE IDENTIFICATION

Antonella Delmestri and Nello Cristianini

September 2010

Technical Report # DISI-10-048

Robustness and Statistical Significance of PAM-like Matrices for Cognate Identification

Antonella Delmestri[1] and Nello Cristianini[2]

1. Department of Information Engineering and Computer Science, University of Trento, Italy

2. Intelligent Systems Laboratory, University of Bristol, U.K.

Abstract: This paper tests the influence of the training dataset dimension on a recently proposed orthographic learning system, inspired from biological sequence analysis and successfully applied to cognate identification. This system automatically aligns a given set of cognate pairs producing a meaningful training dataset, learns from it substitution parameters using a PAM-like technique and utilises them to recognise cognate pairs. The results show no difference in the performance when training the system with about 650 cognate pairs extracted from 6 Indo-European languages or with about 62,000 cognate pairs extracted from 76 Indo-European languages. In both cases the system outperforms all comparable orthographic and phonetic methods previously proposed in the literature. This paper also investigates the statistical significance of these results when compared with earlier proposals. The outcome confirms that the performance reached by this system with both training datasets is significantly higher than the one achieved by all the other methods. Indeed, the training dataset dimension seems not to influence either the accuracy or the statistical significance of this learning system that needs only a very small amount of data to reach an outstanding performance.

Key words: Cognate identification, substitution matrices, string similarity measures.

1. Introduction

Cognates are words that derive from a common ancestor and share the same etymological origin. Cognate identification represents one of the most promising applications of computational techniques to historical linguistics. The synergy between cognate identification and phylogenetic inference, which uses cognateness information to identify genetic relationships between natural languages, may allow the tracing of language evolution and the investigation of the origin of language. Successful applications of cognate recognition to computational linguistics include dialectology [19,38,49,42] and language reconstruction [6,5,22,39,23]. However, computational linguistics is not the only field where cognate identification has been successfully employed. In fact it has been beneficially applied to many different areas of natural language processing including semantic word clustering [1], bilingual lexicography [3,18], machine translation [17,26], lexicon induction [32,21,43,36], parallel corpora sentence alignment [44,4,34,35], parallel corpora word alignment [47,24], cross-lingual information retrieval [41] and confusable drug name detection [25].

Because of the close analogy between language evolution and species evolution [7,2], evolutionary biology and historical linguistics have evolved in exceptionally similar ways. In automatic cognate identification as in computational molecular analysis, strings may be studied successfully by inexact string matching techniques that allow their similarity to be measured and their optimal alignment to be found through dynamic programming [16]. Pairwise alignment algorithms generally consist of a two-step procedure that starts calculating the maximum similarity score between two strings and ends tracing back their optimal alignment. If the algorithm is focussed on optimally aligning the complete strings, it produces a *global alignment* [37,15], while if it is targeted on any substring alignment that can reach the highest score, it performs a *local alignment* [45,15]. Any alignment algorithm that calculates the maximum similarity score between two strings utilises a scoring scheme that greatly influences the significance of the alignments it produces. In bioinformatics, where strings are generally addressed as sequences, substitution matrices are scoring schemes extensively used to analyse protein or nucleic acid sequences [12].

Given an alphabet $\mathcal{A}$ with $|\mathcal{A}| \geq 2$, a *substitution matrix* $|\mathcal{A}|$-by-$|\mathcal{A}|$ over $\mathcal{A}$ represents the scores at which each character of $\mathcal{A}$ may transform into another character of $\mathcal{A}$. The substitution matrix alphabet must reflect the application involved. Substitution matrices can be constructed by collecting a large sample of verified pairwise alignments, or multiple sequence alignments, and deriving the scores using a probabilistic model. Ideally, the scores in the matrix should be proportional to the true probabilities of mutations occurring through a period of evolution that the alignments try to represent [16]. On the other hand, the score of the alignments should give a measure of the relative likelihood that the sequences are related as opposed to being unrelated, which is generally achieved through a *log-odds-ratio* [12].

In cognate identification, the alphabet may be *orthographic*, when cognates are employed in their orthographic format, or *phonetic*, when cognates are represented in a phonetic notation. The former does

not require any phonetic transcription, relies on the traces that sound changes leave in the orthography and assumes that alphabetic character correspondences represent in some way sound correspondences. The latter depends on phonetic transcriptions of texts, whose attainment is still a very time consuming and demanding task, but benefits from the phonetic features present in phonemes. For this reason a phonetic approach is supposed to be more accurate than an orthographic one in the task of cognate identification, but several recent results seem to prove the opposite [30,27,11]. In this paper we investigate the robustness and the statistical significance of a learning system for cognate identification proposed by Delmestri and Cristianini [11] that includes a substitution matrix generator based on a scoring scheme very successfully used for biological sequence analysis, the Point Accepted Mutation (PAM) matrices [8,9,10].

The rest of the paper is organised as follows. Section 2 describes the learning system, section 3 explains the experimental design and section 4 includes the results of our investigation, discusses them and compares them with others previously proposed in the literature. Finally, section 5 reports the conclusions drawn by this study.

2. The Learning System

The learning system proposed by Delmestri and Cristianini [11] has been inspired by biological sequence analysis, which has been historically supported by golden standard substitution matrices in the discovery of sequence similarities. The system utilises orthographic data extracted from Indo-European Swadesh lists based on the Latin alphabet, but it may easily be adapted to any alphabetic system, including the phonetic alphabet, if data were available. The architectural design consists of three main modules.

The first component is a global pairwise aligner [37] that aligns sensibly cognate pairs and prepares a meaningful training dataset, guided by a symmetric, linguistic-inspired substitution matrix shown in Table 1. For readability, only the lower triangular matrix is filled in.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
A	2																									
B	-3	2																								
C	-3	-3	2																							
D	-3	0	-3	2																						
E	0	-3	-3	-3	2																					
F	-3	0	0	0	-3	2																				
G	-3	0	0	0	-3	-3	2																			
H	-3	-3	0	0	-3	0	-3	2																		
I	0	-3	-3	-3	0	-3	-3	0	2																	
J	0	-3	-3	0	0	-3	0	0	0	2																
K	-3	-3	0	-3	-3	-3	0	0	0	0	2															
L	-3	-3	0	0	-3	0	0	0	0	0	-3	2														
M	-3	0	-3	0	-3	-3	-3	-3	-3	0	-3	-3	2													
N	-3	-3	0	0	-3	-3	0	-3	-3	-3	0	0	0	2												
O	0	-3	-3	-3	0	-3	-3	-3	0	0	-3	-3	-3	-3	2											
P	-3	0	0	-3	-3	0	-3	-3	-3	-3	0	0	0	-3	-3	2										
Q	-3	-3	0	-3	-3	0	0	0	-3	-3	0	-3	-3	-3	-3	0	2									
R	-3	-3	-3	0	-3	-3	-3	0	-3	-3	0	0	0	0	-3	-3	-3	2								
S	-3	0	0	0	-3	-3	0	0	-3	0	0	0	-3	-3	-3	-3	-3	0	2							
T	-3	0	0	0	-3	-3	-3	0	-3	-3	0	0	0	0	-3	0	-3	0	0	2						
U	0	-3	-3	-3	0	-3	-3	0	0	0	-3	0	-3	-3	0	-3	-3	-3	-3	-3	2					
V	-3	0	0	-3	-3	0	0	0	-3	-3	0	-3	0	-3	-3	0	0	-3	-3	-3	0	2				
W	0	0	-3	-3	0	-3	-3	-3	0	0	-3	-3	-3	-3	0	-3	0	-3	-3	-3	0	0	2			
X	-3	-3	0	-3	-3	-3	0	0	-3	0	0	-3	-3	-3	-3	-3	-3	-3	0	-3	-3	-3	-3	2		
Y	0	-3	-3	-3	0	-3	-3	0	0	0	-3	0	-3	-3	0	-3	-3	-3	-3	-3	0	-3	-3	-3	2	
Z	-3	-3	0	0	-3	-3	0	-3	-3	0	-3	-3	-3	0	-3	-3	-3	0	0	0	-3	-3	-3	0	-3	2

Table 1. The linguistic-inspired substitution matrix

This 26-by-26 matrix aims to represent the a-priori likelihood of transformation between each character of the alphabet into another and tries to code well known systematic sound changes left in written Indo-European languages, including Grimm's and Verner's Law, Centum-Satem division, rhotacism, assimilation, dissimilation, lenition, fortition and L-vocalisation. A value of 2 is given to all the elements of the main diagonal that represent no change, 0 values to all the character transformations considered "possible", a value of -3 to all the character transformations considered "impossible". A gap penalty of -1 is applied for insertion and deletion (*indels*). If the aligner finds more than one optimal alignment with the same rate, it chooses one of them through an alternate trace back (↖←↑ | ↖↑← | ←↖↑ | ←↑↖ | ↑←↖ | ↑↖←) in an attempt to eliminate possible bias caused by always giving priority to the same conditional predicates and therefore assuring a more balanced learning process.

The second component of the learning system is a generator of symmetric PAM-like substitution matrices that uses a technique similar to the PAM method developed by Margaret Dayhoff [8,9,10] and widely used for amino acid sequence analysis. The PAM approach aims to learn substitution parameters from global alignments between closely related sequences and then to extrapolate from this data longer evolutionary divergences. A family of ten PAM-like matrices appears to be adequate for studying the divergence time of the languages considered.

The third component of this system is a cognate identifier that benefits from the PAM-like matrices and from a family of parameterised string similarity measures [11] to rate the cognate pairs. The similarity measures derive from different normalisations of a generic scoring algorithm and take into account the similarity of each string with itself with the aim of eliminating, or at least reducing, the bias due to different string length. For example, given two strings, S_1 and S_2, and a generic similarity rating algorithm AL, the similarity measure $sim_1(S_1, S_2, AL)$ normalises the rate of the algorithm AL by the arithmetic mean of the similarity rates calculated by AL applied to each string with itself. Possible scoring algorithms include the Needleman-Wunsch algorithm for global alignment [37,15] and the Smith-Waterman for local alignment [45,15].

$$sim_1(S_1, S_2, AL) = \frac{2 * AL(S_1, S_2)}{AL(S_1, S_1) + AL(S_2, S_2)}$$

3. Experimental Design

It has already been proved [11] that the learning system described in section 2 is successful in cognate identification. In modelling our experiments in this paper, our aim has been to study the influence of the training dataset dimension on the performance of the system in terms of accuracy. Moreover, we have been focussed on analysing its statistical significance as the dataset dimension varies. In doing so, we have trained with a 76 Indo-European languages dataset, two families of PAM-like matrices, one based on the Roman alphabet and one based on its extension with gap. The gap symbolises the null character for indels and is represented by the symbol '-'. We have called these learning models DAY76 and DAY76b respectively.

3.1 Data

The *training dataset* has been sourced from the *Comparative Indo-European Database* by Dyen et al. [13] which is in orthographic format without diacritics and uses the 26 letters of the Roman alphabet. This corpus consists of 200-word Swadesh lists [46] of basic, culture independent and time resistant meanings from eighty-four current Indo-European languages. The words are classified by meaning and by certain or uncertain cognateness. We have extracted from this corpus all the speech varieties that did not overlap with the test dataset and in doing so we have excluded English, German, French and the five varieties of Albanian, for a total of 76 Indo-European languages. Among these languages we have considered only the word pairs reported by Dyen et al. [13] as certain cognates and only the first cognate pair when more words were provided for the same meaning in the same language, for a total of about 62,000 cognate pairs. A few apparent mistakes have been amended. We have then automatically aligned these word pairs as described in section 2.

The *test dataset* has been built from the orthographic form of the 200-word Swadesh lists of English, German, French, Latin and Albanian provided by Kessler [20] together with his cognateness information.

The training dataset and the test dataset did not present any overlap in their language sets.

3.2 Evaluation Methodology

In order to assess our cognate identification system, we have used ten language pairs built from the combination of the five languages forming the test dataset. For each language pair and for all the word pairs with the same meaning in two languages, we have evaluated the likelihood that the two words were cognates by calculating a score. The scores of each language pair have then been sorted and, when more word pairs have showed the same rate, the alphabetic order has been considered as well to avoid random results.

To evaluate the capacity of our string similarity system in the task of cognate identification, we have utilised an evaluation metric called *11-point interpolated average precision* [31]. Originally built for ranking computation in the field of Information Retrieval, this measure has the benefit of not using a threshold which may be arbitrary or application specific. This measure has also been frequently used in the field of cognate recognition by other systems with which we wanted to be properly comparable [30,27]. We have engaged the families of PAM-like matrices in the alignment and rating process of the test dataset using the standard sequence alignment algorithms [37,45,15] and the family of parameterised similarity measures proposed [11]. We have applied a unary gap penalty in the alignment algorithms for DAY76 which is based on the Latin alphabet without gap.

4. Results

In order to evaluate the performance of DAY76 and DAY76b trained with about 62,000 cognate pairs from 76 Indo-European languages, we have computed the 11-point interpolated average precision for each of the ten language pairs of our test dataset with each PAM-like matrix and with each similarity measure described in section 2. We have then calculated their average, standard deviation, variance and median.

Both the two models DAY76 and DAY76b have achieved very good performance. DAY76b, which utilises the Latin alphabet extended with gap, has performed better than DAY76 and has produced equal top rating results using either global or local alignment. Table 2 reports PAM4 generated by DAY76b as an example and, for readability, only the lower triangular matrix is filled in. It is worth noting that this matrix contains substitution parameters that reproduce linguistics sound changes left in the written orthography including vowel changes, Grimm's Law (e.g. B→P, P→F; D→T, T→S/Z; G→C/K, C/K→H), Verner's Law (e.g. F→V, H→G, S→Z), Centum-Satem division (e.g. K→C/Q/S), rhotacism (e.g. D/L/S→R), lenition (e.g. F/S/X→H), fortition (e.g. J→G) and dissimilation (e.g. R→D/L, Q→C).

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**	**-**
A	2.38																										
B	-3.31	10.89																									
C	-2.78	1.91	3.73																								
D	-3.49	1.06	0.23	7.05																							
E	1.97	-2.94	-2.39	-3.06	1.94																						
F	-3.26	7.1	4.17	-0.63	-2.89	8.13																					
G	-1.65	1.18	1.92	0.44	-1.31	0.43	5.87																				
H	-0.68	-0.62	0.99	-0.25	-0.43	0.3	0.53	1.48																			
I	1.72	-2.92	-2.12	-3.04	1.77	-2.75	-1.18	-0.13	1.99																		
J	0.6	-1.61	-0.44	-0.37	0.69	-1.55	1.1	0.35	0.73	0.86																	
K	-2.51	-0.63	3.72	-1.08	-2.07	1.62	3.05	1.61	-1.45	0.29	6.03																
L	-3.08	-3.38	-1.09	-1.03	-2.71	-1.98	-0.09	-0.25	-1.66	-1.07	-2.15	10.13															
M	-3.8	-1.6	-1.97	-1.56	-3.38	-2.03	-1.62	-1.84	-3.44	-1.12	-2.56	-3.29	9.63														
N	-2.84	-2.87	-0.67	-0.39	-2.48	-2.54	-0.23	-1	-2.48	-1.18	-0.76	-0.33	3.82	5.62													
O	2.24	-3.27	-2.84	-3.58	1.91	-3.28	-1.66	-0.65	1.72	0.67	-2.56	-2.98	-3.87	-2.95	2.5												
P	-4.74	7.15	4.99	-2.02	-4.28	8.83	-0.11	-0.45	-4.12	-2.58	2.13	-1.55	-1.49	-3.23	-4.82	10.6											
Q	-2.18	2.52	3.91	-1.14	-1.86	4.87	1.91	1.11	-1.57	-0.3	4.14	-1.81	-2.13	-1.24	-2.21	5.54	4.4										
R	-3.58	-3.7	-1.89	0.06	-3.19	-3.45	-1.75	-0.13	-3.08	-1.82	-1.56	2.18	-1.28	-0.38	-3.69	-4.58	-2.24	7.27									
S	-1.98	-1.45	1.61	0.27	-1.63	-0.69	0.74	1.7	-1.47	0.16	2.05	-1	-2.56	-1.3	-2.05	-1.38	1.01	0.2	4.61								
T	-3.08	-0.15	2.01	3.95	-2.69	-0.04	-0.12	0.35	-2.6	-0.89	0.97	-0.55	-0.98	0.15	-3.18	0.06	0.85	-0.06	0.98	4.49							
U	1.58	-1.65	-1.81	-2.76	1.43	-1.91	-0.86	-0.07	1.38	0.61	-1.64	-0.47	-2.83	-2.24	1.75	-3.18	-1.3	-2.74	-1.31	-2.36	1.58						
V	-0.94	4.74	1.32	-1.55	-0.75	3.41	1.66	0.49	-0.66	-0.03	0.87	-2.13	1.26	-1.16	-0.75	3.2	1.89	-2.41	-0.5	-1.02	0.44	4.14					
W	0.81	2.03	-0.56	-2.01	0.8	0.49	0.22	0.14	0.78	0.41	-0.69	-2.11	-1.26	-1.64	0.94	-0.44	-0.02	-2.38	-0.86	-1.69	1.13	2.45	1.89				
X	-1.85	-1.17	2.42	-0.65	-1.48	0.45	2.24	2.05	-1.12	0.45	3.74	-1.49	-2.61	-1.23	-1.9	0.05	2.44	-0.99	3.6	0.54	-1.03	0.35	-0.52	4.3			
Y	1.68	-2.82	-2.06	-2.78	1.69	-2.62	-1.09	-0.08	1.81	0.73	-1.63	-1.67	-3.19	-2.28	1.73	-3.98	-1.57	-2.85	-1.31	-2.4	1.38	-0.65	0.74	-1	1.73		
Z	-1.73	-1.04	0.95	2.76	-1.4	-0.87	2.3	0.5	-1.38	0.81	0.61	-1.46	-1.46	0.3	-1.77	-1.69	0.27	-0.22	2.09	1.48	-1.23	-0.42	-0.74	1.6	-1.2	3.97	
-	0.11	-0.36	-0.06	-0.26	0.13	-0.42	0.21	0.11	0.15	0.11	0.08	-0.61	-0.48	-0.1	0.12	-0.78	-0.09	-0.41	0.24	-0.15	0.12	0.09	0.12	0.3	0.13	0.13	0.13

Table 2. PAM4 generated by DAY76b

Fig. 1 and Fig. 2 show the averaged 11-point interpolated average precision achieved over the ten language pairs of our test dataset using the first ten PAM-like matrices of DAY76b. For completeness, a PAM0 matrix, which represents 0 evolutionary distances and coincides with the identity matrix, has been included. This outcome has been reached using the family of similarity measures introduced [11] based respectively on the Needleman-Wunsch algorithm (NW) and on the Smith-Waterman algorithm (SW). All the similarity measures proposed have performed consistently well and sim_1, sim_3, sim_5 and sim_6 have shown to achieve the better accuracy. Among the PAM-like matrices, PAM4, PAM5 and PAM6 all have proved to be able to represent the appropriate evolutionary divergence for the test dataset.

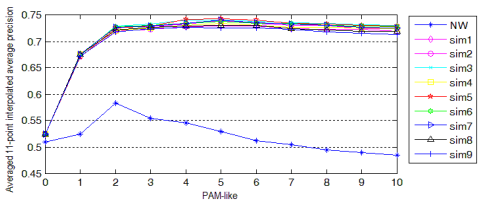

Fig. 1. Averaged 11-point interpolated average precision for DAY76b using NW

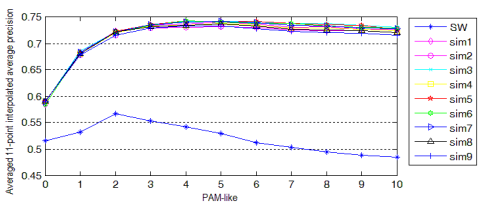

Fig. 2. Averaged 11-point interpolated average precision for DAY76b using SW

4.1 Related Works

Kondrak [22] developed ALINE, a phonetic sequence aligner specifically designed for cognate recognition and, based on linguistic knowledge, he represented phonemes as vectors of multi-valued phonetic features. His algorithm calculated the similarity between phonetically transcribed words through the sum of the similarity scores of their phonemes which were optimally aligned by a dynamic programming procedure.

Mackay [29] presented a suite of Pair Hidden Markov Models and training algorithms for cognate identification based on an automata originally proposed in [12] on the task of biological sequence analysis. The training dataset was extracted from the *Comparative Indo-European Database* by Dyen et al. [13] and consisted of about 120,000 word pairs in orthographic format. This system was tested on the dataset provided by Kessler [20] and, due to the phonetic transcriptions provided, compared by *Mackay and Kondrak* [30] with ALINE [22] and other methods formerly introduced in the literature. The results showed that all the Pair Hidden Markov Models outperformed the other methods in terms of averaged 11-point interpolated average precision. Hereinafter we shall call the one that performed better, simply, PHMM.

Kondrak and Sherif [27] for cognate identification proposed several models of a Dynamic Bayesian Net previously introduced for computing word similarity on the task of pronunciation classification [14]. About 180,000 word pairs in orthographic format were extracted from the *Comparative Indo-European Database* by Dyen et al. [13] and used twice to train the system and enforce the scoring symmetry. A set of other phonetic and orthographic algorithms, including ALINE [22] and PHMM, were assessed and tested on the dataset proposed by Kessler [20]. One of the Dynamic Bayesian Nets, which we shall refer to as DBN, reached only a slightly better averaged 11-point interpolated average precision than PHMM.

Delmestri and Cristianini [11] presented the learning system described in section 2 on the task of cognate recognition. This system was trained with about 650 cognate pairs belonging to 6 Indo-European languages extracted from the *Comparative Indo-European Database* by Dyen et al. [13] in orthographic format and automatically aligned using the linguistic-inspired substitution matrix shown in Table 1. Two models, named DAY6 and DAY6b, were developed based respectively on the Latin alphabet and on its extension with gap. The lists proposed by Kessler [20] were used as a test dataset. The results showed that DAY6 and DAY6b outperformed all the phonetic and orthographic comparable methods previously proposed in the literature increasing the averaged 11-point interpolated average precision by about 5%.

4.2 Comparison

In the task of cognate identification, the two models proposed in this paper, DAY76 and DAY76b, produce very similar results when compared with DAY6 and DAY6b [11]. This is a particularly notable outcome because of the big difference in the training dataset dimension between the two model groups. In fact DAY6 and DAY6b have been trained with only about 650 sensibly aligned cognate pairs, extracted from Italian, Portuguese, Spanish, Dutch, Danish and Swedish. DAY76 and DAY76b have been trained with approximately 62,000 sensibly aligned cognate pairs extracted from 76 very diverse Indo-European speech varieties that include the 6 languages used to train DAY6 and DAY6b.

	DAY6 NW	DAY6 SW	DAY6b NW	DAY6b SW	DAY76 NW	DAY76 SW	DAY76b NW	DAY76b SW
Average	0.729	0.735	0.743	0.749	0.729	0.740	0.743	0.743
Standard deviation	0.159	0.158	0.149	0.144	0.161	0.161	0.143	0.146
Variance	0.025	0.025	0.022	0.021	0.026	0.026	0.020	0.021
Median	0.752	0.768	0.783	0.780	0.756	0.786	0.770	0.770

Table 3. 11-point interpolated average precision of DAY6, DAY6b, DAY76 and DAY76b

Table 3 reports the performance produced by DAY6, DAY6b, DAY76 and DAY76b. It is worth noting that the average, standard deviation, variance and median of the 11-point interpolated average precision scores are impressively stable. The four models show a similar behaviour in relation to the alphabet and the alignment algorithm employed. In fact DAY6 and DAY76, that utilise the Latin alphabet, behave very similarly to each other when using respectively the Needleman-Wunsch algorithm and the Smith-Waterman algorithm. The same happens to DAY6b and DAY76b that use the Latin alphabet extended with gap.

Indeed, this outcome would suggest that when using PAM-like matrices with the family of similarity measures proposed [11], the dimension of the training dataset is not particularly influential on the system performance in terms of accuracy, if the cognate word pairs are sensibly aligned.

Table 4 shows a comparison of the top comparable phonetic and orthographic methods reported in the literature with the best results achieved by DAY6b and DAY76b using NW and SW. The results produced by NEDIT, the edit distance [28,48] normalised by the length of the longer word, are also shown and used as a baseline. The 11-point interpolated average precision achieved by ALINE [22], PHMM [30] and DBN [27] is reported as in the literature. The outcome shows that the learning system described in section 2 consistently outperforms phonetic static algorithms like ALINE [22] and orthographic learning models like PHMM [30] and DBN [27].

Indeed, DAY6b and DAY76b produce an averaged 11-point interpolated average precision approximately 5% higher than PHMM and DBN, 18% higher than ALINE and 28% higher than NEDIT.

It is worth noting that not only the average of the 11-point interpolated average precision scores is higher, but also their standard deviation and variance are much lower, suggesting that this learning system is also more stable in its performance across various language pairs.

Languages		NEDIT	ALINE	PHMM	DBN	DAY6b NW	DAY6b SW	DAY76b NW	DAY76b SW
English	German	0.907	0.912	0.930	0.927	0.929	0.934	0.933	0.935
French	Latin	0.921	0.862	0.934	0.923	0.921	0.924	0.914	0.918
English	Latin	0.703	0.732	0.803	0.822	0.823	0.826	0.810	0.818
German	Latin	0.591	0.705	0.730	0.772	0.770	0.772	0.777	0.779
English	French	0.659	0.623	0.812	0.802	0.836	0.830	0.823	0.823
French	German	0.498	0.534	0.734	0.645	0.796	0.788	0.763	0.760
Albanian	Latin	0.561	0.630	0.680	0.676	0.690	0.721	0.692	0.698
Albanian	French	0.499	0.610	0.653	0.658	0.607	0.625	0.666	0.663
Albanian	German	0.207	0.369	0.379	0.420	0.553	0.552	0.566	0.554
Albanian	English	0.289	0.302	0.382	0.446	0.503	0.518	0.486	0.485
Average		0.584	0.628	0.704	0.709	0.743	0.749	0.743	0.743
Standard deviation		0.231	0.193	0.194	0.176	0.149	0.144	0.143	0.146
Variance		0.054	0.037	0.038	0.031	0.022	0.021	0.020	0.021
Median		0.576	0.627	0.732	0.724	0.783	0.780	0.770	0.770

Table 4. 11-point interpolated average precision of several methods

Student's t-test			
Sample1	**Sample2**	**p-value**	**Statistical significance**
Main comparisons			
DAY6b	DBN	0.030	Good evidence
DAY76b	DBN	0.028	Good evidence
Secondary comparisons			
DAY6b	NEDIT	0.0004	Strong evidence
DAY6b	ALINE	0.001	Strong evidence
DAY6b	PHMM	0.025	Good evidence
DAY76b	NEDIT	0.0004	Strong evidence
DAY76b	ALINE	0.0004	Strong evidence
DAY76b	PHMM	0.029	Good evidence

Table 5. Statistical significance of DAY6b and DAY76b using SW

4.3 Statistical Significance

In order to understand if our results represent a statistically significant improvement or have been achieved by chance, we have run some paired two-sample Student's t-tests [40]. A Student's t-test determines whether two samples having a comparable average are likely to have come from the same population or from two different populations. We have assumed that the two samples are normally distributed but we did not suppose that the variances are equal because the sample size of the two compared groups is the same. This assures that the Student's t-test is highly robust to the presence of unequal variances [33]. Each sample has consisted of the ten 11-point interpolated average precision scores between language pairs produced by one of the systems reported in Table 4. We have conducted paired tests, which calculate the difference between arithmetic means of paired samples, because the samples to compare were not independent. For each test, our experimental hypothesis has been that our sample contained higher 11-point interpolated average precision scores than the sample we wanted to compare with. As a consequence, the null hypothesis we have tested for rejection has been that our sample did not contain 11-point interpolated precision scores higher than the sample with which we wanted to compare. Because the null hypothesis states a predicted direction of outcome, we have run one-tailed t-tests, meaning that our interest is only in one tail of the Student's distribution.

Table 5 shows the p-values and the consequent statistical significance of the t-tests that we have run to compare the best results obtained by DAY6b and DAY76b using SW, with the other systems reported in Table 4. All the t-tests have rejected the null hypothesis with strong or good evidence and have confirmed the experimental hypothesis. This validates the statistical significance of our results in the task of cognate identification that outperform those achieved by other systems previously reported in the literature [22,30,27].

It is worth noting that the statistical significance has remained stable with the enlargement of the training dataset dimension. In fact, we have run a t-test between the best results of DAY6b and DAY76b to check any possible statistical difference between the two. The p-value found, which is 0.199, has given no evidence of any statistical difference between DAY6b and DAY76b samples. This would suggest that the dimension of the training dataset for the learning system does not influence its statistical significance.

5. Conclusions

We have studied the influence of the training dataset dimension on a learning system developed using PAM-like matrices and a family of similarity measures in the task of cognate identification to show its robustness. In doing so, we have trained the system with about 62,000 cognate word pairs from 76 Indo-European languages after having automatically aligned them using global alignment and a linguistics-inspired substitution matrix. We have compared the obtained averaged 11-point interpolated average precision with previous results produced by the same system trained with only 650 cognate word pairs from a subset of 6 Indo-European languages, sensibly aligned as described. The results have shown to be remarkably consistent and did not present any relevant difference between the two model groups that both outperform comparable orthographic and phonetic systems previously proposed in the literature.

We have also investigated the statistical significance of our results when compared with earlier proposals and with each other. The outcome has shown, with strong and good evidence, that the averaged 11-point interpolated average precision of our system is approximately 5% higher than those achieved by comparable orthographic and phonetic systems. Moreover, when the results obtained by the models trained respectively with about 62,000 and with about 650 cognate pairs have been tested against each other, they have shown no evidence of any statistical difference. Indeed, the training dataset dimension seems not to influence either the performance or the statistical significance of this learning system.

This hypothesis is very encouraging because it would overcome one of the limits of learning systems, which is the need of a large training dataset. If a small group of sensibly aligned cognate pairs is able to train properly this system, not only may it help to discover distant relationships between languages or language families when there is no consensus, but may also be particularly useful in the study of those languages that do not benefit from large cognate datasets. This may be the case with extinct languages and their relationships in the field of historical linguistics, but also with less studied and less spoken speech varieties in several fields of natural language processing like machine translation, parallel bilingual corpora processing and lexicography.

The future and fascinating aim of this work is to apply the described methodology to the study of language evolution starting with the Indo-European language family.

References

[1] G. W. Adamson, J. Boreham, "The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles," *Information Storage & Retrieval*, vol. 10, no. 7-8, pp. 253-260, 1974.

[2] Q. D. Atkinson, R. D. Gray, "Curious Parallels and Curious Connections - Phylogenetic Thinking in Biology and Historical Linguistics," *Systematic Biology*, vol. 54, no. 4, pp. 513-526, 2005.

[3] C. Brew, D. McKelvie, "Word-pair extraction for lexicography," in *Proceedings of the 2nd International Conference on New Methods in Language Processing*, Ankara, 1996, pp. 45-55.

[4] K. W. Church, "Char_align: a program for aligning parallel texts at the character level," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-1993)*, Columbus, Ohio, U.S.A., 1993, pp. 1-8.

[5] M. A. Covington, "Alignment of Multiple Languages for Historical Comparison," in *Proceedings of COLING-ACL'98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 1998, pp. 275-280.

[6] M. A. Covington, "An Algorithm to Align Words for Historical Comparison," *Computational Linguistics*, vol. 22, no. 4, pp. 481–496, December 1996.

[7] C. R. Darwin, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London, U.K.: John Murray, 1859.

[8] M. O. Dayhoff, R. V. Eck, "A Model of Evolutionary Change in Proteins," *Atlas of Protein Sequence and Structure 1967-1968*, vol. 3, pp. 33-41, 1968.

[9] M. O. Dayhoff, R. V. Eck, C. M. Park, "A Model of Evolutionary Change in Proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 89-99, 1972.

[10] M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, "A Model of Evolutionary Change in Proteins," *Atlas of Protein Sequence and Structure*, vol. 5, no. 3, pp. 345-352, 1978.

[11] A. Delmestri, N. Cristianini, "String Similarity Measures and PAM-like Matrices for Cognate Identification," accepted for publication in *Bucharest Working Papers in Linguistics*, vol. XII, no. 2, 2010.

[12] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis*. Cambridge, U.K.: Cambridge University Press, 1998.

[13] I. Dyen, J. B. Kruskal, P. Black, "An Indoeuropean classification: A lexicostatistical experiment," in *Transactions of the American Philosophical Society*, vol. 82, part 5, 1992.

[14] K. Filali, J. Bilmes, "A Dynamic Bayesian Framework to Model Context and Memory in Edit Distance Learning: An Application to Pronunciation Classification," in *Proceedings of ACL 2005*, 2005, pp. 338-345.

[15] O. Gotoh, "An improved algorithm for matching biological sequences," *Journal of Molecular Biology*, vol. 162, no. 3, pp. 705-708, December 1982.

[16] D. Gusfield, *Algorithms on Strings, Trees and Sequences*. New York, U.S.A.: Cambridge University Press, 1997.

[17] J. BM. Guy, "An algorithm for identifying cognates in bilingual word-lists and its applicability to machine translation," *Journal of Quantitative Linguistics*, vol. 1, no. 1, pp. 35-42, 1994.

[18] D. Inkpen, O. Frunza, G. Kondrak, "Automatic Identification of Cognates and False Friends in French and English," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgaria, 2005, pp. 251-257.

[19] B. Kessler, "Computational dialectology in Irish Gaelic," in *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*

(EACL), San Francisco, California, U.S.A., 1995, pp. 60-66.

[20] B. Kessler, *The Significance of Word Lists*. Stanford, California, U.S.A.: CSLI Publications, 2001.

[21] P. Koehn, K. Knight, "Knowledge sources for word-level translation models," in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001, pp. 27-35.

[22] G. Kondrak, "A New Algorithm for the Alignment of Phonetic Sequences," in *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, vol. 4, Seattle, Washington, 2000, pp. 288-295.

[23] G. Kondrak, "Algorithms for Language Reconstruction," University of Toronto, Canada, PhD Thesis 2002.

[24] G. Kondrak, "Cognates and Word Alignment in Bitexts," in *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, Phuket, Thailand, 2005, pp. 305-312.

[25] G. Kondrak, B. J. Dorr, "Identification of Confusable Drug Names: A New Approach and Evaluation Methodology," in *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 2004, pp. 952-958.

[26] G. Kondrak, D. Marcu, K. Knight, "Cognates Can Improve Statistical Translation Models," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003) companion volume*, Edmonton, Alberta, 2003, pp. 46-48.

[27] G. Kondrak, T. Sherif, "Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification," in *Proceedings of the COLING-ACL Workshop on Linguistic Distances*, Sydney, Australia, 2006, pp. 43-50.

[28] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[29] W. Mackay, "Word Similarity using Pair Hidden Markov Models," University of Alberta, Canada, Master's thesis 2004.

[30] W. Mackay, G. Kondrak, "Computing word similarity and identifying cognates with Pair Hidden Markov Models," in *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*, Ann Arbor, Michigan, U.S.A., 2005, pp. 40–47.

[31] C. D. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts, U.S.A.: The Massachusetts Institute of Technology (MIT) Press, 1999.

[32] G. S. Mann, D. Yarowsky, "Multipath Translation Lexicon Induction via Bridge Languages," in *Proceedings of NAACL 2001*, 2001, pp. 151-158.

[33] C. A. Markowski, E. P. Markowski, "Conditions for the Effectiveness of a Preliminary Test of Variance," *The American Statistician*, vol. 44, no. 4, pp. 322-326, November 1990.

[34] T. McEnery, M. Oakes, "Sentence and Word Alignment in the CRATER Project," in *Using Corpora for Language Research*, Jenny Thomas and Mick Short, Eds.: Longman, 1996, ch. 13, pp. 211-231.

[35] D. Melamed, "Bitext maps and alignment via pattern recognition," *Computational Linguistics*, vol. 25, no. 1, pp. 107-130, 1999.

[36] A. Mulloni, V. Pekar, "Automatic detection of orthographic cues for cognate recognition," in *Proceedings in the 5th international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 2006, pp. 2387-2390.

[37] S. B. Needleman, C. D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443-453, March 1970.

[38] J. Nerbonne, W. Heeringa, "Measuring dialect distance phonetically," in *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97)*, Madrid, 1997, pp. 11-18.

[39] M. P. Oakes, "Computer Estimation of Vocabulary in Protolanguage from Word Lists in Four Daughter Languages," *Journal of Quantitative Linguistics*, vol. 7, no. 3, pp. 233-243, 2000.

[40] L. R. Ott, M. Longnecker, *An Introduction to Statistical Methods and Data Analysis*, 5th ed. Pacific Grove, California: Duxbury Press, 2001.

[41] A. Pirkola, J. Toivonen, H. Keskustalo, K. Visala, K. Järvelin, "Fuzzy Translation of Cross-Lingual Spelling Variants," in *Proceedings of the 26th Annual International ACM SIGIR'03 Conference on Research and Development in Information Retrieval*, Toronto, Canada, 2003, pp. 345–352.

[42] J. Prokić, M. Wieling, J. Nerbonne, "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, Athens, Greece, 2009, pp. 18–25.

[43] S. Schulz, K. Markó, E. Sbrissia, P. Nohama, U. Hahn, "Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon," in *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, vol. 2, Geneva, Switzerland, 2004, pp. 813-819.

[44] M. Simard, G. F. Foster, P. Isabelle, "Using Cognates to Align Sentences in Bilingual Corpora," in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, Montreal, Canada, 1992, pp. 1071-1082.

[45] T. F. Smith, M. S. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195-197, March 1981.

[46] M. Swadesh, "Lexico-Statistic Dating of Prehistoric Ethnic Contacts," *Proceedings of the American Philosophical Society*, vol. 96, no. 4, pp. 452-463, August 1952.

[47] J. Tiedemann, "Automatic construction of weighted string similarity measures," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 213-219.

[48] R. A. Wagner, M. J. Fisher, "The String-to-String Correction Problem," *Journal of the Association for Computing Machinery*, vol. 21, no. 1, pp. 168-173, January 1974.

[49] M. Wieling, J. Prokić, J. Nerbonne, "Evaluating the Pairwise String Alignment of Pronunciations," in *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage Social Sciences, Humanities, and Education*, Athens, Greece, 2009, pp. 26-34.