A STATISTICAL ANALYSIS OF LANGUAGE EVOLUTION

MARCO TURCHI
Department of Information Engineering , University of Siena, Via Roma 36,
Siena, 53100, Italy
turchi@dii.unisi.it

NELLO CRISTIANINI
Department of Statistics, University of California Davis,One Shields Ave,
Davis CA,95616, US
nello@support-vector.net

We propose to address a series of questions related to the evolution of languages by statistical analysis of written text. We develop a ”statistical signature” of a language, analogous to the genetic signature proposed by Karlin in biology, and we show its stability within languages and its discriminative power between languages. Using this representation, we address the question of its trajectory during language evolution. We first reconstruct a phylogenetic tree of IE languages using this property, in this way showing that it also contains enough information to act as a ”tracking” tag for a language during its evolution. One advantage of this kind of phylogenetic trees is that they do not depend on any semantic assessment or on any choice of words. We use the ”statistical signature” to analyze a time-series of documents from four romance languages, following their transition from latin. The languages are italian, french, spanish and portuguese, and the time points correspond to all centuries from III bC to XX AD.

1. Introduction

In this paper we consider an aspect of language evolution, namely the process by which a language slowly changes by accumulation of many ”neutral mutations”, that is mutations that do not affect its effectiveness as a means of communication. The resulting ”drift” can be studied as a trajectory in a space, as we will describe below.

Biological evolution is the process by which all forms of life change slowly over time because of slight variations in the genetic sequences that one generation passes down to the next. It has been known for some time, now, that the majority of molecular mutations are selectively neutral, that is do not affect the fitness of the phenotype and hence are free to accumulate. The corresponding statistical model of sequence evolution (The Neutral Theory of Evolution, by Motoo Kimura) is a centerpiece of modern genomics. In that model, evolution corresponds to a trajectory in the space of all possible DNA sequences, with most steps being neutral with respect to selection, and mostly equivalent to a random walk. That neutral mutations can reach fixation for purely statistical reasons has been known for a long time.

Similar considerations can be made for the evolution of languages: neutral mutations accumulate, and some can become fixed in the population, over time. This creates a random walk, that can partly be reconstructed by simply keeping track of some statistical markers in the sequence, as done in DNA sequence evolution.

In this paper we investigate the use of statistical properties of languages to analyze linguistic evolution. We call them statistical language signatures (SLS) and we investigate how they evolve over time, how well their reflect ancestral relations between languages, and if they can be used to obtain language trees that are independent of any subjective choice. This approach by-passes any semantic assessment of word similarity or any arbitrary choice of words to be compared. It is repeatable automatically and hence objectively by simply performing statistical comparisons between text documents. Then we use SLS representation to analyze a time series of romance languages, from early latin to modern times. The approach is entirely data-driven. We make use of 3 datasets to independently validate our choice of features (SLS) and to analyze aspects of language evolution. A first dataset (containing 50 news stories written in 5 languages) is used to test the hypothesis that out representation is sufficiently stable and sensitive to characterize a language, at least within the domain of the indo-european (IE) family. The second corpus contains translations of the same document (”The universal declaration of human rights”) in 34 modern languages. And the third dataset contains literary works from early latin to modern romance languages, covering the past 22 centuries.

The fundamental observation is that the SLS of a text does not depend on its semantic content, but rather on the language in which it is written. In other words, all documents in a language have similar statistical signature. Another key observation is that all languages we examine have their characteristic SLS, and that they can be reliably identified by it. We test both these observations on the first dataset, with high statistical confidence.

The consequence of these two - apparently conflicting - observations is that the SLS evolves slowly, drifting over time, and diverging as the languages diverge from a common ancestor. In this, it behaves similarly to the genomic signatures introduced by Karlin and on which our analysis is based (Karlin, Mrzek, & Campbell, 1997). To test this hypothesis, we used the second corpus, and standard phylogenetic reconstruction algorithms, to reconstruct a tree of the IE family. The resulting tree, entirely based on statistical properties, is generally in agreement with the commonly accepted view of the IE family, although some exceptions are discussed in the Conclusions.

Finally, we focus on the process of drift of a language in statistical space. We model language evolution as a trajectory in the space of all possible statistical signatures, from an ancestral state to the current one. Modeling this drift is an important long term research goal, and we can only outline our approach in this paper. We use the third dataset to measure the distance covered by certain romance languages in the past 22 centuries. We notice some abrupt change points corresponding to known transitions from latin to national languages. At the end we outline a series of open problems, or research objectives, for this project.

In our current analysis we are limited by the use of texts available in the latin alphabet, and hence we focus mostly on european languages. However we believe that the methods can be exported to more general situations, perhaps using standard transliteration methods or - later - even phonetic representations.

1.1. *Statistical Language Signature*

It has been known for a long time that the probability of observing a certain character in a linguistic sequence depends strongly on the previous characters, and also is highly dependent on the language in consideration (Shannon, 1951). The frequency with which di-grams (pairs of letters) appear in a language is a very stable property of that language, as is a related quantity known as Karlin's odds ratio in genome analysis. If we remove all punctuation from a text document, all that is left is 26 letters and blank spaces separating them. So every document is a sequence from an alphabet of 27 letters. We denote by $C(i,j)$ the number of times that the di-gram (ij) is observed in the document. We can then define a di-gram frequency matrix as the matrix whose entry $D(i,j) = \frac{C(i,j)}{(n-1)}$ (where n is the document length). The odds-ratio matrix is defined as follows:

$$K(i,j) = \frac{C(i,j)}{C(i)C(j)}$$

where $C(i) = \sum_j C(i,j)$.

We want to investigate the use of D and K as statistical signatures of a language. We will also use them to assess the proximity between languages, and this means that we need to introduce a concept of distance that is appropriate in the space of matrices $\Re^{27\times 27}$. We are in this way defining a metric space where we ”embed” a language, and we model language evolution as a trajectory in that space.

We will use two simple distances. Other choices are naturally possible, and should be investigated separately.

- Frobenius Distance:
  $D_F(M^1, M^2) = \|M^1 - M^2\|_F = \langle (m^1_{i,j} - m^2_{i,j}), (m^1_{i,j} - m^2_{i,j}) \rangle = \sqrt{\sum_{i=1}^{27} \sum_{j=1}^{27} |m^1_{ij} - m^2_{ij}|^2}$

- Kalin (1-norm) Distance:
  $D_{L1}(M^1, M^2) = \frac{1}{(27)^2} \sum_{ij} |m^1_{i,j} - m^2_{i,j}|$

With these definitions, we can model a language as a point in a space, and its evolution as a trajectory in that space. We could even measure its rate of movement, in principle, since we have a notion of distance. Certainly we can define language similarity, and use that as a proxy in phylogenetic reconstruction. All this can make sense, however, only if these features are stable: they should be properties of the language, and not of the given document; and they should be able to distinguish between languages. If that can be proven, we can analyze phylogenetic relations between languages in this representation.

1.2. *Suitability of SLS as Features*

Each language has its own statistical signature. In english, digrams such as ”th” and ”ed” are very frequent, in italian the typical endings in vowels can be seen as high frequencies of digrams ”a-”, ”e-” etc (where we represented the blank symbol by ”-”). These differences, that reflect grammatical, phonetic and historical factors, can be readily seen in the feature matrices of the two languages.

To test the stability of these features within a language, as well as their reliability as discriminators between languages, we have used our first corpus: a set of 50 documents (10 each for English, German, Spanish, Italian and French). We computed the average pairwise distance for documents in the same language and for documents in different languages. We than compared their ratio with the same quantity measured for randomly created sets of 10 documents. We repeated this 10,000 times, and each time the resulting ratio was larger: with p-value $<$ 0.0001 this representation is well correlated to the difference between languages. Indeed, this quantity has been used to implement language classification systems for a long time (Beesley, 1988).

1.3. *Language Evolution in* $\Re^{27\times 27}$

If the SLS is a stable property of a language, and it is significantly different in related languages, it must be drifting over time. If this drift resembles a random walk (a hypothesis that should be tested in future work), then its *net* amount of drift should be proportional to the time dividing two languages, though a number of statistical corrections should be applied to the distance measured in feature space to really reconstruct the actual time since divergence. In this project we settle for a simpler test, using the pairwise distance matrix obtained with the expressions above to reconstruct a phylogenetic tree. We used the standard algorithm Neighbor Joining (Saitou & Nei, 1987), that is fairly tolerant to violations of the molecular clock assumption (genetic distance being proportional to time).

The dataset used for this part of the study is a subset of that used in (Benedetto, Caglioti, & Loreto, January 2002), our corpus being formed by 34 translations of the ”Universal Declaration of Human Rights” (UNResol, 1948) into modern languages from Romance, Celtic, German, Slavic, Baltic families, and the Basque language included as an outgroup. Also (Benedetto et al., January 2002) produced phylogenetic trees, using information theoretic tools.

The fact that each document is a translation of the ”Universal Declaration of Human Rights” offers the advantage that they all have roughly the same length, which facilitates our statistical analysis. The disadvantage however, is that in very close languages, the translation of the same word can be the same, or have the same root. This means that our estimate distances for adjacent/far languages might be biased.

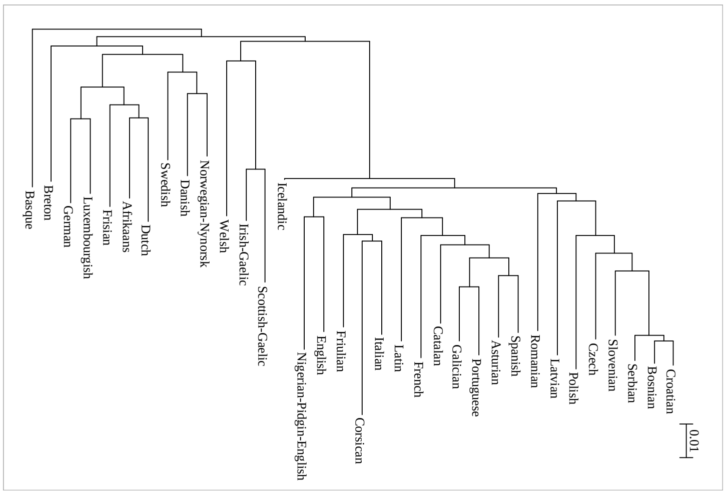

Figure 1. Language Evolution tree using the relative frequency of di-grams as features, and the Frobenius distance

The trees obtained with both SLSs (figures 1 and 2) are mostly compatible with the standard organization of the IE family, with the Karlin odds representation giving better results than the digrams. That means that not only can our SLS characterize a language, but can act as tags to track its evolution over long periods of time. Clearly this quantity seems to be changing slowly, and we can see from the fine organization of the slavic family or from the organization of languages in the iberian peninsula, it seems to also have a fairly steady drift. It is interesting to note that also the violations of the accepted topology of the tree can give us information about language evolution. For example, languages such as Romanian and English clearly are the result of massive borrowing from nearby languages, and an no longer be assigned to their original family (at least not their lexicon, which

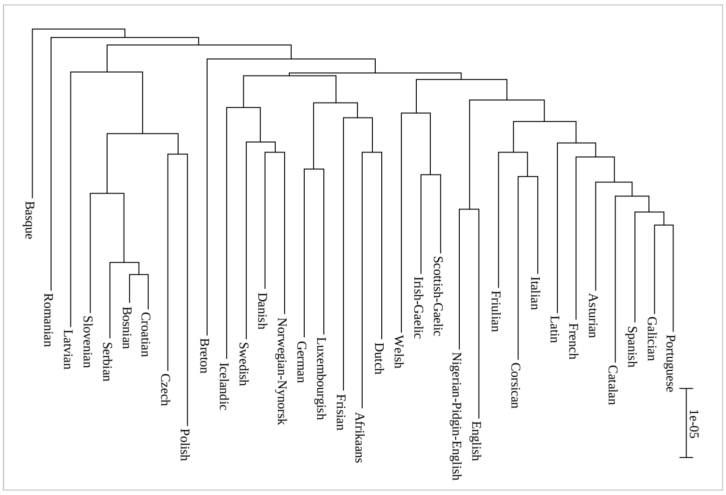

Figure 2. Language Evolution tree using the Odds ratios as features, and the Karlin distance

is what is captured mostly by this representation). In the di-grams representation there are various problems in assigning icelandic (which is instead correctly assigned by Karlin odds) and english in all cases seems to be attracted by french. This is better seen in the Multidimensional Scaling plot of the 34 languages.

Notice that we simplified the text to force it into a 26 letters alphabet, in so doing removing significant information, such as that coming from special letters in various languages. In particular we mapped the letters to their nearest english-alphabet counterpart, without using a linguistic criterion. Our assumption was that given the inherently statistical nature of the approach, we could ignore at a first approximation the effects of this arbitrary step, modeling them as a small perturbation of the signal. This has been the case for most languages, but in some cases, however, this rough simplification has proven to be sufficient to mislead the algorithms (see for example Breton). In the future, we are planning to make use of the phonetic alphabet, to reduce this effect.

1.4. *Time Series Analysis.*

The third experiment focused on time series analysis of documents spanning 22 centuries within the romance family. We constructed a dataset containing 119 different documents, written in Latin, Italian, Spanish, Portuguese and French, start from 200 BC and including the 20th century. Documents are mostly literary works, chosen to cover uniformly every period and every language. The non latin languages start mostly in the XI century, and have about 12 documents per century.

We measure the distance of each document from the oldest one, and we plot

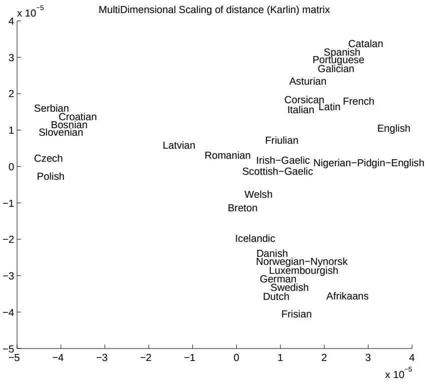

Figure 3. Multi Dimensional Scaling of some IE Languages

this distance as a function of time. The obvious change point observed in the XI century could be an artefact due to the fact that we could not find earlier documents in the non-latin languages, and is clearly a draw-back of using written language as opposite to spoken one. A more careful choice of the data might help us to reduce the gap in that transition. Also we can find written latin documents throughout the entire period, but we have stopped the latin series more or less where the national languages series started. What is more interesting is that the distance from the origin is in all languages more or less comparable: they all seem to have moved of a comparable amount, in the 22 centuries, although not smoothly (see figure 5). We can see the distances between these languages also in figure 4, although this multidimensional scaling representation can be misleading (projects into 2 dimensions a 27^2 dimensional dataset).

1.5. *Conclusions*

Various conclusions can be drawn from the experimental results we obtained: the first one is that some aspects of historical linguistics can indeed be investigated by using statistical tools. This rises hopes of applying the same tools to ancient texts, so as to look further back in time. But at the same time, a number of problems with this approach are visible in the results, directly suggesting various improvements.

First, it is not always the case that this statistical approach is robust enough to ignore the effect of alternative spelling conventions (as seen in the case of Breton and Icelandic). This can be addressed by moving future investigations to documents written using the IPA (international phonetic alphabet). Notice however that it can be argued that even spelling conventions evolve, and are part of the phylogenetic signal we are trying to analyze, as we focus on the evolution of written text. Second, we see the effect of borrowings (as seen in the case of English

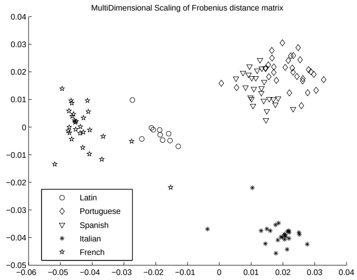

Figure 4. Multi Dimensional Scaling of some Romance Languages

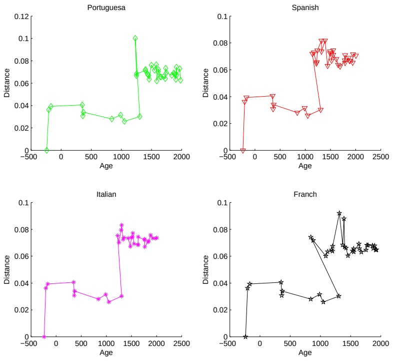

Figure 5. Time Series Analysis of some Romance Languages

and Romanian): in many cases the assumption that the evolutionary history of languages can be represented by a tree is not justified, at least with respect to their lexicon. This can be addressed by using tools from evolutionary biology aimed at reconstructing ”phylogenetic networks” rather than trees.

Because of the inherently statistical nature of this approach, however, to a first approximation we believe that all the above effects can be treated as random perturbations, and for most languages they are not sufficient to corrupt the phylogenetic signal. As we refine the method, we expect to find cleaner and more informative patterns in the data.

References

Beesley, K. R. (1988). Language identifier: A computer program for automatic natural-language identification of on-line text. *The 29th Annual Conference of the American Translators Association*, 4754.

Benedetto, D., Caglioti, E., & Loreto, V. (January 2002). Language trees and zipping. *Physical Review Letter*, *88*(4).

Karlin, S., Mrzek, J., & Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology*, *179*(12), 38993913. (0021-9193/97/04.0010)

Saitou, N., & Nei, M. (1987). The neighbour-joining method: a new method for constructing phylogenetic trees. *Mol. Biol. Evol.*

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell Systems Technical Journal*(30), 50-64.

Universal declaration of human rights. (1948, December). (United Nations General Assembly Resolution)