

# String Similarity Measures and PAM-like Matrices for Cognate Identification

Antonella Delmestri  
Department of Information Engineering  
and Computer Science  
University of Trento  
Italy  
antonella.delmestri@disi.unitn.it

Nello Cristianini  
Intelligent Systems Laboratory  
University of Bristol  
U.K.  
nello@support-vector.net

## Abstract

We present a new automatic learning system for the identification of cognates, words that derive from a common ancestor and share the same etymological origin. Our approach combines and adapts several techniques developed for biological sequence analysis to the natural language processing environment. We design a linguistic-inspired matrix to align sensibly our training dataset. We introduce a PAM-like technique, similar to the one successfully used in biological sequence alignment, in order to produce substitution matrices. We propose a novel family of parameterised string similarity measures and we apply them together with the PAM-like matrices to the task of cognate identification. We develop and test our proposal on standard datasets of Indo-European languages in orthographic format based on the Latin alphabet, but it could easily be adjusted to datasets using any other alphabet, including the phonetic alphabet if data in phonetic transcription were available. We compare our system with other models reported in the literature and the results show that our method outperforms in terms of precision both orthographic and phonetic approaches formerly presented.

## Keywords

Cognate identification, substitution matrices, string similarity measures.

## 1. Introduction

Language is a defining feature that distinguishes modern humans from all the other species, is a carrier of culture and plays a key role in communication. The analogy of language evolution with species evolution, predicted by Charles Darwin in his “*On the Origin of Species*” [2] has aroused a growing interest in the scientific community following the amazing progress of computational molecular biology in the field of genomes. Bioinformatics techniques are now applied to the field of natural language processing where they are making significant contributions and presenting exciting opportunities for further investigation.

Natural languages that originate from a common ancestor are genetically related, words are the backbone of any natural language and *cognates* are words sharing the same ancestor and etymology. Therefore cognate

identification represents the foundation for discovering the evolutionary history of languages. However, cognate recognition has proved to be useful not only in historical linguistics, but also in the very diverse fields of natural language processing. Applications that benefit from cognate identification include lexicography [1], parallel bilingual corpora processing, such as sentence alignment [23], word alignment [27,14] and lexicon translation [22], statistical machine translation [17], and confusable drug name detection [16].

In historical linguistics, cognates are also called *strict or genetic cognates* as they derive from a “*vertical*” transmission and they do not include borrowings. *Borrowings or loans* are words borrowed from other languages through a “*horizontal*” transmission and for this reason do not follow the same phonological changes that occur over time. In many disciplines of natural language processing, the term cognates or *broad cognates* has a wider meaning and also includes borrowings.

When relatedness between cognates have to be evaluated, the methodologies applied can be either *orthographic*, where cognates are analysed in their writing form of graphemes, or *phonetic*, where cognates have to be represented in a phonetic notation in order to be examined. The *orthographic approach* relies on the fact that alphabetic character correspondences represent in some way sound correspondences, as sound changes leave traces in the orthography. However it does not require any phonetic transcription whose attainment is still a very time consuming and challenging task. On the other hand, in evaluating word relatedness, *phonetic methods* depend on phonetic transcriptions of texts but benefit from the phonetic characteristics and features of phonemes that can be decomposed into vectors of phonetic attributes. Even if for the task of cognate identification a phonetic approach is supposed to be more accurate than an orthographic one for its understanding of phonetic changes, the debate is still open and a comparative evaluation of several recent results seems to prove the opposite [20,18,15].

Another differentiating feature between methods applied to the assessment of word relatedness is the capacity to adapt to different contexts, and, based on that, evaluation systems can be either *static* or *active*. A

*static system* is based on manually designed and incorporated knowledge, does not require any supervision and is not able to learn by processing data. On the other hand, an *active system* has the capacity to learn and adjust but may need supervision.

Several different approaches to the cognate identification problem have been proposed and orthographic or phonetic methodologies have been applied as well as learning algorithms or manually-designed procedures. In this paper we consider some authoritative methods proposed in the literature and compare them with our novel system.

The remainder of the paper is organized as follows. In Section 2 we introduce alignments and substitution matrices to the task of word relatedness. In Section 3 we propose our new learning system, including a linguistic-inspired matrix to align the training dataset, a PAM-like technique to produce scoring schemes and a novel family of string similarity measures. In Section 4 we describe the experimental design including the datasets used and the evaluation methodology. Finally, in Section 5 we present and discuss the results of our investigation and compare them with others in the literature.

## 2. Word relatedness

Cognate words can be studied by string matching techniques and cognate recognition represents a typical *inexact string matching* problem [10]. By adopting this approach to determine the relatedness of two strings, it is possible to either measure their distance, evaluating how distant the two strings are from each other, or to measure their similarity, calculating instead how similar the two strings are. The distance method leads to a minimisation problem because its aim is to find the minimum distance between two strings, while the similarity method guides towards a maximisation problem as its target is to find the maximum similarity between two strings.

### 2.1 Alignments

The task of calculating the distance or the similarity between two strings is closely related to the task of finding an optimal alignment between the two strings: dynamic programming algorithms can perform both tasks [10]. *Global* or *local alignment* algorithms, widely used in biological sequence analysis<sup>1</sup>, usually consist of a scoring system for measuring distance or similarity between the characters of the alphabet employed and a procedure for finding the optimal alignment. Even if the small length of the cognate words could make global alignment apparently more appropriate, local alignment can be useful in order to focus on the word roots, disregarding inflectional and derivational affixes [13]. Local alignment is only appropriate under the similarity approach. The dynamic programming algorithm for

solving the problem of *global sequence alignment* is known as the *Needleman-Wunsch algorithm* [24], but the more efficient version generally used was introduced by Gotoh [9]. The dynamic programming algorithm for solving the problem of *local sequence alignment* is called the *Smith-Waterman algorithm* [25], but the more efficient version generally used is again the one proposed by Gotoh [9].

### 2.2 Substitution matrices

*Substitution matrices* or *scoring matrices* are widely used in bioinformatics in the context of protein or nucleic acids sequence alignments. The significance of the resulting alignment depends greatly on the chosen scoring scheme, that is generally symmetric, and no one method is right for all types of applications [10].

Given an alphabet  $\mathcal{A}$  with  $|\mathcal{A}| \geq 2$ , each character of  $\mathcal{A}$  is more or less likely to transform into several other characters over time. A *substitution matrix*  $|\mathcal{A}|$ -by- $|\mathcal{A}|$  over  $\mathcal{A}$  represents the rates at which each character of  $\mathcal{A}$  may change into another character of  $\mathcal{A}$ . These rates in principle can be costs when they signify distances or can be scores when they signify similarities. Ideally, substitution matrices should reflect the true probabilities of mutations occurring through a period of evolution and should contain values proportional to these probabilities.

There are many different ways to construct a substitution matrix but the general approach is to assemble a large sample of verified pairwise alignments, or multiple sequence alignments, and derive the scores using a probabilistic model. Ideally, the scores in the matrix should reveal the phenomena that the alignments try to represent and the target is to assign a score to the alignments that gives a measure of the relative likelihood that the sequences are related as opposed to being unrelated [6]. In doing this, the *log-odds-ratio* is considered, that is the logarithm of the ratio of the probability that the sequences are associated as opposed to being random. The choice of the logarithm base is generally not important. In the related or *match model*, aligned pairs of residues occur with a joint probability, and the probability for the whole alignment is the product of these joint probabilities. In the unrelated or *random model*, the probability of the two sequences is just the product of the probabilities of each character, because the model assumes that each character occurs independently. The scores arranged in a matrix constitute the substitution matrix.

## 3. A new learning system

In order to study word relatedness, we have decided to choose the similarity approach which is the standard in biological sequence analysis and frequently used in natural language processing. Similarity allows local alignment, as well as global alignment, to be performed and it leads to the maximisation problem of finding the highest scoring alignment of the two words. We have developed a new learning system utilising orthographic

---

<sup>1</sup> In biological sequence analysis, strings are generally addressed as sequences.

data based on the Latin alphabet, but our proposal may be easily adapted to any alphabetic system, including the phonetic alphabet.

### 3.1 A linguistic-inspired matrix

In order to generate automatically a sensibly aligned training dataset, we have produced a linguistic-inspired substitution matrix based on knowledge of orthographic changes in the Indo-European languages. We have considered the 26 letters of the Latin alphabet and we have prepared a symmetric 26-by-26 matrix that contains a-priori likelihood of transformation between each character of the alphabet into another. We have given a value of 2 to all the elements of the main diagonal, because it is likely that a character preserves itself. We have assigned zero values to all the character transformations considered “possible”, a score of -3 to all the character transformations considered “impossible” and a gap penalty of -1 for insertion and deletion (*indels*), not to have overlaps between two indels and an “impossible” match. We have tried to represent in the linguistic-inspired matrix the traces that systematic sound changes left in written languages. Vowel shift chain, consonant shift chain including Grimm’s and Verner’s laws, rhotacism, assimilation, dissimilation, lenition and L-vocalisation have been considered. We have used this substitution matrix to perform pairwise alignment on cognate pairs by the Needleman-Wunsch algorithm [24], which is the standard for sequence global alignment. If more optimal alignments have been found, one of them has been randomly chosen.

### 3.2 PAM matrices

We have investigated PAM matrices that have been the standard and sole substitution matrices for amino acid alignments up until the advent of BLOSUM matrices [11]. The term PAM is an acronym for “Accepted Point Mutation” and refers to a family of amino acid substitution matrices, developed by Margaret Dayhoff et al. [3,4,5], that encode and summarise expected evolutionary changes of amino acids. An accepted point mutation in a protein is a replacement of one amino acid by another that has been accepted by natural selection and passed on to its progeny. The name PAM is also used as a measure unit to express the evolutionary divergence between two amino acid sequences. In this way, a PAM0 matrix coincides with the identity matrix where each character is considered maximally similar to itself, but not able to transform into any other character. The foundation of Dayhoff and co-workers approach is to obtain substitution rates from global alignments between closely related proteins and then to extrapolate this data to longer evolutionary divergence. The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence and all other PAM matrices are extrapolated from PAM1. This approach assumes that the frequencies of the amino acids remain constant over time and that the mutational process

causing replacements in an interval of 1 PAM unit operate the same for longer periods [10].

### 3.3 PAM-like matrices

Due to the lack of supervised and organised databases of cognate words and to the small length of words compared with the length of biological sequences, we have been forced to make some choices that differentiate partially our method, from the one Margaret Dayhoff and co-workers used to create the PAM matrices for biological sequence analysis. Their starting point was to identify a group of protein families where each pair of sequences showed amino acid diversity up to 15% and from them they built hypothetical phylogenetic trees with the parsimony method [3]. The group of cognate families showing up to 15% of identity that we have been able to extract from our dataset has been completely useless as composed of a few families of nearly identical words where the only mismatches were due to *indels*. Increasing the identity threshold up to 25% or 35% has not produced any substantial improvement. For example, the cognate words Italian *fiore* and French *fleur*, that are clearly closely related, present a diversity of 80% as 4 letters over 5 are different. We have decided to use the whole dataset available and due to the small dimension of the cognate families we have compared the cognate words with each other and not with their hypothetical ancestors. We have then followed the Dayhoff method to produce a family of PAM-like matrices based on a non symmetric matrix M of mutation probabilities, where M(i,j) contains the probability that character  $\mathcal{A}_j$  mutates to character  $\mathcal{A}_i$  in 1 PAM unit. Firstly a matrix A of accepted point mutation has been calculated ignoring the evolutionary direction meaning that A(i,j) and A(j,i) were incremented every time character  $\mathcal{A}_i$  was replaced by  $\mathcal{A}_j$  or vice-versa. Then the relative mutability m(j) of each character  $\mathcal{A}_j$  has been calculated as the ratio of observed changes to the frequency of occurrence. Finally M has been calculated as follows:

$$M(i,j) = \frac{\mu * m(j) * A(i,j)}{\sum_i A(i,j)} \quad \forall i \neq j$$

$$M(j,j) = 1 - \mu * m(j) \quad \forall j$$

where  $\mu$  is a proportionality constant we set to 1.

To generate scoring matrices suitable for longer times, we have produced matrices  $M^n$  by multiplying matrix M by itself  $n$  times that gives the probability that any particular character mutates to another one in  $n$  PAM units. Each PAMn matrix was obtained by the following log-odds ratios where f(i) and f(j) are the observed frequencies of character  $\mathcal{A}_i$  and  $\mathcal{A}_j$  normalized respectively by the number of all mutations.

$$PAMn(i,j) = 10 * \log_{10} \frac{f(j)*M^n(i,j)}{f(i)*f(j)} = 10 * \log_{10} \frac{M^n(i,j)}{f(i)}$$

We have not scaled the values in the PAM-like matrices and we have left the final scores with two decimal numbers. Because we have not limited the identity percentage within the cognate family considered, 10 PAM-like matrices have shown to be sufficient for modelling the divergence time of the languages considered.

### 3.4 A family of string similarity measures

We have proposed a family of parameterised string similarity measures obtained through different normalisations of a generic similarity rating algorithm in the aim to take into account the similarity of each string with itself and to eliminate, or at least reduce, the bias due to different string length. Indeed, alignments of two identical strings do not have a constant rate under the similarity approach because the score depends on the length of the string but also on the substitution rates of the characters involved.

Given two strings,  $S_1$  and  $S_2$ , and a generic similarity rating algorithm  $AL$ , we have defined the family of string similarity measures in Table 1 by normalising in various ways the similarity rate between them using the similarity rates of each string with itself.

The similarity measure  $sim_1(S_1, S_2, AL)$  normalises the rate of a similarity rating algorithm  $AL$  applied to calculate the similarity of  $S_1$  with  $S_2$  by the *arithmetic mean* of the rates given by the same algorithm applied to calculate the similarity of each string with itself. The similarity measure  $sim_2(S_1, S_2, AL)$  does the same but normalises the rate by the *weighted arithmetic mean* that considers also the length of the two strings. The similarity measures  $sim_3$  and  $sim_4$  employ a normalisation by the *geometric mean* and the *weighted geometric mean* respectively, while  $sim_5$  and  $sim_6$  normalise by the *harmonic mean* and the *weighted harmonic mean*. The *Heronian mean* is used to normalise the rate in  $sim_7$ , the *root mean square* in  $sim_8$  and the *contraharmonic mean* in  $sim_9$ .

Following the idea of employing the similarity of each string with itself in calculating string similarity, other similarity measures may be added to this set. We have used the family of new similarity measures with the *Needleman-Wunsch* algorithm [24] for global alignment and with the *Smith-Waterman* algorithm [25] for local alignment, but the new measures may be used with any other similarity rating algorithm. All the similarity measures proposed significantly outperform the basic algorithms they are based on as shown in Section 5.

<i>String similarity measures</i>	<i>Normalised by</i>
$sim_1(S_1, S_2, AL) = \frac{2 * AL(S_1, S_2)}{AL(S_1, S_1) + AL(S_2, S_2)}$	<i>Arithmetic Mean</i>
$sim_2(S_1, S_2, AL) = \frac{(\text{len}(S_1) + \text{len}(S_2)) * AL(S_1, S_2)}{(\text{len}(S_1) * AL(S_1, S_1) + \text{len}(S_2) * AL(S_2, S_2))}$	<i>Weighted Arithmetic Mean</i>
$sim_3(S_1, S_2, AL) = \frac{AL(S_1, S_2)}{\sqrt{AL(S_1, S_1) * AL(S_2, S_2)}}$	<i>Geometric Mean</i>
$sim_4(S_1, S_2, AL) = \frac{AL(S_1, S_2)}{\sqrt{\text{len}(S_1) + \text{len}(S_2)} \sqrt{AL(S_1, S_1)^{\text{len}(S_1)} * AL(S_2, S_2)^{\text{len}(S_2)}}}$	<i>Weighted Geometric Mean</i>
$sim_5(S_1, S_2, AL) = \frac{(AL(S_1, S_1) + AL(S_2, S_2)) * AL(S_1, S_2)}{2 * AL(S_1, S_1) * AL(S_2, S_2)}$	<i>Harmonic Mean</i>
$sim_6(S_1, S_2, AL) = \frac{(\text{len}(S_1) * AL(S_2, S_2) + \text{len}(S_2) * AL(S_1, S_1)) * AL(S_1, S_2)}{(\text{len}(S_1) + \text{len}(S_2)) * AL(S_1, S_1) * AL(S_2, S_2)}$	<i>Weighted Harmonic Mean</i>
$sim_7(S_1, S_2, AL) = \frac{3 * AL(S_1, S_2)}{AL(S_1, S_1) + \sqrt{AL(S_1, S_1) * AL(S_2, S_2)} + AL(S_2, S_2)}$	<i>Heronian Mean</i>
$sim_8(S_1, S_2, AL) = \frac{AL(S_1, S_2)}{\sqrt{(AL(S_1, S_1))^2 + AL(S_2, S_2)^2} / 2}$	<i>Root Mean Square</i>
$sim_9(S_1, S_2, AL) = \frac{(AL(S_1, S_1) + AL(S_2, S_2)) * AL(S_1, S_2)}{AL(S_1, S_1)^2 + AL(S_2, S_2)^2}$	<i>Contraharmonic Mean</i>

Table 1 - A family of string similarity measures

## 4. Experimental design

We have designed our experiments for the task of cognate identification with the aim of generating an automatic system able to learn meaningful information such as traces of sound correspondences left in the words orthography.

### 4.1 Datasets

In order to develop our system, we have employed a training dataset and a test dataset with no intersection between the languages they are composed of.

The *training dataset* for our learning system has been extracted from the *Comparative Indo-European*

*Database* by Dyen et al. [7]. It contains 200-word Swadesh lists [26] of universal, non cultural and stable meanings from eighty-four contemporary Indo-European speech varieties. In it, each word is presented in orthographic format without diacritics, using the 26 letters of the Roman alphabet. The data are grouped by meaning and cognateness, which is reported as certain or doubtful. From all the languages available, we have extracted only the reported certain cognate word pairs belonging to a group of six languages, Italian, Portuguese, Spanish, Dutch, Danish and Swedish for a total of about 700 cognate pairs. We have corrected a few evident errors. We have then automatically aligned these word pairs as described in Section 3.1. We have chosen three Romance languages and three Germanic languages to have a balanced training dataset able to learn traces of sound correspondences of most of the language families the test dataset is made of, but contemporarily avoiding overlap between the languages of the training and test datasets.

The *test dataset* consists of the orthographic form of the 200-word Swadesh lists of English, German, French, Latin and Albanian provided by Kessler [12] enhanced with his cognateness information. We have discovered two inconsistencies<sup>2</sup> related to the cognation of two French – German word pairs, as the author has confirmed. To make our results comparable with others reported in the literature [20,18] where the same test datasets have been used, we have not corrected the mentioned errors.

## 4.2 Evaluation methodology

Cognate identification is an excellent method of measuring the ability of a word similarity evaluation system. We have examined pairs of words belonging to different languages but having the same meaning, for which the cognateness is known information. Ten language pairs deriving from the combination of the five languages present in the test dataset have been considered.

We have produced two families of PAM-like matrices, one based on the Roman alphabet and one on its extension with gap, as proposed in Section 3.3. The two learning models, trained on the 6 language dataset, have been named respectively DAY6 and DAY6b.

We have employed these families of PAM-like matrices to align and rate the word pairs of the test dataset with the basic sequence alignment algorithms [24,25] and the family of parameterised similarity measures proposed in Section 3.4. For the model based on the Roman alphabet a unary gap penalty has been applied in the alignment algorithms. Our aim has been to assign a score to each word pair that represents how likely the words are to be cognates. The calculated rates, which

<sup>2</sup> The Latin word “folium”, meaning “leaf”, is reported to be cognate with the French “feuille” and the German “Blatt”, but the latter are not reported as cognates with each other. The same happens to the Latin word “collum”, meaning “neck” with the French “cou” and German “Hals”.

are relative to each other and do not reflect any universal scale, have then been ordered. When more word pairs showed the same rate, the alphabetic order has been considered as well. We have expected to find high density of true cognates or *true positive* at the top of the list and low density of true cognates at its bottom. To appraise our string similarity system on the task of cognate identification, we have not used a score threshold that may be influenced by the type of application, the method used and the degree of language relatedness [15]. Instead we have employed an evaluation metric called *11-point interpolated average precision*, borrowed from the field of *Information Retrieval*, designed specifically to calculate rankings [21]. This measure is also frequently used by other systems in the field of cognate recognition [20,18] and we wanted to make our results properly comparable. For the same reason, we have not distinguished between cognates and borrowings.

## 5. Experimental results

We have employed the Needleman-Wunsch algorithm for global alignment and the Smith-Waterman algorithm for local alignment with the novel family of string similarity measures to evaluate the performance of our cognate identification system. For each PAM-like matrix and for each similarity measure, we have computed the 11-point interpolated average precisions for each of the ten language pairs of our test dataset and then we have calculated their average, standard deviation and median. As the identity matrix can be considered as a PAM matrix at 0 evolutionary distance, it has been included in the tests for completeness. The set of similarity measures proposed consistently outperforms the basic algorithm it is based on and the group that performs better in both models, not surprisingly, utilises local alignment. Figure 1 and Figure 2 plot the results produced by DAY6 and DAY6b using the Smith-Waterman algorithm (SW).

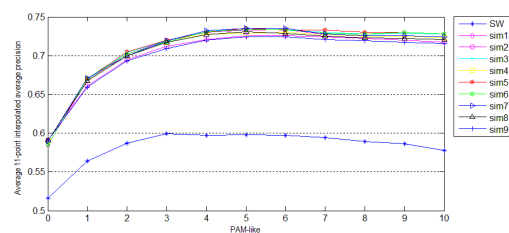


Figure 1 - Average 11-point interpolated average precision for DAY6 using SW

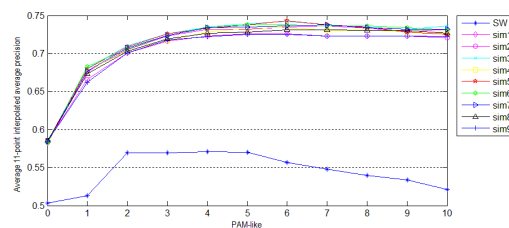


Figure 2 - Average 11-point interpolated average precision for DAY6b using SW

Matrix	SW	sim <sub>1</sub>	sim <sub>2</sub>	sim <sub>3</sub>	sim <sub>4</sub>	sim <sub>5</sub>	sim <sub>6</sub>	sim <sub>7</sub>	sim <sub>8</sub>	sim <sub>9</sub>
PAM0	0.503	0.583	0.585	0.585	0.586	0.585	0.583	0.585	0.586	0.585
PAM1	0.513	0.676	0.666	0.681	0.677	0.680	0.683	0.677	0.673	0.662
PAM2	0.569	0.705	0.701	0.710	0.704	0.709	0.706	0.707	0.703	0.701
PAM3	0.569	0.723	0.716	0.726	0.717	0.726	0.724	0.724	0.719	0.718
PAM4	0.571	0.732	0.723	0.735	0.728	0.733	0.735	0.734	0.727	0.722
PAM5	0.570	0.732	0.726	0.739	0.733	0.738	0.738	0.735	0.728	0.725
PAM6	0.557	0.735	0.726	0.742	0.734	<b>0.743</b>	0.738	0.736	0.731	0.725
PAM7	0.548	0.736	0.723	0.738	0.730	0.738	0.736	0.738	0.731	0.723
PAM8	0.540	0.733	0.723	0.735	0.731	0.735	0.736	0.734	0.730	0.723
PAM9	0.534	0.729	0.723	0.732	0.728	0.728	0.734	0.732	0.730	0.723
PAM10	0.521	0.732	0.721	0.736	0.728	0.725	0.731	0.732	0.727	0.722

Table 2 - Average 11-point interpolated average precision for DAY6b using SW

Table 2 displays in a tabular format the same results of the plot in Figure 2 i.e. the average of the 11-point interpolated average precisions over the ten language pairs of our test dataset achieved using DAY6b and the similarity measures based on the Smith-Waterman algorithm. Due to lack of space the results reached by DAY6 and DAY6b with the Needleman-Wunsch algorithm, that are slightly lower, are not reported.

## 5.1 Discussion

The two models DAY6 and DAY6b achieve very good results especially when employing local alignment with the Smith-Waterman algorithm. Between the two models, DAY6b, that utilises the extended alphabet, achieves the better results suggesting that the system is also able to learn appropriate gap penalties. Among the PAM-like matrices, PAM6 seems to represent the appropriate evolutionary divergence for the test dataset. All the similarity measures proposed perform consistently well and *sim*<sub>5</sub> presents the higher precision when combined with PAM6. These results suggest that our learning system outperforms all comparable orthographic and phonetic systems reported in the literature [20,18,15] as shown in Section 5.3.

## 5.2 Related works

Mackay [19] on the task of cognate identification followed the orthographic approach and developed a suite of Pair Hidden Markov Model (PHMM) variations and training algorithms based on a model originally presented by Durbin et al. [6]. The training dataset consisted of about 120,000 word pairs extracted from the *Comparative Indo-European Database* by Dyen et al. [7]. A development dataset was used to determine several parameters of the models. Mackay and Kondrak [20] tested this system on the dataset proposed by Kessler [12], that provides also word phonetic transcriptions, and they compared it with ALINE [13]. This is an algorithm for phonetic sequence alignment based on linguistic knowledge developed by Kondrak. Mackay and Kondrak tested the PHMMs also against the Levenshtein distance with Learned Weights (LLW)

method formerly proposed by Mann and Yarowsky to investigate the induction of translation lexicon via bridge languages [22]. The method employed the alphabet-weight edit distance [10] with modified costs for edit operations learned by a stochastic transducer from the same orthographic training dataset. Mackay and Kondrak showed that all the PHMMs outperformed the other methods and that a Viterbi-like log odds algorithm [6] gave significantly better results. We will call it hereinafter simply PHMM.

Kondrak and Sherif [18] working on orthographic data developed four different models of a Dynamic Bayesian Net (DBN) previously proposed by Filali and Bilmes for computing word similarity [8]. In order to train their system on the task of cognate recognition, Kondrak and Sherif extracted from the *Comparative Indo-European Database* by Dyen et al. [7] about 180,000 word pairs. They used them twice to enforce the symmetry of the scoring and they built up a development dataset to set-up the parameters of their system. They also evaluated a group of other phonetic and orthographic algorithms, including ALINE [13], LLW, and PHMM, and tested them on the dataset proposed by Kessler [12]. One of the DBN outperformed all the other systems including PHMM, but not significantly. We will call it only DBN.

Kondrak [15] investigated identification of cognates and recurrent sound correspondences testing several phonetic methods on the test dataset provided by Kessler [12]. His best result was achieved combining ALINE [13] with a sound correspondence-based method trained using a six languages development dataset. This set was extracted from the orthographic *Comparative Indo-European Database* by Dyen et al. [7] and then manually transcribed into a phonetic notation. This system improved the performance of ALINE, but did not outperform the orthographic PHMM and DBN previously described.

All the results presented in this section are quite remarkable because they suggest that orthographic learning models can outperform systems specifically designed for the task of phonetic alignment, like ALINE [13] and its variations [15], given enough training data.

### 5.3 Comparison

Both our models, DAY6 and DAY6b, when using global alignment as well as local alignment, consistently outperform both PHMM and DBN in term of precision on the task of cognate identification.

Table 3 shows the results produced by DAY6b, our best model, which outperforms PHMM and DBN. Moreover, the standard deviation calculated on our sample is much lower suggesting that our system is also more consistent in its performance across the different language pairs. This guess is confirmed by a higher median which indicates the central tendency. DAY6b utilises PAM6 as scoring matrix and  $sim_5$  as a similarity measure based on SW, as documented in Table 2.

The 11-point interpolated average precision achieved by PHMM and DBN is reported as in the literature [20,18]. The average results of PHMM and DBN are very close to each other maybe because a Hidden Markov Model can be considered as the simplest Dynamic Bayesian Network. However, DBN standard deviation is much lower showing a better data distribution. The proportion of cognate words present in each language pair of the test dataset is also reported.

Languages		Cognate proportion	PHMM	DBN	DAY6b
English	German	0.590	0.930	0.927	0.934
French	Latin	0.560	0.934	0.923	0.923
English	Latin	0.290	0.803	0.822	0.820
German	Latin	0.290	0.730	0.772	0.778
English	French	0.275	0.812	0.802	0.828
French	German	0.245	0.734	0.645	0.775
Albanian	Latin	0.195	0.680	0.676	0.674
Albanian	French	0.165	0.653	0.658	0.620
Albanian	German	0.125	0.379	0.420	0.566
Albanian	English	0.100	0.382	0.446	0.510
AVERAGE		0.284	0.704	0.709	<b>0.743</b>
Standard deviation		0.168	0.194	0.176	<b>0.145</b>
Median		0.260	0.732	0.724	<b>0.777</b>

Table 3 - 11-point interpolated average precision

We have used the same source for the training dataset and the same test dataset that Kondrak and co-workers have used in the design of PHMM and DBN. However, there are several aspects that differentiate deeply our learning approach, including the dimension of the training dataset used, its quality and its meaningfulness. In fact our system has employed less than 1% of the data they utilised and we have considered only word pairs reported by Dyen et al. [7] as certain cognates. Moreover, we have automatically aligned the cognate pairs the system has to learn from, using a substitution matrix that incorporates some linguistic knowledge in an attempt to generate a meaningful training dataset. These aspects may have contributed to generate a better

performance. It is also worth to notice that our system accommodates quite well the Albanian language that makes the test dataset challenging. In fact Albanian constitutes its own branch of the Indo-European language family and it is not part of the language families our system has been trained with.

### 6. Conclusion

We have developed a learning system for the task of cognate identification and we have proved its effectiveness against other orthographic models that have already been shown to outperform phonetic systems proposed in the literature. Our results reinforce the hypothesis that orthographic learning systems may recognise traces of sound correspondences left in the words orthography and can perform better than phonetic static models. This idea is very encouraging considering that phonetic transcriptions are very difficult to produce and frequently performed by hand with the consequent loss of time and the possible lack of accuracy and uniformity. Our PAM-like matrices, together with our new family of similarity measures, may help to discover distant relationships between languages, where controversies still exist, and to analyse less studied language families.

Our future objective is to continue investigating substitution matrices for the tasks of word similarity and cognate recognition. In particular we would like to model our PAM-like matrices employing a training dataset including more languages to study the influence of the training set dimension on the performance. Another step forward would be to experiment with other successful models borrowed from biological sequence analysis, like the BLOSUM matrices.

### Acknowledgments

We would like to thank Ana Fortun for her contribution to the linguistic-inspired matrix and Grzegorz Kondrak for providing his version of the test dataset.

### References

- [1] Chris Brew and David McKelvie, "Word-pair extraction for lexicography," in *Proceedings of the Second International Conference on New Methods in Language Processing*, Kemal Oflazer and Harold Somers editors, Ankara, September 1996, pp. 45-55.
- [2] Charles 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.
- [3] Margaret O. Dayhoff and R. V. Eck, "A Model of Evolutionary Change in Proteins," *Atlas of Protein Sequence and Structure 1967-1968*, vol. 3, pp. 33-41, 1968.

- [4] Margaret O. Dayhoff, R. V. Eck and C. M. Park, "A Model of Evolutionary Change in Proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 89-99, 1972.
- [5] Margaret O. Dayhoff, R. M. Schwartz and B. C. Orcutt, "A Model of Evolutionary Change in Proteins," *Atlas of Protein Sequence and Structure*, vol. 5, no. 3, pp. 345-352, 1978.
- [6] Richard Durbin, Sean R. Eddy, Anders Krogh and Graeme Mitchison, *Biological Sequence Analysis*. Cambridge, U.K., Cambridge University Press, 1998.
- [7] Isidore Dyen, Joseph B. Kruskal and Paul Black, "An Indo-European classification: A lexicostatistical experiment," in *Transactions of the American Philosophical Society*, vol. 82, part 5, 1992.
- [8] Karim Filali and Jeff 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.
- [9] Osamu Gotoh, "An improved algorithm for matching biological sequences," *Journal of Molecular Biology*, vol. 162, no. 3, pp. 705-708, December 1982.
- [10] Dan Gusfield, *Algorithms on Strings, Trees and Sequences*. New York, U.S.A., Cambridge University Press, 1997.
- [11] Steven Henikoff and Jorja G. Henikoff, "Amino acid substitution matrices from protein blocks," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, 1992, pp. 10915-10919.
- [12] Brett Kessler, *The Significance of Word Lists*. Stanford, California, U.S.A., CSLI Publications, 2001.
- [13] Grzegorz Kondrak, "A New Algorithm for the Alignment of Phonetic Sequences," in *Proceedings of NAACL 2000*, vol. 4, 2000, pp. 288-295.
- [14] Grzegorz Kondrak, "Cognates and Word Alignment in Bitexts," in *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, 2005, pp. 305-312.
- [15] Grzegorz Kondrak, "Identification of Cognates and Recurrent Sound Correspondences in Word Lists," *Traitement automatique des langues*, vol. 50, no. 2, pp. 201-235, October 2009.
- [16] Grzegorz Kondrak and Bonnie 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, August 2004, pp. 952-958.
- [17] Grzegorz Kondrak, Daniel Marcu and Kevin 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)*, Edmonton, Alberta, 2003, pp. 46-48.
- [18] Grzegorz Kondrak and Tarek 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.
- [19] Wesley Mackay, "Word similarity using Pair Hidden Markov Models," University of Alberta, Master's thesis 2004.
- [20] Wesley Mackay and Grzegorz 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.
- [21] Christopher D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, The Massachusetts Institute of Technology (MIT) Press, 1999.
- [22] Gideon S. Mann and David Yarowsky, "Multipath Translation Lexicon Induction via Bridge Languages," in *Proceedings of NAACL 2001*, 2001, pp. 151-158.
- [23] Dan Melamed, "Bitext maps and alignment via pattern recognition," *Computational Linguistics*, vol. 25, no. 1, pp. 107-130, 1999.
- [24] Saul B. Needleman and Christian 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, pp. 443-453, 1970.
- [25] T. F. Smith and M. S. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195-197, March 1981.
- [26] Morris Swadesh, "Lexico-statistic Dating of Prehistoric Ethnic Contacts," *Proceedings of the American Philosophical Society*, vol. 96, no. 4, pp. 452-463, August 1952.
- [27] Jörg 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.