A Computational Model of the Phonetic Space and Its Applications

Anil Kumar Singh* — **Taraka Rama*** — **Pradeep Dasigi***

* *Language Technologies Research Centre, IIIT, Hyderabad, India*
{anil@research, taraka@students}.iiit.ac.in, pradeep.dasigi@gmail.com

RÉSUMÉ. *Nous présentons un modèle informatique de l'espace phonétique en termes de contextes phonétiques. Nous commençons par un modèle qui définit des contextes en termes de phonèmes. Nous considérons alors un modèle directement en termes de dispositifs phonétiques ou articulatoires, supposant d'abord que les dispositifs sont indépendants et puis considérants le cas quand ils ne sont pas indépendants. Utilisant un tel modèle de l'espace phonétique, nous pouvons calculer l'une ou l'autre similitude distributionnelle ou calculer la probabilité d'ordre pour le texte écrit dans une notation phonétique. Les deux, indépendamment du contexte modèle, peut être utiles pour beaucoup d'applications informatiques. Nous présentons des résultats pour quatre applications, à savoir construisant les arbres phylogénétiques avec des langues, de l'identification apparentée et de l'estimation d'origine de mot. Pour ces applications nous pouvions obtenir des résultats au moins comparables aux méthodes du dernier cri. Une autre application que nous discutons brièvement étudie la variation phonétique. Comme prolongation aux modèles, nous discutons également comment des mots peuvent être représentés dans l'espace phonétique pendant que les surfaces tridimensionnelles dans les trellis contraints de point et suggèrent quelques directions pour les travaux futurs.*

ABSTRACT. *We present a computational model of the phonetic space in terms of phonetic contexts. We start with a model that defines contexts in terms of phonemes. We then consider a model directly in terms of the phonetic or articulatory features, first assuming that the features are independent and then considering the case when they are not independent. Using such a model of phonetic space, we can calculate either distributional similarity or calculate sequence probability for text written in a phonetic notation. Both of these, apart from the context models, can be useful for many computational applications. We present results for four applications, namely constructing phylogenetic trees of languages, cognate identification and word origin guessing. For these applications we were able to get results at least comparable to the state of the art methods. Another application we briefly discuss is studying phonetic variation. As an extension to the models, we also discuss how words can be represented in the phonetic space as*

L'objet. Volume 8 – n°2/2005, pages 1 à 15

three dimensional surfaces in constrained point lattices and suggest some directions for future work.

MOTS-CLÉS : *Modélisation informatique, dispositifs phonétiques, similitude distributionnelle, probabilité d'ordre, distances de langue, arbres phylogénétiques, identification apparentée, origine de mot, variation phonétique, trellis de point*

KEYWORDS: *Computational modeling, phonetic features, distributional similarity, sequence probability, language distances, phylogenetic trees, cognate identification, word origin, phonetic variation, point lattices*

1. Introduction

Quite apart from speech processing proper, there is now a substantial quantity and variety of work in Natural Language Processing (NLP) that is, in some way, phonetic processing of text. Computational techniques have been developed for applications like estimating distances between languages for automatically constructing phylogenetic trees (Dyen *et al.*, 1992b; Nerbonne *et al.*, 1997; Kondrak, 2002a; Ellison *et al.*, 2006b), cognate identification (Ribeiro *et al.*, 2001), letter to phoneme conversion (Bartlett *et al.*, 2008), transliteration (Knight *et al.*, 1997; Haizhou *et al.*, 2004; AbdulJaleel *et al.*, 2003) etc. Some of these techniques derive from established work in linguistics (Swadesh, 1952), while some are mostly statistical (Marchand *et al.*, 2000; Damper *et al.*, 2004). However, majority of the work perhaps lies somewhere in between (Virga *et al.*, 2003; van den Bosch *et al.*, 2006).

The motivation for this paper is to find a common ground on which all these techniques can be based such that this common ground is both linguistically valid as much as possible and also directly relevant for computational purposes. We see this common ground (at least partly) as the phonetic space of languages. By computationally modeling this phonetic space and using this model for practical applications, we try to show that it is possible to come to a common ground for phonetic processing of text.

We present a model of the phonetic space in terms of phonetic contexts, which are in turn defined in terms of phonetic or articulatory features, with and without an independence assumption. We describe how this model can be used for calculating distributional similarity as well as sequence probability. We also use the model to solve the problems of constructing phylogenetic trees, cognate identification and word origin guessing. The results that we have obtained are comparable to or better than the state of the art. Near the end, we briefly present the initial results of a pilot study on semi-automatic study of phonetic variation and change.

The approach presented in the paper is not quite the classical supervised Machine Learning approach because for several applications for which these models can be used, no labeled data is needed. Also, we are making as much use of the linguistic information available as possible, without committing to any specific linguistic theory or framework or even a specific list of articulatory features. The idea is to use only as many features and values as are required to distinguish one phoneme from another for a certain purpose. Statistical techniques are used only after the available linguistic information has been modeled, instead of relying mostly on statistical techniques to induce all linguistic information.

Rather than presenting a unified approach for a large number of phonetics based models, the aim is to formally model all the linguistic information about phonetics as applied to written text and then suggest ways of using this information along with the time tested statistical techniques and their variations for our purposes. The motivation is to achieve simplicity in linguistic as well as statistical modeling for the same level of performance as obtained by more complex state of the art techniques.

The paper is organized as follows. In the next section we briefly consider the recent debate about the discreteness and universality of the phonetic space and explain how it does not affect our work. In Section 3, we start modeling the phonetic space by defining phonemes in terms of phonetic features. In Section 4, we present a model in terms of phonemes, followed by Section 5, in which we present the model in terms of phonetic features. In Section 6, we generalize the model to drop the independence assumption. In the next section, we consider a special case of the general model, viz. phonetic feature based n-grams. In Section 8, we present two of the ways of using the model of the phonetic space for computational purposes, viz. by estimating distributional similarity and by calculating sequence probability. In Section 9, we present the results of our experiments on using the model for constructing phylogenetic trees. In Section 10, we use the model for cognate identification and present the results. In the next section, we present the results of an experiment on word origin guessing. In Section 12 we describe an initial application of the model for studying phonetic variation and change. Section 13 goes back to extending the model and describes how words can be represented as three dimensional surfaces in constrained point lattices which represent the phonetic space. In Section 14, we suggest some directions for future work and conclude.

2. Phonetic Space: Discreteness and Universality

For computational applications such as estimation of language distance, we can potentially use crosslingual as well as intra-lingual comparison. Several of the earlier attempts (Nerbonne *et al.*, 1997; Kondrak, 2002a) were based on crosslingual comparison of phonetic forms, but some researchers have argued against the possibility of obtaining meaningful results from crosslingual comparison of phonetic forms. This is related to the idea of a common phonetic space. Port and Leary (Port *et al.*, 2005) have argued against it. Ellison and Kirby (Ellison *et al.*, 2006b) argue that even if there is a common space, language specific categorization of sound often restructures this space. They conclude that if there is no language-independent common phonetic space with an equally common similarity measure, there can be no principled approach to comparing forms in one language with another. They suggest that language-internal comparison of forms is better and psychologically more well-grounded. These and other objections require some clarification about the validity of computationally modeling the phonetic space, implicitly assuming that there is a common phonetic space and crosslingual comparison of phonetic forms can be meaningful for our purposes.

Although the idea of a phonetic space is quite an old one, it was redefined to be innate for all humans (Chomsky *et al.*, 1968) and this innate universal and discrete phonetic space has been the foundation of a lot of work in linguistics (Patricia Keating *et al.*, 1983; Hardcastle *et al.*, 1999; Wolfram *et al.*, 1998). Not everyone, however, agrees with this idea. Perhaps the most important and comprehensive criticism against the idea of a discrete symbolic universal phonetic space is from Port and Leary (Port *et al.*, 2005) in the article titled *Against Formal Phonology*. We will not go into this

debate here, but would only point out that this criticism is not very relevant for our (i.e., computational) purposes. Even Port and Leary concede that discrete representation of phonemes can be useful for some practical purposes:

> If a loose approximation to a formal system is all that is required, for example, if one were designing a practical orthography for a language or trying to facilitate adult language learning, then a simplified formal approximation to phonology (as used by most phonologists) is likely to be quite useful.
>
> ...
>
> We do not deny that the phonologies of languages exhibit symbol-like properties, such as reusable and recombinable sound patterns. A small inventory of segment-sized, graphically represented phonological categories can provide a practical scheme for representing most languages on paper.
>
> ...
>
> Certainly, there are still some generalizations to be drawn across languages, such as, say, the tendency of [ki] to evolve historically into [ci].

In psychology also, there has been a long debate about a similar problem which can be stated in terms of a common chromatic space (Saunderson *et al.*, 1946; Lucey *et al.*, 2002), possibly defined in terms of common chromatic features. Do humans in different cultures see the same colors? There is still no conclusive answer to this, but many computational techniques have been tried to solve real world problems like classifying human faces, seemingly with the implicit assumption that there is a common chromatic space. Such computational techniques have shown some success (Yang *et al.*, 1996; Chen *et al.*, 2003), thus providing empirical evidence for the validity of a common chromatic space at least from the computational point of view. This can serve as a useful analogy for our case where the validity of universal phonetic space can have direct implications for computational processing of written text. Conversely, successful computational techniques which are based on the idea of a common phonetic space could be a kind of empirical evidence in support of the idea of a common phonetic space.

To summarize, most of the criticism against the idea of a universal phonetic space is about how humans process, perceive and produce speech, whereas our attempt is mainly aimed at phonetic processing of written text for practical applications. Also, since our model is probabilistic, some of the criticism against discreteness does not apply to our model.

3. Modeling of the Phonetic Space

Before describing the three models with increasing coverage or generality, we define some symbols.

Let the universal set of phonemes found in human languages be:

$$\Phi = \{\phi_1,\ \phi_2,\ \ldots,\ \phi_{|\Phi|}\} \quad [1]$$

And the universal set of phonemic features which define these phonemes be:

$$F = \{f_1,\ f_2,\ \ldots,\ f_{|F|}\} \quad [2]$$

These phonemic features can take the following values:

$$V = \{v_1,\ v_2,\ \ldots,\ v_{|V|}\} \quad [3]$$

A phoneme is defined as a set of feature-value pairs:

$$\phi = \{(f_1, v_1),\ (f_2, v_2),\ \ldots,\ (f_{|f|}, v_{|f|})\} \quad [4]$$

For a specific language L, the set of phonemes that occur in that language will be a subset of Φ:

$$\Phi_L = \{\phi_1, \phi_2, \ldots, \phi_{|\Phi_L|}\},\ \Phi_L \subset \Phi \quad [5]$$

Similarly, the set of features which are applicable for the above set of phonemes will be a subset of F:

$$F_L = \{f_1, f_2, \ldots, f_{|F_L|}\},\ F_L \subset F \quad [6]$$

The model being probabilistic, everything occurs with a probability, such that $p(x)$ denotes the probability of x and $P(x)$ denotes the probability distribution over x.

In the next section we describe a model that is directly in terms of phonemes, i.e., treats phonemes as atomic units.

4. Phonemic Model

In essence, we define the model $\mho$ of (the structure of) the phonetic space as the collection of all possible contexts (c) and their frequencies or probabilities.

Thus, we can define such a model for a language L in terms of phonemes as:

$$\mho_L = \{c_1, c_2, \ldots, c_{|\mho_L|}\}, |\mho_L| \leq \Phi_L^T \quad [7]$$

where Φ_L^T is the number of phoneme tokens.

Note that the symbols represent types, not tokens, unless specifically marked by the superscript T. In the case of sequences (Section 8.2), however, there is a difference in notation as the elements of the sequences are tokens and the subscripts denote the phoneme index in the sequence.

The first model is a simple model in that it ignores the existence of phonemic features and models the phonetic space directly in terms of phonemes. If the words of a language are available in an IPA-like notation, then a simple phonemic n-gram model for that language can be easily created. Cavnar (Cavnar *et al.*, 1994) had shown that in a letter based n-gram of text, the top 300 or so n-grams represent the identity of the language and therefore can be used for language identification. This insight directly applies to phonemes as letters are implicitly supposed to represent the phonemes. We can say that the top few hundreds of phoneme based n-grams would not only represent the identity of the language, but they would do so because they roughly model the structure of the phonetic space of that language, as distinct from the structure of the phonetic space of another language. And the idea of universal phonetic space (even if as an approximation for practical purposes) means that both these spaces are part of the universal phonetic space.

The context c in this model is defined as a triple consisting of the phoneme, the left context c^λ and the right context c^ρ:

$$c = (\phi,\ c^\lambda,\ c^\rho) \quad [8]$$

The left and the right context themselves are ordered sequences of phonemes:

$$c^\lambda = (\phi_1, \phi_2, \ldots, \phi_{|c^\lambda|}) \quad [9]$$

$$c^\rho = (\phi_1, \phi_2, \ldots, \phi_{|c^\rho|}) \quad [10]$$

Since this model does not take into account the articulatory features, it fails to model the phonemes themselves. It works only with the information that is directly available on the surface. As a result, it cannot be used for, say, drawing generalizations about phonetic variation.

In the next section we define a model in terms of articulatory or phonemic features but with the assumption that the features are independent.

5. Independent Feature Model

The phonemic model does not capture a lot of information about the phonetic space of the language. For example, in Historical Linguistics (Hock, 1991), the laws of sound change are in terms of phonemic (or articulatory) features, not phonemes. Therefore a better model would be one that is based on phonemic features. Such a model can be statistically constructed from the same data as the previous model. However, instead of phoneme contexts, we would have phonemic feature contexts. At this stage, we can have two variations of the model: one which assumes that the features are independent of each other and the other that does not make this assumption. In reality, the features are not independent of each other as we can see by going back to the laws of sound change. These laws make no assumption of independence and most of them, in fact, involve more than one feature in the same law. This indicates that the independent feature model is not a good model of the phonetic space as it ignores the dependence between different features. However, we present this model as a step towards building the more general model described in the next section. An illustration of this model is provided in Figure 1.

The context c in this model is defined as a triple consisting of the feature-value pair, the left context c^{λ} and the right context c^{ρ}:

$$c = ((f,\ v),\ c^{\lambda},\ c^{\rho}) \qquad [11]$$

and the left and the right context themselves are ordered sequences of feature-value pair:

$$c^{\lambda} = ((f, v_1),\ (f, v_2),\ \ldots,\ (f, v_{|c^{\lambda}|})) \qquad [12]$$

$$c^{\rho} = ((f, v_1),\ (f, v_2),\ \ldots,\ (f, v_{|c^{\rho}|})) \qquad [13]$$

The independence assumption means that the value of a feature in a phoneme sequence depends only on the values of the same feature in the context. Since there will be as many feature sequences for a given phoneme sequence as there are feature types, the model can be defined as:

$$\mho_L = \{c_1, c_2, \ldots, c_{|\mho_L|}\}, |\mho_L| \leq \Phi_L^T \times |F| \qquad [14]$$

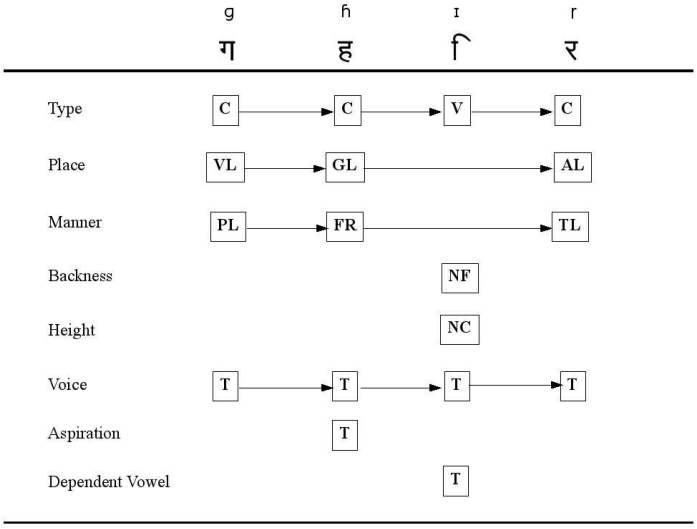

Figure 1. *An example of independent feature model showing an Avadhi word. Sequences can be composed of only the same feature. The values might be the same or different, depending on the phonemes. Note that the implicit shwa after the first consonant is not shown as the text is in Devanagari script. Feature codes used in the figure are: C is consonant, V is vowel, T is true, VL is velar, GL is glottal, AL is alveolar, PL is plosive, FR is fricative, TL is trill, NF is near-front and NC is near-close.*

6. Dependent Feature Model

This more general model of the phonetic space drops the independence assumption and captures the information about which feature occurs in the context of which other features. Therefore, in this model, context sequences would consist of a feature-value pairs such that any feature can precede or follow any other feature.

This model is very different from the previous model because, in the previous model, the elements of a sequence have the same feature but possibly different values, whereas in this model, any feature can follow any feature because the feature independence assumption has been dropped. This difference can be seen by comparing Figure 1 with Figure 2.

The left and the right context would now be defined as:

$$c^{\lambda} = ((f_1, v_1),\ (f_2, v_2),\ \ldots,\ (f_{|c^{\lambda}|}, v_{|c^{\lambda}|})) \qquad [15]$$

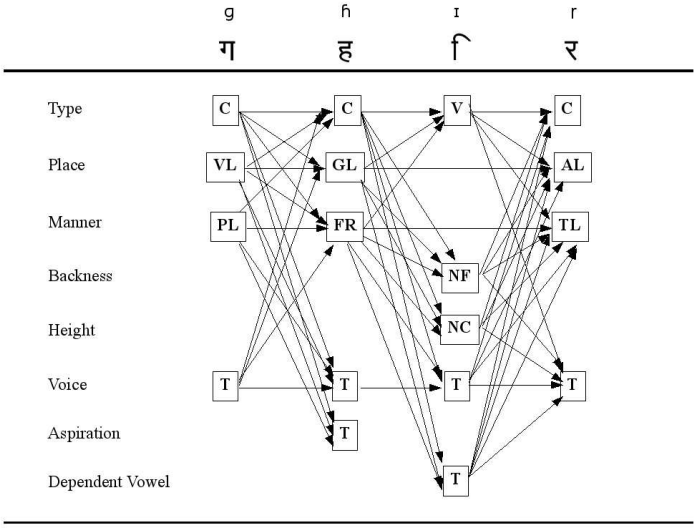

Figure 2. *An example of dependent feature model showing the same Avadhi word. Sequences can be composed of all possible combinations of feature value pairs for the given phonemes.*

$$c^{\rho} = ((f_1, v_1),\ (f_2, v_2),\ \ldots,\ (f_{|c^{\rho}|}, v_{|c^{\rho}|})) \qquad [16]$$

And the model would be:

$$\mho_L = \{c_1, c_2, \ldots, c_{|\mho_L|}\}, |\mho_L| \leq \Phi_L^T \times (|F| \times |V|)^{|c^{\lambda}|+|c^{\rho}|+1} \qquad [17]$$

In the most likely case, we would keep the lengths of the context fixed, i.e., c^{λ} and c^{ρ} will be constants. Therefore, the size of the model will not grow exponentially with length of the phoneme sequence. And in most of the practical cases, it might be possible to use further optimizations to reduce the size of the model and the complexity of the problem. It may be noted that the number of features and their values are finite and small enough and a lot of them do not occur together in any given language. Usually, we would be interested only in the most frequent contexts, so that most of the contexts can be pruned.

It may be stated here that the context discussed above is for a particular phoneme. In Section 8.2, which is about calculation of sequence probability, subscripts are used

to indicate the phoneme for a given feature-value pair, so that this information is not lost.

In Section 12, we discuss an obvious application of this phonetic context based model, viz. studying phonetic variation and change.

One possible problem with this model is that it is too general. Though the features are not independent, their dependence is also not completely arbitrary. This model does not explicitly or systematically take into account the nature of this limited dependence.

In the next section we consider a special case of this model.

7. Feature-Value N-Grams as a Special Case

The model described in the previous section is defined in terms of contexts in order to account for the nature of the phonetic space. In that model, the context is centered at a particular element (either a phoneme or a phonemic feature), i.e., the context is the context of a particular element. However, for many practical applications, it may not be required to have such 'centered contexts'. Instead, we can treat contexts as sequences of elements found in the language. Such a model defined in terms of 'centerless contexts' (basically n-grams) becomes a special case of the model in terms of centered contexts. Thus, in the case of phonemes, the context will be defined as:

$$c = (\phi_1,\ \phi_2,\ \ldots,\ \phi_{|c|}),\ where\ |c| < |\mho_L| \qquad [18]$$

Similarly, for dependent phonemic features, the centerless context would be:

$$c = ((f_1, v_1),\ (f_2, v_2),\ \ldots,\ (f_{|}c|, v_{|}c|)),\ where\ |c| < |\mho_L| \qquad [19]$$

These simplified models are basically n-gram models, either in terms of phonemes or phonemic feature-value pairs (whether dependent or independent). However, only phonemic n-grams have been tried previously. We suggest that phonemic feature n-grams can be used fruitfully for many applications. Since we are trying to relate contexts, i.e., centered contexts to n-grams, we have used the term 'centerless context' for n-grams. 'Centerless context' is the same as an n-gram, but it can also be seen as a special case of 'centered context'.

8. Using the Models of Space

As mentioned earlier, the model we are presenting is probabilistic. Phonemes, features, feature-value pairs and context can all be assigned probabilities based on a cor-

pus of text in phonetic notation. Thus, the model $\mho_L$ is ultimately a probability distribution where every context occurs with some probability. Given this fact, two of the general ways in which the models can be used are:

1) **By calculating distributional similarity**: We have two corpora in phonetic notation. We build a distributional model for both. Then, we use some distributional similarity measure such as KL-divergence (Sibun *et al.*, 1996) to calculate the distributional similarity of the two models. One of the corpora can, in fact, be just one word. This would happen for problems like word origin guessing. This way of using the model is similar to language identification. Both the corpora can be a long sequence or lists of words for problems like calculating language distances and for building phylogenetic trees of languages.

2) **By calculating sequence probabilities**: We again have two corpora in phonetic notation, but in this case we build a model out of only one. The other is a sequence of phonemes and we want to calculate the probability of the phoneme sequence. Whereas the first way (distributional similarity) is more suited for applications like calculating language distances, this way is better for checking whether a generated phoneme sequence is valid or not. One possible application is transliteration to those languages which use very phonetic scripts (e.g. Brahmi origin scripts) in the sense that the mapping between letters and words is almost one to one. Sequence probability calculation can also be useful for other problems like word origin guessing.

There are other ways in which the model can be used and we will consider one of them in Section 13. But before that we will describe how distributional similarity and sequence probability can be calculated and then we will briefly present the results of using the model for four applications.

8.1. *Calculating Distributional Similarity*

We would want to calculate distributional similarity if we have two different (phonetic) corpora and we have to find how similar or distant they are. Note that the word corpora here does not imply large amount of data. We could even get some meaningful results when both the corpora are just words, e.g. in the first step of the cognate identification problem.

Let the two corpora be C_1 and C_2. We first build (dependent) phonetic n-gram models from these, which are $\mho_1$ and $\mho_2$ such that the context are defined by the eqn. 19. Since this model does not assume independence of features, any feature can follow any feature. Once the n-grams ('centerless contexts') are compiled, we merge the n-grams of different sizes (unigram, bigrams etc.) and prune the model to retain only top G n-grams. Only after this we calculate the relative probabilities of individual n-grams for the model $\mho_1$ simply as follows:

$$p(c) = \frac{|c|}{\sum |c|} \qquad [20]$$

Note that the above probabilities are different from the standard n-gram (conditional) probabilities because no distinction is made between the n-grams of different sizes. The reason for doing this is that in this case we are going to find distributional similarity, not sequence probabilities. Standard n-gram probabilities would be required in the latter case.

Similarly, we can calculate probabilities $q(c)$ for the model $\mho_2$. Now we can calculate the distributional similarity of the two corpora or the two models by using a measure like relative entropy or KL-divergence (Sibun *et al.*, 1996):

$$\Psi(C_1, C_2) = \Psi(\mho_1, \mho_2) = \sum_{c=c_1=c_2} p(c) \frac{log\ p(c)}{log\ q(c)} \quad [21]$$

In the usual case (for problems like word origin guessing and language distance calculation), given one model, we would wish to find (from out of a set of other models) the model that maximizes the above similarity:

$$\hat{\mho} = \arg\max_i \Psi(\mho, \mho_i) \quad [22]$$

8.2. *Calculating Sequence Probabilities*

The n-gram probabilities for the dependent feature based model can be calculated in a way similar to that for standard n-grams. For feature unigrams, the equation will be:

$$p(f, v) = \frac{|(f, v)|}{\sum |(f, v)|} \quad [23]$$

For longer n-grams c_j, we will have:

$$p((f_j, v_j)|(f_{j-1}, v_{j-1})) = \frac{|c_j|}{\sum_j\ |(c_j|c_{j-1})|} \quad [24]$$

$$where\ c_j = ((f_1, v_1),\ (f_2, v_2),\ \ldots,\ (f_j, v_j)) \quad [25]$$

Now suppose we want to estimate the probability of a phoneme sequence given by:

$$S_\phi = (\phi_1, \phi_2, \ldots, \phi_s) \quad [26]$$

Since each phoneme is mapped to a set of feature-value pairs, the way to calculate the probability of the sequence will be somewhat different from that for standard (e.g. phoneme, letter or word) n-grams. We will have to take into account the probabilities of all the possible feature-value contexts or feature-value n-grams for the given phoneme sequence. Given a corpus, we can build a kind of language model of the feature-value n-grams. Then, for a phoneme sequence, the sequence probability, or more accurately, the likelihood or score of the sequence $L(S_\phi)$, given the model and assuming $n = 3$, i.e., the model is trigram model, will be:

$$p(S_\phi) \approx p(\phi_1)p(\phi_2|\phi_1)p(\phi_3|\phi_1, \phi_2)p(\phi_3|\phi_1, \phi_2)\ldots p(\phi_{|S_\phi|}|\phi_{|S_\phi|-1}, \phi_{|S_\phi|-2}) \quad [27]$$

However, estimating the terms in the above equation will be more complicated than in the case of phoneme n-grams. We will describe the method of estimation below.

Before proceeding further, it can be mentioned that the feature-value pairs of a single phoneme are also not independent, but they can be assumed to be dependent in a fixed predictable way. This is because the features can be arranged hierarchically, i.e., we know (given a language) that one feature can be given precedence over another and that this another feature will occur only if the feature above it in the hierarchy has occurred. For example, if have this non-exhaustive list of features: *consonant*, *vowel*, *manner*, *place*, *voicing*, *aspiration*, *length* and *height*, then for a language like Hindi, the following can be observed:

– *Consonant* and *vowel* have the highest precedence

– *Manner*, *place*, *voicing* and *aspiration* occur 'after' *consonant*

– *Manner* and *place* have higher precedence than *voicing* and *aspiration*

– *Length* and *height* occur 'after' *vowel*

Using this kind of information, we can prepare a precedence list of features. In this list, a feature only depends on the preceding feature. Also, two or more features can have the same precedence. This condition allows us to calculate the conditional probabilities of a feature-value pair, from which we can calculate the prior probabilities of feature-value pairs:

$$p_p(f_k, v_k) = \prod_{k\ to\ 1} p_p((f_k, v_k)|(f_{k-1}, v_{k-1})), k = precedence\ position \quad [28]$$

Such that:

$$p_p(f_1, v_1) = p(f_1, v_1) \quad [29]$$

For example, if the precendence list for a particular phoneme is *consonant*, *place*, *manner*, *voicing* and , *aspiration*, then k_1 will be *consonant* and k_5 will be *aspiration*. There can be two main precedence lists, one for the consonants and one for the vowels.

Note that $p(f_1, v_1)$ is calculated from the corpus, while $p_p(f_1, v_1)$ is the modified value we calculate above to take into account the known dependence of features within a phoneme and we use it only for calculating the probabilities of feature-value pair sequences. These dependencies are very few, quite predictable, unambiguous and easily enumerable by anyone who is selecting the features and values. That is why it may not be necessary to derive them automatically, though it is possible to do that.

The modified value of the feature-value pair probability will now be:

$$p_m(f, v) = p_p(f, v)p(f, v) \quad [30]$$

Also note that $p_m(f, v)$ is not, strictly speaking, a probability, but it can be normalized to be treated as 'probability'. However, that is not required for our purposes as we are only using it to calculate the 'sequence probability', which also (in the strict sense) is not really a probability, but an estimate of how likely a sequence is given the model.

The modified value of the conditional 'probability' for a feature-value pair bigram will be:

$$p_m((f_2, v_2)|(f_1, v_1)) = p_p(f_2, v_2)p((f_2, v_2)|(f_1, v_1))$$

It will be similar for larger n-grams, i.e., the precedence probability of only the last feature-value pair in the n-gram will be used to modify the value.

Now, if we denote the feature-value pair (f, v) as f^v, then the above can be used to calculate the probabilities of feature-value pair sequences:

$$p(S_{f^v}) \approx p_m(f_1^v)p_m(f_2^v|f_1^v)p_m(f_3^v|f_1^v, f_2^v)p_m(f_3^v|f_1^v, f_2^v)\ldots p_m(f_{|S_{f^v}|}^v|f_{|S_{f^v}|-1}^v, f_{|S_{f^v}|-2}^v) \quad [31]$$

Now, we define the probability of a phoneme n-gram as the sum of the probabilities of all the possible feature-value sequences for that n-gram.

For example, for a phoneme trigram:

$$p(\phi_3|\phi_1,\phi_2) = \sum_{1\ to\ Q} p(S^q_{f^v}) \qquad [32]$$

where Q is the number of all possible feature-value sequences for the trigram. The probabilities of phoneme n-grams, when substituted in eqn. 27 will give the probability of a phoneme sequence.

While using sequence probabilities, complexity is worth considering. In practice, this should not be a major issue in most of the cases where these models could be used because we are dealing mainly at the word level where only the length of the word can cause a problem. The number of features and the number of feature values are both small enough (less than 10). There can be a problem only when we try to enumerate all possible sequences of feature-value pairs. But as explained above, this is not required if we apply 'trigram approximation' to the sequence probability calculation.

Another question about these models could be about the requirement of large data sets. This problem would occur only to the extent it would occur with any other data oriented approach. In our case it might actually be less because working at the feature level rather than at letter or phoneme level means that there is more data and therefore relatively less data sparsity.

In the next section we present the results of using the model (via distributional similarity) for the first application.

9. Application to Phylogenetic Trees

Establishing relationships between languages which have been in contact for a long time has been a topic of interest in historical linguistics (Campbell, 2004). However, this topic has been relatively less explored in the computational community. Most of the previous work is focused on reconstruction of phylogenetic trees for a particular language family using handcrafted word lists (Gray *et al.*, 1995; Atkinson *et al.*, 2006; Nakhleh *et al.*, 2005) or using synthetic data (Barbançon *et al.*, 2007). But as Singh and Surana (Singh *et al.*, 2007a) have showed, corpus based measures can be successfully used for comparative study of languages.

Ellison and Kirby (Ellison *et al.*, 2006a) discussed establishing a probability distribution for every language through intra-lexical comparison using confusion probabilities. The distance between every language is estimated through KL-divergence and Rao's distance. The experiments are conducted on Dyen's (Dyen *et al.*, 1992a) classical Indo-European dataset. The estimated distances are used for constructing the phylogeny of the Indo-European languages.

Bouchard-Cote et al. (Bouchard-Cote *et al.*, 2007), in a novel attempt, combined the advantages of classical comparative method and the corpus-based probabilis-

tic models. The word forms are represented by phoneme sequences which undergo stochastic edits along the branches of a phylogenetic tree. The robustness of the model is proved when it selects the linguistically attested phylogeny.

In another novel attempt, Singh and Surana (Singh *et al.*, 2007a) used simple corpus based measures to show that corpus can be used for comparative study of languages. They use both character n-gram distances and Surface Similarity (Singh *et al.*, 2007b) to identify the possible cognates, which in turn are used to estimate the inter-language distance.

9.1. *Our Experiments*

In our experiments, we used the character n-gram based measure used by Singh and Surana as the baseline. Note that since the experiments are on Indian languages, which use Brahmi origin scripts, character n-gram based model is almost identical to the phoneme n-gram based model. We compare the results for character or phoneme n-grams with a model (described earlier in this paper) based on phonemic feature n-grams. For calculating language distance, we use the same distributional similarity measure as used by Singh and Surana, viz. Symmetric Cross Entropy:

$$d_{sce} = \sum_{g_l = g_m} (p(g_l)\ log\ q(g_m) + q(g_m)\ log\ p(g_l)) \qquad [33]$$

where p and q are the probability distributions for the two languages and g_l and g_m are n-grams in languages l and m, respectively.

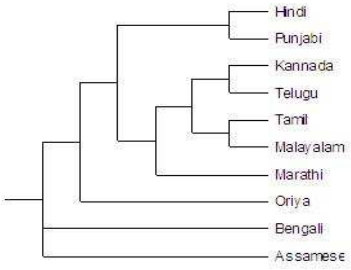

Figure 3. *Phylogenetic tree using phoneme n-grams*

The feature n-grams are computed as follows. For a given word, each letter is first converted into a vector consisting of the feature-value pairs which are mapped to it. Then, from the sequence of vectors of features, all possible sequences of features up to the length 3 (the order of the n-gram model) are computed. All these sequences of features (feature n-grams) are added to the n-gram model. Finally the model is pruned to keep only the top N_g n-grams, which is based on Cavnar's observations and the results reported in the literature for language identification.

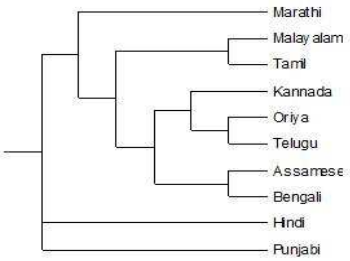

Figure 4. *Phylogenetic tree using feature-value n-grams*

9.2. *Experimental Setup*

Although the languages that we selected belong to two different language families, there are a lot of similarities among them. As argued by Emeneau (Emeneau, 1956), they all belong to the Indian *linguistic area*, i.e., an area in which languages have been so much in contact and for so long that they have influenced each other heavily, even in syntax, and even if they belong to different genetic families.

The corpora used for our experiments are part of CIIL multilingual corpus. Initially the word types with their frequencies are extracted from the corpus. Then the word types are sorted based on their corresponding frequency. Only the top N_w of these word types are retained. This is done with the aim of including as much core vocabulary as possible for comparing the languages[1].

We calculate the distance between every pair of languages available. We compare the results between the two methods (phoneme based and feature-value based) discussed above by constructing trees using these methods. The trees are constructed using the NEIGHBOR program in the PHYLIP package[2].

9.3. *Results*

Figures 3 and 4 show the trees generated by the two methods, respectively. Table 1 gives the language distances calculated using feature-value n-grams. The tree given by feature-value n-grams can be said to be better for the following reasons. It correctly positions Marathi between Hindi/Punjabi and the other languages. It also correctly groups Bengali together with Assamese. It groups Oriya with Telugu and Kannada, which is wrong in terms of genetic families, but is correct in the sense that there are a lot of similarities between Oriya and Telugu and even more between Telugu and Kannada.

1. For our experiments we fixed N_w at 50,000
2. `http://evolution.genetics.washington.edu/phylip/phylip.html`

	BN	HI	KN	ML	MR	OR	PA	TA	TE
AS	0.02	0.06	0.07	0.12	0.09	0.05	0.09	0.13	0.05
BN		0.06	0.07	0.13	0.08	0.04	0.09	0.11	0.02
HI			0.09	0.09	0.06	0.08	0.03	0.15	0.13
KN				0.10	0.09	0.02	0.08	0.10	0.03
ML					0.13	0.13	0.11	0.07	0.15
MR						0.08	0.06	0.13	0.09
OR							0.07	0.10	0.00
PA								0.14	0.07
TA									0.08
AS: Assamese, **BN:** Bengali, **HI:** Hindi, **KN:** Kannada **ML:** Malayalam, **MR:** Marathi, **OR:** Oriya, **PA:** Punjabi, **TA:** Tamil, **TE:** Telugu									

Tableau 1. *Inter-language distances among ten major South Asian languages using feature-value n-grams*

10. Application to Cognate Identification

Cognates are words of the same origin that belong to different languages. For example, the English word *beaver* and the German word *biber* are cognates descending from Proto-Germanic **bebru* and Proto-Indo-European **bher*. Identification of cognates is a major task in Historical Linguistics. Cognates usually have similar phonetic (and possibly orthographic) forms where string similarities can be used as the first step for identifying them, the second step being eliminating 'false friends'. We are not attempting the second step of the cognate identification process in this paper. We use a distributional similarity based measure for identifying the (potential) cognates.

10.1. *Related Work*

Kondrak (Kondrak, 2002b) proposed algorithms for aligning two cognates, given the phonetic transcriptions, based on phonetic feature values. The system which he calls ALINE (Kondrak, 2000) assigns a similarity score to the two strings being compared. In another paper (Kondrak, 2001), he combines semantic similarity with the phonemic similarity to identify the cognates between two languages. Bergsma et al. (Bergsma *et al.*, 2007) use character-based alignment features as an input for the discriminative classifier for classifying the word pairs as cognates or non-cognates.

10.2. *Cognate Identification*

For a given word pair, feature-value n-grams and their corresponding probabilities are estimated for each word by treating each word as small corpus and compiling

feature-value based n-gram model. For each word, all the n-grams irrespective of their sizes (unigram, bigram etc.) are merged in one vector, as mentioned in earlier sections. Now that we have two probability distributions, we can calculate how similar they are using any information theoretic or distributional similarity measure. For our experiments, we used normalized symmetric cross entropy as given in eqn. 33.

10.3. *Experimental Setup*

The data for this experiment was obtained from Dravidian Etymological Dictionary[3]. Word lists for Tamil and Malayalam were extracted from the dictionary. Only the first 500 entries in each word list were manually verified. The candidate pair set was created by generating all the possible Tamil-Malayalam word pairs. The electronic version of the dictionary was used as the gold standard. The task was to identify 329 cognate pairs out of the 250,000 candidate pairs (0.1316%).

The standard string similarity measures such as Scaled Edit Distance (SED), Longest Common Subsequence Ratio (LCSR) and the Dice measures were used as baselines for the experiment. The system was evaluated using *11-point interpolated average precision* (Manning *et al.*, 1999). The candidate pairs are reranked based on the similarity scores calculated for each candidate pair. The 11-point interpolated average precision is an information extraction evaluation technique. The precision levels are calculated for the recall levels of 0%, 10%, 20%, 30%,.....,100%, and then averaged to a single number. The precision at recall levels 0% and 100% are uniformly set at 1 and 0 respectively.

10.4. *Results*

The results for the four measures are given in the Table 2. The precision is the highest for feature-value pair based n-grams, in spite of the fact that the measure used by us is a distributional similarity measure, whereas the other three are sequence similarity measures. The results show that feature-value based model can outperform phoneme based model for certain applications.

	SED	LCSR	DICE	*Feature-Value n-Gram*
Genetic Cognates	49.32%	52.02%	51.06%	53.98%

Tableau 2. *Results for cognate identification using distributional similarity for feature-value pair based model as compared to some other sequence similarity based methods*

3. `http://dsal.uchicago.edu/cgi-bin/philologic/getobject.pl?c.0:1:3.burrow`

11. Application to Word Origin Guessing

We also conducted an experiment on word origin guessing, which is also a task that is often performed in Historical Linguistics and is closely related to the task of cognate identification. For the experiment, we prepared lists of about 150 words each of Sanskrit, Persian and English origin. All these words were either inherited (from Sanskrit) or borrowed (from Persian and English). These are three languages from which the major Indian languages have inherited or borrowed the largest number of words. The data was divided into training and testing parts roughly in the ratio 80-20 for five-fold cross validation.

We used a distributional similarity measure for phoneme n-grams as well as feature-value pair n-grams, using a method similar to the one used for cognate identification and language distance calculation, as described in earlier sections. The second measure used was 'distributional difference' (cardinality of the set difference if the two sets are sets of n-grams). The third measure was sequence probability as defined in Section 8.2 with the difference that we did not take into account the precedence probability. The sequence probability for phoneme n-grams was calculated in the usual way. The experiment was conducted with and without one extra heuristic: marking the boundary of the word with special characters such that the values of all the features for these special characters are START or END depending on whether the character marks the beginning or the end of the word.

	Without Word Boundary			**With Word Boundary**		
	Sanskrit	**Persian**	**English**	**Sanskrit**	**Persian**	**English**
Phonemes (DS)	72.00	87.33	84.00	81.33	85.33	90.67
Features (DS)	48.67	97.37	74.00	84.00	95.33	79.33
Phonemes (DD)	78.67	81.34	81.33	82.67	86.67	83.33
Features (DD)	64.00	90.67	74.65	84.00	86.67	76.00
Phonemes (SP)	52.00	66.00	78.67	58.00	68.67	85.33
Features (SP)	47.33	68.00	77.33	50.67	70.67	86.00

DS: Distributional Similarity
DD: Distributional Difference
SP: Sequence Probability

Tableau 3. *Results for experiments on word origin guessing using distributional similarity, distributional difference and sequence probability for feature-value pair n-grams as compared to corresponding phoneme n-grams based measures. The numbers indicate the precision with which the origin of Hindi and/or Telugu words was correctly identified.*

The results are shown in Table 3. Without the word boundary heuristic, phoneme n-grams consistently perform better. However, when the word boundary is marked, there is a significant change in the results so that feature n-grams are better in five case and the same in one. Distributional similarity seems to be somewhat better than distributional difference. Sequence probability does not perform as well as expected

for Sanskrit and Persian. This could be due to the way it is calculated or it might be due to the lack of the right smoothing technique or tuning. It will be worth exploring whether assigning proper feature weights while calculating the sequence probability can improve the results.

Errors are mostly of three kinds. The first are due to word length being too small (4 or less), i.e., data sparsity on the testing side. The second are due to inherent ambiguity as many words are phonetically valid in more than one languages. The third category is of 'genuine' errors which could perhaps be eliminated with better tuning of the current techniques.

12. Studying Phonetic Variation and Change

We have also conducted a pilot study on using the model in terms of 'centered contexts' for studying phonetic variation and change. One of the tasks in Historical Linguistics and in the study of synchronic as well diachronic phonetic variation is to identify patterns of phonetic variation in terms of articulatory features (Hock, 1991; Campbell, 2004). These patterns can be useful for descriptive linguistics, but are also used to arrive at laws of sound change.

In this section we describe a simple method using the phonetic contexts as defined in Sections 5 and 6 to make the task of identifying such patterns of variation easier. For our experiments we consider only contexts of length 1 to 3, which usually account for the most common patterns[4]. Note that the total maximum context length is 3, which means that the maximum size of left and right contexts is only 1.

The probability of a context representing a pattern or being an instance of a pattern of (regular) variation can be defined, as a first step, simply as:

$$p(f_t|f_s,\ c_s^{\lambda},\ c_s^{\rho}) = \frac{C(f_t,\ f_s,\ c_s^{\lambda},\ c_s^{\rho})}{C(f_s,\ c_s^{\lambda},\ c_s^{\rho})} \qquad [34]$$

Even in this constrained situation, there can be four kinds of patterns: $p(f_t|f_s,\ c_s^{\lambda},\ c_s^{\rho})$ (the left as well as the right context is present), $p(f_t|f_s,\ c_s^{\lambda})$ (only the left context is present), $p(f_t|f_s,\ c_s^{\rho})$ (only the right context is present), $p(f_t|f_s)$ (neither the left nor the right context is present).

We prepared an Avadhi-Hindi small parallel word list (217 word pairs) and compiled a feature based model from it. To extract the patterns, we first need to align the phonemes within the words. For this we used a Dynamic Time Warping based algorithm where each node (phoneme) in the trellis represents a vector of feature-value pairs for the phoneme. Alignment is on the basis of features (Singh, 2006). Then,

4. We do not present any empirical evidence for this claim about the context length, but it may be noted that most of the laws of sound change are in terms of contexts of maximum length 3.

using the above equation, we tried to identify the patterns of variation that have a relatively high count (5 or more) and a high probability (more than 0.85). Of course, only those patterns are retained where the aligned center phonemes (whose contexts we are looking at) are not the same since we are interested in variation.

Although we have not performed proper evaluation for this task, but using just the simple method described above, we were able to extract patterns such that many of them were correct patterns of variation as manually checked. Some example of these patterns (Avadhi to Hindi) are given below:

```
[length=short] > [length=long]
          as in {\em d(u)sara} to {\em d(U)sarA}
[length=short] > [length=long] / [+voiced]
          as in {\em d(u)sara} to {\em d(U)sarA}
[length=short] > [length=long] / [type=vowel]
          as in {\em deKA(i)} to {\em diKA(I)}
[length=short] > [length=long] / [+voiced]_[type=consonant]
          as in {\em d(e)vAra} to {\em d(I)vAra}
[manner=trill] > [manner=approximant] / [type=vowel]
          as in {\em bAda(r)a} to {\em bAda(l)a}
[manner=trill] > [manner=approximant] / [+voiced]
          as in {\em bAda(r)a} to {\em bAda(l)a}
[manner=trill] > [manner=approximant] / [type=vowel]_[+voiced]
          as in {\em uje(r)e} to {\em ujA(l)e}
[place=alveolar] > [place=post-alveolar] / _[type=vowel]
          as in {\em di(s)i} to {\em di(S)A}
[place=alveolar] > [place=post-alveolar] / _[+voiced]
          as in {\em di(s)i} to {\em di(S)A}
[place=alveolar] > [place=post-alveolar] / [+voiced]_[+voiced]
          as in {\em di(s)i} to {\em di(S)A}
```

The notation followed above is the same as that used for laws of sound change, i.e., A > B / X_Y means that A changes into B when preceded by X and followed by Y. Though we have presented the variation patterns from Avadhi to Hindi, the direction can be easily reversed.

Apart from the fact that there are many false positives (most of them because of wrong alignment of phonemes), there are some other problems with this simple technique. As can be observed from above examples, many patterns are highly overlapping. How to derive generalizations from such patterns can be a topic for further research.

13. Point Lattice Representation of the Phonetic Space

Coming back to the modeling of the phonetic space, we observe that the phonetic space can be seen as made up of discrete points. This space is three dimensional. The

first axis is formed by the features and the second by their values. Both these are of finite length as the number of features and their values is assumed to be finite in our model. The third axis extends indefinitely and on it can be situated the phonemes. Thus, the phonetic space can be described as a constrained point lattice that extends indefinitely along the phoneme axis. Given such a space, a specific phoneme can be represented as a two dimensional polygon along the feature-value axes. Following the same logic and applying the feature precedence condition, a phoneme sequence or a word can be represented in the point lattice as a three dimensional surface.

If we represent words in the phonetic space as describe above, then comparing words or phoneme sequences would be equivalent to comparing three dimensional surfaces in constrained point lattices. We might be able to use techniques from three dimensional geometry and matrix theory for this purpose. And if we have a 'parallel corpus' of words from different languages or dialects, we can try to induce the rules of phonetic change or variation as attempted in the previous section. Also, by comparing the three dimensional surfaces, we could calculate the Surface Similarity (Singh *et al.*, 2007b), i.e., phonetic and orthographic similarity of two words. Figures 1 and 2 point to the way the point lattice representation can be visualized, except that the feature value dimension is not shown as it is vertical to the paper.

One possibility of confusion here is with regard to the dimension representing features because it might seem that there is no preferred linear ordering of features. Although the features dimension is nominal, the the surfaces can still have a useful meaning because we just have to select one particular feature ordering and use it consistently. As long as this ordering is consistent, we can still calculate Euclidean distances. Also, as explained in Section 8.2, feature do have a precedence order. Features as used by us can be arranged in a hierarchy, which gives the features higher in the hierarchy greater precedence. Features which are at the same level can be arranged in any order, but this order can be made consistent for the purposes of point lattice representation. Moreover, features can be assigned weights according to the order of precedence. This would mean that the features dimension need not be just nominal.

14. Conclusion

As phonetic processing of text is now being resorted to for diverse purposes, we presented a computational model of the phonetic space that can partly serve as a common ground for such processing. We started with a model defined in terms of phonemes and then moved on to two models defined in terms of articulatory features, first with the independence assumption and then without it. We used this model for four applications, namely constructing phylogenetic trees of languages, cognate identification, word origin guessing and studying phonetic variation and change, with results comparable to or better than the state of the art. We also discussed how a word can be represented as a three dimensional surface in a constrained point lattice representing the phonetic space.

The work presented here points to the possibility of future work in several directions. The first is improving the general model of Section 6 so that the exact nature of dependence of features is also modeled adequately. The second is improving the estimation of sequence probability so that it consistently works better than distributional similarity for the relevant applications. The third is extending the initial work on phonetic variation presented in Section 12. There might also be considerable scope for using the point lattice representation of the phonetic space for applications like inducing the rules of phonetic change and variation.

15. Bibliographie

AbdulJaleel N., Larkey L. S., « Statistical transliteration for english-arabic cross language information retrieval », *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, New York, NY, USA, p. 139-146, 2003.

Atkinson Q., Gray R., « How old is the Indo-European language family? Progress or more moths to the flame », *Phylogenetic Methods and the Prehistory of Languages (Forster P, Renfrew C, eds)*p. 91-109, 2006.

Barbançon F., Warnow T., Evans S., Ringe D., Nakhleh L., An experimental study comparing linguistic phylogenetic reconstruction methods, Technical report, Technical Report 732, Department of Statistics, University of California, Berkeley, 2007.

Bartlett S., Kondrak G., Cherry C., « Automatic Syllabification with Structured SVMs for Letter-to-Phoneme Conversion », *Proceedings of ACL-08: HLT*, ACL, Columbus, Ohio, p. 568-576, June, 2008.

Bergsma S., Kondrak G., « Alignment-Based Discriminative String Similarity », *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, Prague, Czech Republic, p. 656-663, June, 2007.

Bouchard-Cote A., Liang P., Griffiths T., Klein D., « A Probabilistic Approach to Diachronic Phonology », 2007.

Campbell L., *Historical linguistics: an introduction*, MIT Press, 2004.

Cavnar W., Trenkle J., « N-gram-based text categorization », *Ann Arbor MI*, vol. 48113, p. 4001, 1994.

Chen D. S., Liu Z. K., « A Novel Approach to Detect and Correct Highlighted Face Region in Color Image », *AVSS '03: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, IEEE Computer Society, Washington, DC, USA, p. 7, 2003.

Chomsky N., Halle M., *Sound Pattern of English*, Harper & Row, 1968.

Damper R., Marchand Y., Marseters J., Bazin A., « Aligning Letters and Phonemes for Speech Synthesis », *Fifth ISCA Workshop on Speech Synthesis*, ISCA, 2004.

Dyen I., Kruskal J., Black P., « An Indoeuropean classification: a lexicostatistical experiment », *American Philosophical Society*, 1992a.

Dyen I., Kruskal J., Black P., « An Indo-European classification: A lexicostatistical experiment », *Transactions of the American Philosophical Society, 82:1-132*, 1992b.

Ellison T., Kirby S., « Measuring language divergence by intra-lexical comparison », *Proceeding of the 21st ACL Conference*, Morristown, NJ, USA, p. 273-280, 2006a.

Ellison T. M., Kirby S., « Measuring Language Divergence by Intra-Lexical Comparison », *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Sydney, Australia, 2006b.

Emeneau M., « India as a Lingustic Area », *Language*p. 3-16, 1956.

Gray R., Atkinson Q., « Language-tree divergence times support the Anatolian theory of Indo-European origin », *Earth Planet. Sci*, vol. 23, p. 41-63, 1995.

Haizhou L., Min Z., Jian S., « A joint source-channel model for machine transliteration », *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, p. 159, 2004.

Hardcastle W. J., Laver J., *The Handbook of Phonetic Sciences*, Wiley-Blackwell, 1999.

Hock H. H., *Principles of historical linguistics*, Mouton de Gruyter, Berlin, 1991.

Knight K., Graehl J., « Machine transliteration », *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, p. 128-135, 1997.

Kondrak G., « A new algorithm for the alignment of phonetic sequences », *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, p. 288-295, 2000.

Kondrak G., « Identifying cognates by phonetic and semantic similarity », *North American Chapter Of The Association For Computational Linguistics*, Association for Computational Linguistics Morristown, NJ, USA, p. 1-8, 2001.

Kondrak G., Algorithms for language reconstruction, PhD thesis, 2002a. Adviser-Graeme Hirst.

Kondrak G., *Algorithms for language reconstruction*, University of Toronto, Ont., Canada, 2002b.

Lucey S., Sridharan S., Chandran V., « Adaptive mouth segmentation using chromatic features », *Pattern Recognition Letters*, vol. 23, n° 11, p. 1293 - 1302, 2002.

Manning C., Schutze H., « Foundations of Natural Language Processing », 1999.

Marchand Y., Damper R., « A Multistrategy Approach to Improving Pronunciation by Analogy », *Computational Linguistics*, vol. 26, n° 2, p. 195-219, 2000.

Nakhleh L., Ringe D., Warnow T., « Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages », *Language*, vol. 81, n° 2, p. 382-420, 2005.

Nerbonne J., Heeringa W., « Measuring dialect distance phonetically », *Proceedings of SIGPHON-97: 3rd Meeting of the ACL Special Interest Group in Computational Phonology*, 1997.

Patricia Keating M. H., Jackson E., « Vowel allophones and the vowel-formant phonetic space », *Journal of Acoustical Society of America*, vol. 74, n° S1, p. S90, November, 1983.

Port B., Leary A., « Against formal phonology », p. 81(4):927-964, 2005.

Ribeiro A., Dias G., Lopes G., Mexia J., « Cognates alignment », *Machine Translation Summit VIII, Machine Translation in The Information Age*p. 287-292, 2001.

Saunderson J. L., Milner B. I., « Modified chromatic value color space », *Journal of the Optical Society of America*, vol. 36, n° 1, p. 36, 1946.

Sibun P., Reynar J. C., « Language Identification: Examining the Issues », *In Proceedings of SDAIR-96, the 5th Symposium on Document Analysis and Information Retrieval.*, p. 125-135, 1996.

Singh A. K., « A Computational Phonetic Model for Indian Language Scripts », *Proceedings of the Constraints on Spelling Changes: Fifth International Workshop on Writing Systems*, Nijmegen, The Netherlands, 2006.

Singh A. K., Surana H., « Can Corpus Based Measures be Used for Comparative Study of Languages? », *Proceedings of the Ninth Meeting of ACL Special Interest Group on Computational Phonology and Morphology*, Association for Computational Linguistics, Prague, Czech Republic, 2007a.

Singh A. K., Surana H., « Using a Single Framework for Computational Modeling of Linguistic Similarity for Solving Many NLP Problems », *Proceedings of Eurolan Doctoral Consortium*, Iasi, Romania, 2007b.

Swadesh M., « Lexico-dating of prehistoric ethnic contacts », *Proceedings of the American philosophical society, 96(4)*, 1952.

van den Bosch A., Canisius S., « Improved morpho-phonological sequence processing with constraint satisfaction inference », *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL*, p. 41-49, 2006.

Virga P., Khudanpur S., « Transliteration of proper names in cross-lingual information retrieval », *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition*, Morristown, NJ, USA, p. 57-64, 2003.

Wolfram W., Schilling-Estes N., *American English: dialects and variation*, Wiley-Blackwell, 1998.

Yang J., Waibel A., « A real-time face tracker », *Proceeding of WACV'96*, Sarasota, FL, USA, p. 142?147, 1996.

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél. : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com

ANNEXE POUR LE SERVICE FABRICATION

A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER
DE LEUR ARTICLE ET LE COPYRIGHT SIGNE PAR COURRIER
LE FICHIER PDF CORRESPONDANT SERA ENVOYE PAR E-MAIL

1. ARTICLE POUR LA REVUE :
   L'objet. Volume 8 – n°2/2005

2. AUTEURS :
   Anil Kumar Singh[*] *— Taraka Rama*[*] *— Pradeep Dasigi*[*]

3. TITRE DE L'ARTICLE :
   A Computational Model of the Phonetic Space and Its Applications

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :
   Modeling the Phonetic Space

5. DATE DE CETTE VERSION :
   22 octobre 2009

6. COORDONNÉES DES AUTEURS :

– adresse postale :

[*] Language Technologies Research Centre, IIIT, Hyderabad, India

{anil@research, taraka@students}.iiit.ac.in, pradeep.dasigi@gmail.com

– téléphone : 00 00 00 00 00

– télécopie : 00 00 00 00 00

– e-mail : guillaume.laurent@ens2m.fr

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :
   LaTeX, avec le fichier de style `article-hermes.cls`,
   version 1.23 du 17/11/2005.

8. FORMULAIRE DE COPYRIGHT :
   Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :
   `http://www.revuesonline.com`

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél. : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com