

# Structural complexity of phonological systems

*Christophe Coupé, Egidio Marsico and François Pellegrino*

## 1. Introduction

In the linguistic tradition, including phonology, complexity has often been evoked when looking for explanatory arguments (a given phenomenon is rarer because it is more complex than another) or looking for a balance of complexity within subsystems of a language, or directly comparing and ranking several languages according to a linguistic dimension (for a review, see Chitoran and Cohn, and Pellegrino et al. this volume). In this perspective, the concept of complexity is intrinsically relative and necessarily yields to judging something as more or less complex than something else regarding a particular property or even globally. Thus, anyone involved in the enterprise of evaluating phonological complexity faces the tricky issue of, as a first step, defining a set of (phonological) properties and for each property defining a scale of complexity. Then, and only then, is one able to start comparing the phonological complexity of the chosen phonological elements. In that perspective, “to be complex” or not is a (possibly gradient) quality assigned to a particular set of elements. This task is all but straightforward. If choosing the set of properties can be quite simple, characterising them with a scale of complexity is much trickier. Moreover, as one tries to combine several properties of an element to evaluate its overall complexity, the issue of weighting these different dimensions can easily lead to a dead end. In this regard, Maddieson (this volume), is very insightful and presents an excellent summary of where to find and how to define phonological complexity, as well as the limits of this notion.

Interestingly, an alternative conception of complexity has developed for a half-century, stemming from cybernetics, systems theory and systems dynamics (e.g. Abraham, 2001 for an epistemological view). First found in statistical physics, biology and computer science, it has rapidly proven relevant within the field of humanities and social sciences and it is now definitively associated with the notion of “complex system”. In that framework, a system **is or is not** complex according to whether its structure and behaviour satisfies particular characteristics. The picture is thus substan-

tially different from the “arithmetic” view of complexity because one no longer needs to look for the very dimensions on which to compute complexity, more or less objectively, but rather to “just” validate if a system fulfils some properties known *a priori*. This way, complexity is no longer a relative notion. To illustrate what the properties of complex systems can be, we refer to Steels (1997):

“a complex system consists of a set of interacting elements where the behaviour of the total is an indirect, non-hierarchical consequence of the behaviour of the different parts [...]. In complex systems, global coherence is reached despite purely local nonlinear interactions. There is no central control source”.

A system can thus be said to be complex if:

- i) it is structured in different levels;
- ii) the properties of the global level (the systemic ones) differ from those of the elements of the basic level;
- iii) the systemic properties cannot be derived linearly from the basic ones.

Seeds of this new paradigm can be found in Warren Weaver’s seminal article (1948) where he emphasized the understanding of **organized complexity** as one of the key issues to be addressed by modern science. Lying somewhere between the simple problems that could be solved with pre-20<sup>th</sup> century science, and the disorganized complexity that was handled with new statistical and probabilistic tools in the first half of the 20<sup>th</sup> century, this complexity involves dealing with a number of factors that do not behave independently, but interact into what Weaver called an “*organic whole*”. Rather than the basic number of factors or constituents in the system (that would be low for simple problems and potentially high in the case of disorganized complexity), it is the nature of their interrelations that actually matters. The step forward lies precisely in this differentiation of levels where the elements and the properties of each level may differ; and what matters is the way the structure of the systemic level emerges from interactions at the basic level.

This view, stemming from the science of complex systems, leads to modifying the way phonological complexity is addressed: we no longer intend to compare the overall complexity of phonological systems in terms of which one is more complex than the others. Instead, we aim at characterising their structure. Explaining why there are so many different structures seems now even more crucial than knowing if one is more complex than another. After all, all languages seem to work with the same efficiency; no

one has ever reported a language with communicative disabilities or non-impaired children failing to learn a particular language. For all we know, all languages are functionally equal (and all complex enough), and yet as Ferguson (1978, p. 9) wrote:

“As soon as human beings began to make systematic observations about one another's languages, they were probably impressed by the paradox that all languages are in some fundamental sense one and the same, and yet they are also strikingly different from one another”.

Indeed, typological research has shown that despite the fact that certain types of linguistic structures are clearly more frequent than others, even the uncommon ones can be relatively numerous and very different. Thus, this coexistence of numerous viable types of linguistic elements and structures, although unevenly distributed, reveals that language is a system poorly constrained, or at least presenting numerous degrees of freedom.

In this contribution, we develop a study of the structural complexity of the phonological systems of the UPSID database<sup>1</sup> in line with these statements. To set the stage, let us just look at the variation present in the languages of the database. They have from 11 to 141 segments, from 3 to 28 vowels, from 6 to 95 consonants, from 6 to 51 distinctive features. This has to do with the variations of types, but discrepancies are even wider when one looks at tokens. To give a few examples, some segments are present in only one language whereas others can cover up to 90% of the sample; only one language has 28 vowels but more than 20% have five; stop consonants are present in all languages, etc. These two sources of data (types and tokens) offer different kinds of information. Looking at types raises issues regarding the set of possible phonological elements (be they features, segments or systems), and at first glance, the observed diversity could push toward considering phonological systems as simple sets of unorganized segments. However, when compared to the theoretical number of possible combinations of features and segments, the number of attested types is relatively low, showing instead that phonological systems are not randomly composed. Moreover, when looking at tokens, the uneven distribution of types among languages reveals that some systems prevail. Consequently, we need to understand what parameters are making one system more widespread than another. This means also, from a methodological point of view, that frequencies of distribution are not an explanation *per se* and thus should not be considered as inputs in a model, but rather as what is to be explained. They are the emergent properties of an underlyingly organized structure.

The notion of 'emergence' is a key concept of the dynamical complex system framework. As mentioned before, the different structures of the systems are considered as emerging from the specific interactions of their elementary units. To some extent then a system can be seen as the reflection of the constraints at work. The citation below from Björn Lindblom, illustrates this from the diachronic perspective:

"The new form [i.e. the new pronunciation that yields a potential sound change] gets tested implicitly on a number of dimensions: 'articulatory ease', 'perceptual adequacy', 'social value' and 'systemic compatibility'. If the change facilitates articulation and perception, carries social prestige and conforms with lexical and phonological structure, its probability of acceptance goes up. If the change violates the criteria, it is likely to be rejected." (Lindblom, 1998:245).

Again, the notion of "*systemic compatibility*" pushes forward the idea that the whole (the system) is more than the sum of its parts. Following this line of thinking, in a previous paper (Marsico et al., 2004), we have explored phonological inventories (hereafter PI) assuming that it can lead to the (even partial) understanding of their structure. We began with a bottom-up approach where we intended to i) set the different levels of structuration of PI, ii) identify the properties of each level and iii) characterize the relation(s) between the levels. We got reasonable results as far as points (i) and (ii) are concerned, but our approach showed its limits with point (iii), especially when dealing with the systemic level. The main index we used to monitor the systemic behaviour of PI deals with the notion of redundancy. We wanted to evaluate the longstanding idea of PI as being economic systems (i.e. the MUAF principle – Maximal Use of Available Features – first introduced by Ohala, 1980). Our redundancy measure evaluates the average distance between each segment of a PI and its nearest neighbour. Although the quantitative results seem to show that PI are indeed based on a principle of economy favoring systems with minimal phonological oppositions (i.e. based on only one feature); the qualitative analysis of these results revealed that our measure is not really a systemic one. As a matter of fact, our redundancy index deals more with one-to-one relationship between segments than with a collective behaviour. The lowest redundancy index is obtained as long as each segment has its minimal counterpart in the system (i.e. a segment differing by only one feature) without considering any of the underlying systemic principles on which MUAF is based: maximal use of features, consistent series of segments. The lowest index can be obtained

with a system made of what Lindblom calls "*a collection of 'assorted bonbons'*", (Lindblom, 1998:250).

This has led us to change our perspective and to adopt a top-down approach directly based on the systemic level. We will develop this approach in the remainder of this paper. Section 2 deals with a structural approach where PIs are considered as networks of connected phonemes. In Section 3, PIs are modelled by considering the distribution of co-occurrences of phonemes, in order to define attraction and repelling relations between them. These relations are then used to propose a synchronic measure of coherence for the phonological systems, and then diachronically extended to a measure of stability.

## **2. Considering phonological inventories in the light of graph theory**

### 2.1. From a feature-based distance to phonological graphs

#### 2.1.1. *About graph theory*

Mathematical graph theory, also named network theory, during the last decade has had a significant impact in various scientific fields, for two main reasons. The first is the acknowledgment of the range of this theory, which proposes a set of tools and generic concepts that can be applied to a wide range of questions. The second is linked to the theoretical progress made in the understanding of the properties of networks half-way between regular and random networks (e.g. Erdős and Rényi, 1960). If the detailed analyses of these two specific kinds of networks go back several decades, those of intermediate networks are much more recent, and illustrate the difficulty of apprehending Weaver's "organized complexity". The small-world or scale-free networks are by far the most cited today (e. g. Watts and Strogatz, 1998), since they are very commonly encountered in the study of the non-living, living or social phenomena. Several concepts borrowed from graph theory – like the notions of shortest path, robustness, aggregation, hub, or resilience, etc. have led to substantial breakthroughs in a wide range of applications: the functionality and robustness of internet networks, the understanding of the interactions between proteins, or within complex eco-systems, or the propagation of epidemics (Dorogotsev & Mendes, 2001; Pastor-Satorras & Vespignani, 2001). This statement is also correct in linguistics, where scientists have studied the properties of lexical,

syllabic or phonological graphs (Cancho & Solé, 2001; Cancho & al., 2004, Dorogotsev & Mendes, 2001; Solé, 2004).

In general, a graph is defined by a set of nodes and a set of connections. The way nodes are connected leads potentially to graphs of different types but for which a common set of properties may be calculated. While some of these properties depend on the size of the network, others may be invariant, for a given type of network, and regardless of their respective size. In our approach, each phonological system is considered as a set of two networks, one for the vowel segments and one for the consonants (diphthongs are not considered so far). For each graph, the segments are the nodes and the connections are derived from phonetic-phonological relations between them using the algorithm described in the next section.

### 2.1.2. From phonemes to feature-based phonemic distances

One way of quantifying the relation between any two phonemes is to rely on the features that compose them. In this approach, the degree of interaction corresponds to the distance in terms of features, where these features are compared within the natural classes they belong to. The following examples will illustrate this calculation.

	<i>/i/</i>	<i>/u/</i>
<b>Height</b>	<i>high</i>	<i>high</i>
<b>Backness</b>	<i>front</i>	<i>back</i>
<b>Lip Rounding</b>	<i>unrounded</i>	<i>rounded</i>

→ Distance = 2

	<i>/o/</i>	<i>/õ:/</i>
<b>Height</b>	<i>high-mid</i>	<i>high-mid</i>
<b>Backness</b>	<i>back</i>	<i>back</i>
<b>Lip Rounding</b>	<i>rounded</i>	<i>rounded</i>
<b>Length</b>	<i>short</i>	<i>long</i>
<b>Nasality</b>	<i>oral</i>	<i>nasal</i>

→ Distance = 2

	<i>/p/</i>	<i>/v/</i>
<b>Place</b>	<i>labial</i>	<i>labio-dental</i>
<b>Manner</b>	<i>plosive</i>	<i>fricative</i>
<b>Voicing</b>	<i>unvoiced</i>	<i>voiced</i>

→ Distance = 3

This rough distance could certainly be refined by taking the shared features into account as well, but the main problem would remain: the nature of the relation between phonemes is hard to establish *a priori*, first because of the lack of common ground for their internal description (see the first part of this volume, on the phonological primitives) and second because the principles explaining the relations between phonological segments is still a controversial and open issue. The proposed methodology by no means aims at being the ultimate formalism but it provides a reasonably adequate balance between the need for some phonetic rationale and the possibility of being consistently applied to any phonological system.

### 2.1.3. Construction of phonological graphs

With a quantification of the distance between phonemes, we can now turn to the construction of a graph where all the phonemes of a PI would be connected and fulfil the following principles:

- 1) There must be a path between any two phonemes, direct or indirect;
- 2) This path must be minimal in a way compatible with the notion of economy or parsimony.

The first principle is consistent with the promoted view of PI as systems; no phoneme is isolated within a PI, and consequently each phoneme is at least related to one of the other phonemes of the system. This principle stems from the traditional idea of opposition between phonemes. For each phoneme, the second principle aims at selecting the connections occurring within its neighbourhood (in terms of phonetic similarity) since we consider that long-range connections are meaningless. The neighbourhood is not defined using an *a priori* distance (for instance a hard threshold of 3 between segments) but by selecting the path that preserves a minimal cost as illustrated below<sup>2</sup>.

Let us concentrate on the potential paths linking /o:/ and /a/ in the five vowel system given in Figure 1. The direct path (based on the feature distance between the two phonemes) is 4. Besides, there are several indirect paths, such as for example /o:/ => /e:/ => /a/ or /o:/ => /u/ => /a/ or even /o:/ => /u/ => /i/ => /a/. This last one is especially interesting because the biggest “jump” between two nodes only involves a distance of 2 (in fact all the jumps in this path are of a distance of 2). In our approach, this path is then the less costly, since, step by step, it involves skipping from neighbour to neighbour. In this view, the number of jumps doesn’t matter, what counts

is their size. For this reason, the direct path (/o:/ => /a/, distance = 4) has been removed from the network in favor of the indirect one.

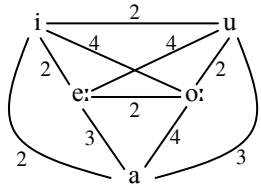
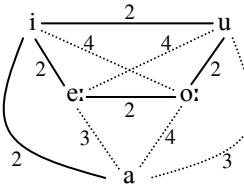
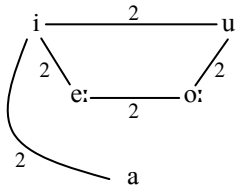
STEP 1	STEP 2	STEP 3
<p>We compute the <u>direct</u> phonetic distance for each phonemes pair.</p>	<p>Identification of pairs of phonemes for which an <u>indirect</u> path requires smaller "jumps" than the direct one.</p>	<p>Suppression of costly <u>direct</u> paths.</p>
		
<p><i>Each node is linked with every other nodes of the network. The values correspond to the phonetic distances.</i></p>	<p><i>Dotted lines show costly paths.</i></p>	<p><i>The resulting network only keeps the less costly connections (direct or indirect).</i></p>

Figure 1. Description of the algorithm of construction of phonological graphs

The inspection of the various networks or graphs built with this approach reveals properties close to the classical ones in phonology in terms of serial or derivative structures. The next figure illustrates this point with the most frequent five vowel system in our data on the left, and a ten vowel system on the right, composed with the same five vowels plus their nasal counterparts. A layered structure is visible: the sub-network consisting of the vowels /i, e, a, o, u/ mirrors the one composed of the nasal counterparts, and is connected to it in a regular fashion.



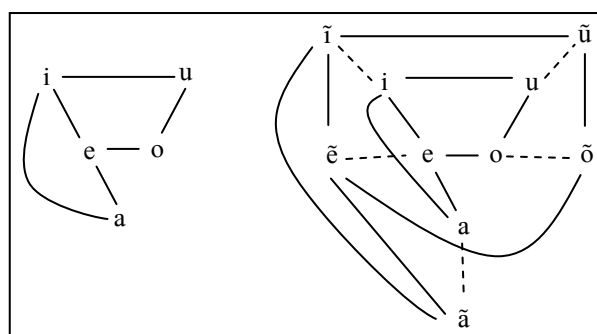


Figure 2. Two examples of phonological graphs

## 2.2. Measuring the structural complexity of phonological inventories

The previous step explained how we have built phonological graphs from PIs; we will now show how we can compare these PIs in terms of structural complexity using a specific measure relying on the corresponding graphs. The interest of this approach lies in the fact that it is anchored outside phonology and linguistics, and is not the result of an *ad hoc* measure based on language comparison. As such, this approach can make possible a comparison of PIs' structural complexity with as little a priori bias as possible. However, estimating the complexity of a graph is not a straightforward matter. Several measures exist (Neel & Orrison, 2006; Jukna, 2006; Bonchev & Buck, 2005) and as always, choosing one over the others seems to depend on implicit considerations.

### 2.2.1. The notion of “off-diagonal complexity”

Among the various possible measures found in the literature, our choice fell on the *off-diagonal complexity* proposed by Claussen (2004). This measure offers different characteristics that parallel simple intuitions linguists have on PIs. Indeed, this measure:

- Doesn't explicitly take into account graph size (i.e. its number of nodes or connections). This method consequently does not postulate that a large PI will be more complex than a small one;
- Is sensitive to the presence of hierarchical sub-structures in the network. This can happen for example when a whole primary system is con-

trasted by a secondary feature (see above the ten vowel system in figure 2, or below, the system of Chipewyan);

- Is minimal for regular graphs and maximal for free-scale graphs. It thus provides a benchmarking for PIs' structural complexity (for which free-scale structure is very unlikely).

The calculation of the offdiagonal complexity follows several steps:

1. Calculation of the degree of each node by counting its connections;
2. Construction of a matrix M defined by  $M(k_1, k_2)$ =number of connections existing between nodes of  $k_1$  degree and nodes of  $k_2$  degree;
3. Calculation of the entropy *C* of the distribution of normalized sums,  $m_i$ , of the values of the minor diagonals of M with the following formula:

$$C = - \sum_{i=0}^{k_{\max}} m_i \log m_i ; k_{\max} \text{ being the max degree of a phoneme of the graph.}$$

Such a measure can seem complicated, but it is actually able to detect the structural regularities existing at the level of the relations between nodes. Figure 3 gives an example.

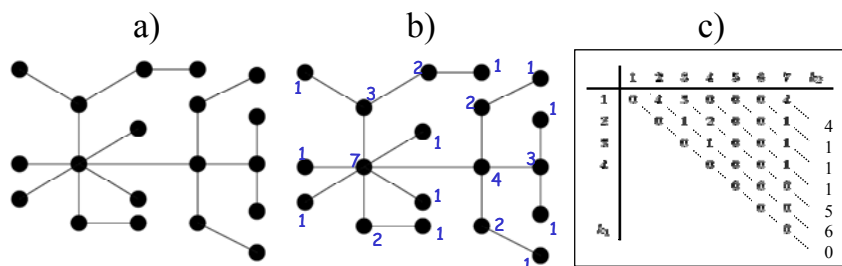


Figure 3. The different steps of calculation of the offdiagonal complexity.

In (a) we have the initial graph with 19 nodes and 18 connections. In (b) we added the degree for each node. (c) presents the corresponding matrix M and the sum of the values of the diagonal. The resulting offdiagonal complexity is thus:

$$C = - \left( \frac{4}{18} \log \frac{4}{18} + \frac{1}{18} \log \frac{1}{18} + \frac{1}{18} \log \frac{1}{18} + \frac{1}{18} \log \frac{1}{18} + \frac{1}{18} \log \frac{1}{18} + \frac{5}{18} \log \frac{5}{18} + \frac{6}{18} \log \frac{6}{18} + 0 \right) = 1.538$$

As the preceding graphs have shown, offdiagonal complexity can only be calculated for non-valued graphs, i.e. graphs where the connections have no intrinsic value or weight. This is a serious limitation since, in our approach, the connections stand for distances between phonemes and are thus naturally valued. Because the Claussen measure doesn't allow taking this information into account, the only use we make of it is when pruning the full graph by removing the costly connections.

Figure 4 gives the offdiagonal complexity of several relatively simple PIs whereas Figure 5 illustrates the possibility to do so with a much more complicated system.

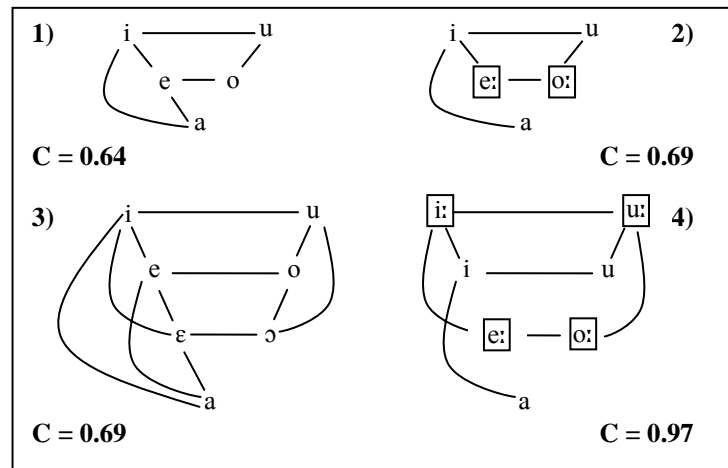


Figure 4. Simple vowel systems and values of their offdiagonal complexity.

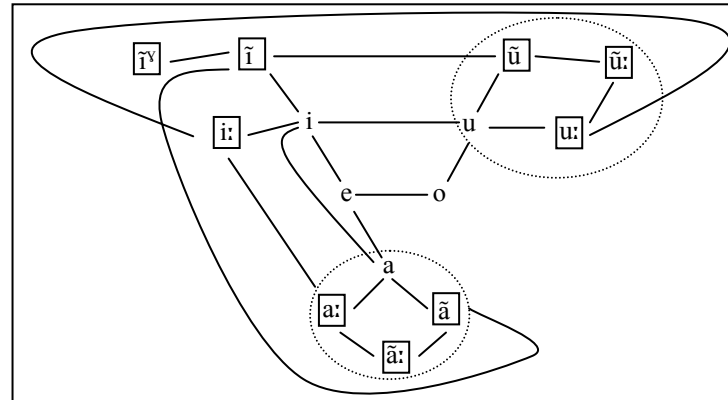


Figure 5. Off-diagonal complexity of the Chipewyan 14 vowel system; C=0.89.

These few examples are clear evidence of the absence of a direct relation between number of phonemes and value of the offdiagonal complexity. Figure 4 compares systems of the same size (4.1 vs. 4.2 for 5-vowel systems, and 4.3 vs. 4.4 for 7-vowel system) but with different complexity, and systems with the same complexity but different cardinal (4.2 vs. 4.3). The PI in 4.1 with only primary phonemes (or “basic” ones, according to Marsico et al. (2004) is less complex than the one in 4.2 with the same number of vowels but with a secondary non contrastive feature (length). This latter system being as complex as the PI in 4.3 which has two more vowels but all primary ones. Chipewyan (figure 5) presents a smaller complexity than the PI in 4.4 despite having twice as many vowels, due to its more regular structure.

### 2.3. Comparisons between phonological inventories from UPSID and random ones regarding off-diagonal complexity

The table below gives the offdiagonal complexity of vocalic and consonantal systems<sup>3</sup> for the whole set of languages from UPSID.

UPSID	Vowel systems	Consonant systems
C mean	0.794	1.670
C min	0	0
C max	1.700	2.379
Std. Dev.	0.313	0.325

No correlation between vocalic and consonantal complexity was found ( $r^2=0.0006$ ). Thus, there is no compensation between structural complexity of vowels and consonants (no negative correlation); nor any parallel behaviour between them, like the smaller the one the smaller the other (no positive correlation). These results confirm, with a different measure, the ones described in Maddieson (2006).

To assess whether the offdiagonal complexity was really capturing meaningful information on PIs, we compared the 451 UPSID PIs with a set of 451 generated PIs. These PIs were randomly composed by picking phonemes (from the whole set of existing phonemes) respecting the distribution of PI size from UPSID. Thus, every UPSID system was matched with a random one of the same size, but for which the content did not obey any linguistic motivation. Our hypothesis was that if random and actual systems lead to similar distribution of structural complexity, the offdiagonal com-

plexity is pointless. The table below gives the distribution of random systems and is to be compared with the previous one.

RANDOM	Vowel systems	Consonant systems
C mean	1.071	1.965
C min	0	1.045
C max	2.106	2.788
Std. Dev.	0.470	0.316

On average, the complexity of random systems is significantly higher than the one of real systems, both for vocalic ( $t(450) = -10.41$ ;  $p < 0.001$ ) and consonantal systems ( $t(450) = -13.85$ ;  $p < 0.001$ ). These results support the idea that this measure of complexity does capture part of the organization of PIs; random systems are more complex, i.e. they are less structured than real ones. On the other hand, there is a large overlap in the ranges of variation of complexity, especially for vowel systems where both random and real systems have a minimal zero value. A possible interpretation is that the off-diagonal complexity is not discriminative enough, due to a limited number of observed structures for the vowel systems.

In order to further evaluate the performance of the algorithm, we considered the possible variations in terms of complexity among the main linguistic groups to which the UPSID languages belong. Following Maddieson (2006), we separated our sample in the 6 major geographical areas presented in the next table, along with the total number of languages per area and the average vocalic and consonantal complexity value.

Two one-factor ANOVAS, independently run on vocalic and consonantal systems, reveal significant differences among the groups ( $F(5) = 6.02$ ;  $p < 0.001$  and  $F(5) = 23.25$ ;  $p < 0.001$ , respectively). Post-hoc Scheffé's tests reveal furthermore that the structural complexity of the vocalic systems of the area "Australia & New Guinea" is significantly different than the areas "Europe, South and West Asia" and "East and South-East Asia". Regarding consonantal systems, several areas, when considered in pairs, show significant differences. For example, the "Africa" area presents a complexity significantly greater than any other, except for the "Europe, South and West Asia" area which presents a very close average value, as shown in the next table:

	Europe, South and West Asia	East and South- East Asia	Africa	North America	South and Central America	Australia and New Guinea
Number of languages	71	108	74	68	66	64
Structural complexity of vocalic system	0.90	0.87	0.79	0.73	0.81	0.65
Structural complexity of conso- nantal system	1.83	1.61	1.84	1.67	1.51	1.45

#### 2.4. Conclusion

As the previous paragraphs have shown, the offdiagonal complexity seems a promising measure for analyzing the structure of PIs. However, although it coincides pretty well with linguists' intuitions when applied to some specific systems, when the whole set of PIs from UPSID is considered, the distribution of complexity values is very compact, thus limiting the comparison between systems. This is due to the fact that the Claussen measure is more adapted to bigger graphs with more diverse internal structures. In our data, the limited typological structural variation is therefore a problem. One possible improvement could be to take into account the weight of the connections of the graphs (i.e. the phonetic distance between phonemes), but this is not possible yet with this measure. Still, these results suggest the following: (i) there are differences in terms of PI structure among linguistic areas, (ii) there is no relation whatsoever between the complexities of vocalic and consonantal system, and (iii) real PIs display a certain degree of regularity that random systems don't.

### 3. From molecules to phonemes: calculating cohesion and stability for phonological inventories

In this section, we will present an alternative approach aiming at characterizing systemic features of PIs as well. We will propose a measure evaluating the *cohesion* of PIs, by borrowing the concept of energy so familiar in statistical physics. Unlike the measure introduced in the previous section, which was based on the topology of the systems, this one focuses more on the phonemes and their very interactions within systems. We will first present a measure of the interactions between phonemes considered two by two, and then an extension of this measure to the evaluation of the overall cohesion of PIs. Last, an evolutionary model of PIs will be presented and commented on. For reasons of clarity, we will only describe here the case of vocalic systems, but our approach also applies to full systems without separating vowels from consonants, as done in section 2.

#### 3.1. On the notion of attraction and repulsion between phonemes

For the topological approach combined with the off-diagonal complexity, the degree of relationship between phonemes was evaluated using a simple feature-based distance. Here, we propose to measure the interaction between two given segments using their patterns of cooccurrence among languages present in the UPSID database. This approach is based on the assumption that if two segments recurrently appear or don't appear together in PIs, an underlying constraint is probably responsible for this pattern. The study of this kind of regularities in the PIs of UPSID has been partially addressed by Clements (2003), who found convincing arguments in favour of the feature economy theory. To do so, Clements studied contingency matrices (see example below) in order to see whether phonetically close segments have a tendency to attract (to be present in the same PI) or repulse each other (to not appear together). He used a  $\chi^2$  test to ensure that only significant interactions are considered, according to the intrinsic frequency of each phoneme.

Clements' approach can be continued in two directions. First, the  $\chi^2$  test is limited when rare events are at play – a problem Clements did not have to deal with in his study. A solution is to apply the exact Fisher test instead, which can be used with any number of occurrences; actually, the  $\chi^2$  is just an approximation of the Fisher test, less costly in terms of calculation, but for which stronger hypotheses must be met.

A second improvement consists in not only considering the cooccurrence of two phonemes A & B, but more generally the four possibilities **A & B, !A & B, A & !B, !A & !B** where "!" stands for the absence of a phoneme. This allows for capturing a larger set of possibly relevant phenomena than if only considering the case where the two segments are present at the same time. The table below gives the contingency matrix for the phonemes /a/ and /ã/ in UPSID:

	/ã/	!/ã/
/a/	82	310
!/a/	1	58

As we can see, only one language has /ã/ without /a/ whereas 310 others present the reverse situation, /a/ without /ã/. If we only calculate the statistical significance of the cooccurrence between /a/ and /ã/, we are bound to find a rather weak interaction because /a/ has an high intrinsic frequency (/a/ is present in 392 languages out of 451). Nevertheless, the contingency matrix is highly asymmetrical. Taking into account all the four possibilities allows us to measure not only the direct interaction between two phonemes of a system, but also the impact of the presence of one of the two when the other is absent: is the system indifferent or is it going to evolve to "recruit" the missing one or "get rid" of the other?

Other pairs of oral vowels and their nasal counterparts follow the same pattern. This particular distribution may be linked to the mechanism of transphonologisation by which nasal vowels are derived from their oral counterparts by extension of the nasal feature of an adjacent consonant. The nasal vowel cannot appear without the oral one and the rare cases where it does only happen because the oral vowel disappears afterward (usually by quality change<sup>4</sup>). This example clearly illustrates that PIs can reflect diachronic processes although they are only implicitly expressed. The approach we are following is similar to a binarisation of PIs as they are now not only described by the set of phonemes they contain but by the set of all the others they don't contain as well. For example, a system with 5 vowels (out of the 180 ones possible in UPSID) will be described by the presence of these 5 vowels AND by the absence of the 175 others.

To quantify the interaction between two phonemes, we took the logarithm of the exact Fisher test. Since the logarithm of probabilities only provides negative values, it is also necessary to evaluate the direction of the interaction: when two phonemes appear together more often than their re-



spective frequencies would predict if they were independent, a "+" sign is given to their interaction whereas a "-" sign is attributed when the frequency of appearance is less than what is expected under the independence hypothesis. Finally, values have been normalized between -1 and +1, and give the strength of the interaction *I*. Figure 6 below presents some of the strongest interactions found in UPSID PIs.

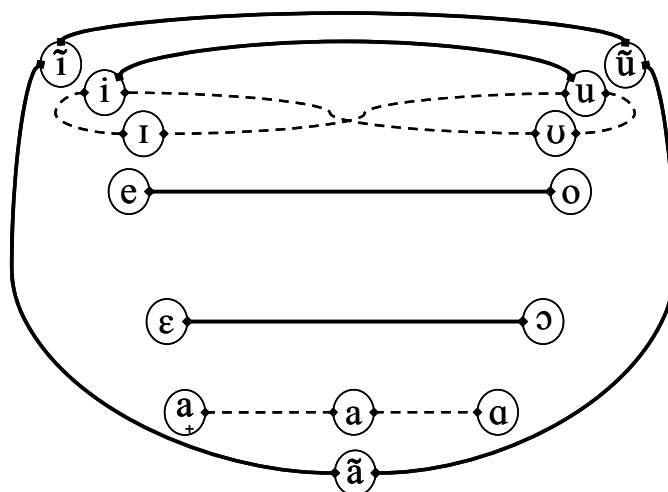


Figure 6. Dashed lines represent the strongest repelling interactions, and solid ones, the strongest attracting interactions.

The relations illustrated in figure 6 reveal that systems have a tendency to harmonize front and back vowels for a given height: /i/ and /u/ attract each other, like /e/ and /o/ or /ɛ/ and /ɔ/. Another positive interaction groups together the three nasal vowels derived from /i, a, u/. Negative interactions are relevant as well. We notice that the three low vowels repel each other; and so do the pairs /i - ɪ/ and /u - ʊ/. The interactions /i - ʊ/ and /ɪ - u/ can be the reflection of the harmony between /i/ and /u/.

In a maybe counterintuitive way, the strongest interactions do not involve /a/ with /i/ or /u/, even though these three segments are present together in a vast majority of languages. This can be explained by the fact that these segments are all very frequent (considered independently), so frequent that the Fisher test does not recognize their interactions as plausible. The same comment applies when the pair of segments involves an extremely rare secondary feature (like breathy-voiced or creaky-voiced for example); the test is then not powerful enough to assign strong interactions. This limitation prevents saying anything about relations between the most

frequent or the rarest segments. It can seem odd, especially for the very frequent segments (/a/, /i/ and /u/ for instance) for which several theories have proposed explanations of their frequency explicitly based on their interaction (in the line of the maximum or adaptive dispersion theory (Liljencrants & Lindblom, 1972)). However, it guarantees that only the information present in the database (and statistically assessed) is considered, without any theoretical a priori. Thus, this approach proposes a theory-neutral point of view that is worth being further explored as a way to access additional information on PIs.

In order to take into account both the interaction and the intrinsic information relative to phoneme frequency in PIs, we also calculate the exact Fisher test for the frequency of distribution of a particular segment compared to a theoretical frequency of 50%. Segments that are present in less than 50% of the languages are given a negative intrinsic value and those present in more than 50% a positive one. These values are obtained by a transformation of the result of the Fisher test similar to the normalization used for the interactions. This intrinsic value  $V$  is linked to the frequency of phonemes through a nonlinear relation that takes the sampling effect into account.

In the current approach, we only consider pairs of segments, but it is theoretically possible to deal with interactions of  $n$ -tuples with  $n > 2$ . Nevertheless, the size of the UPSID database would dramatically limit the number of triplets of phonemes for which significant interactions would be assessed.

### 3.2. From pairs of segments to the whole system

We have defined, on the one hand, the intrinsic value  $V$  for individual segments and on the other hand, the interaction forces  $I$  for pairs of phonemes. Since the exact Fisher test, when applied to the interactions, neutralizes the weight of the intrinsic frequency of the segments, these two measures are statistically independent and thus can be combined for a global measure of cohesion. We now define this measure as  $\mathcal{E}$ :

$$(1) \quad \mathcal{E}(Sv) = \sum_{P_i} V(P_i) + \sum_{\substack{P_i, P_j \\ i \neq j}} I'(P_i, P_j)$$

In this equation,  $Sv$  is a vocalic system,  $P_i$  and  $P_j$  are vowels and  $I'(P_i, P_j) = I(P_i, P_j)$  if  $P_i$  and  $P_j$  are present,  $I(!P_i, P_j)$  if  $P_j$  is present

and  $P_i$  absent etc. This way, we integrate for each pair of segments of a system the relevant combination among the four possible ones (not only present-present). Besides the fact that this doesn't discard useful information, it makes the global cohesion independent of the size of the PI. One potential drawback, though, is the smoothing of the values of  $\mathcal{C}$  and thus the resulting small range of variation between the PIs. However, the study of the PIs' distribution is relevant.

Our approach echoes Pablo Jensen's in a recent economic study about the interactions between retail stores (Jensen, 2006). In his work, the interactions between the various stores, positive or negative, are calculated on the basis of the frequency of their cooccurrences in a close neighbourhood (that plays a similar role to PIs in our approach). All the interactions are then summed to calculate an energy value – corresponding to our value  $I$  – characterizing the organization of an economic and geographic space. This measure can also be calculated for any new potential store in order to evaluate its fitness in the anticipated location.

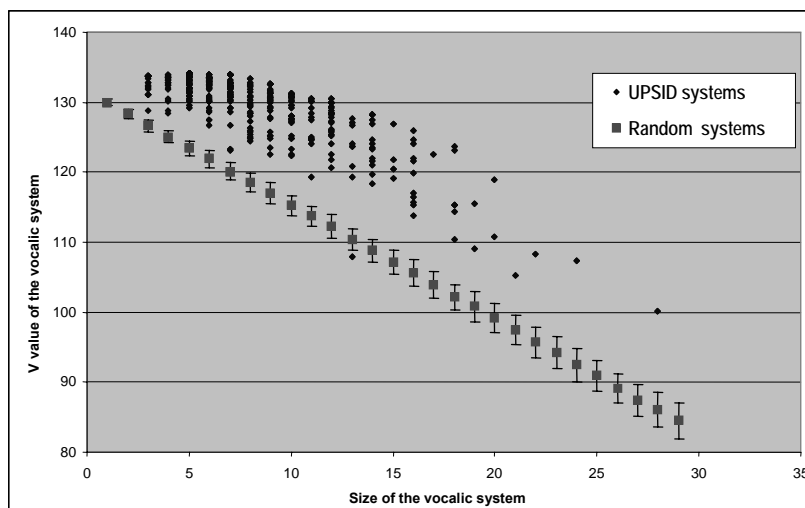
Our approach echoes previous research on maximization of perceptual distance by replacing phoneme-to-phoneme perceptual similarity with synchronic phonemic interactions measured from UPSID. On the one hand, it definitely limits the explanatory power since the interactions revealed probably result from several factors without really identifying them. On the other, it enables us to examine the phonological system as a whole, and not only the primary vowels for instance, since all possible pairs of phonemes can be considered. Moreover, it provides a way to reveal interactions that would have been ignored in other more traditional approaches. Still, an important drawback remains since a kind of circularity is present, because we *a priori* use the frequency of distribution of segments to produce results on the same inventories.

The concept of cohesion, defined as above, may intuitively be connected with a kind of global fitness of PIs: a system consisting of a set of antagonistic phonemes that have a tendency to repel each other would be ill-fitted; vice-versa a well-fitted system would consist of phonemes that go well with each other. Yet, this approach strongly relies on the implicit postulate that summing over the inventory the 1 by 1 interaction within each pair of phonemes is able to capture the complexity of the whole system. We thus hypothesise that we are dealing with a nonlinear second order complexity, and not a higher order one. If, based on this hypothesis, we obtain good results, for example in the predictions of the evolution of PI, then it would seem reasonable to say that PIs are of a relatively “small” complexity compared

to other systems with complexities of higher order. More explicitly, it would indicate that the model based on second-order interactions is a good approximation. On the contrary, if no valid result is reached, higher-order complexity (involving patterns of interactions with 3 or more segments) might be assumed.

### 3.3. Cohesion of the UPSID phonological inventories

The presentation of the results starts with a comparison of the vowel systems of the 451 languages from UPSID with random systems, and by distinguishing the contribution of the intrinsic value  $V$  from the impact of the interactions  $I$  (Figure 7 to Figure 9).



*Figure 7.* Intrinsic values  $V$  for UPSID vowel systems (in dark) and random systems (in grey). Standard Deviation bars are displayed for the distribution of random systems.

Figure 7 shows that the intrinsic values  $V$  are higher for real systems than for random ones. Besides,  $V$  tends to decrease when the size of the system increases, corresponding to the appearance of rarer segments in the system. If we take a look at the maximal values of  $V$  reached for given sizes of the system, one can observe the following hierarchy, given that  $S(n)$  is the system of size  $n$  of maximum intrinsic value:

$$V\{S(5)\} > V\{S(6)\} > V\{S(7)\} > V\{S(4)\} > V\{S(3)\}.$$

Figure 8 deals with the interaction forces.  $I$  is higher, on average, for real systems than for random systems, although the distributions are overlapping. The overlap decreases for larger inventories since  $I$  increases for a significant proportion of real systems while, on average, it monotonically decreases for random systems.

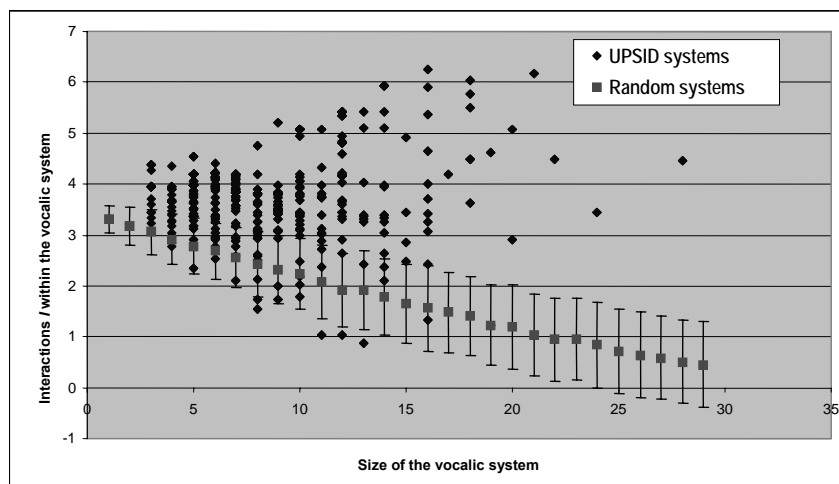


Figure 8. Interaction values  $I$  for UPSID vs. random vowel systems. Color code is the same as for Figure 7.

A plausible explanation is that the more the size of the system increases, the more likely it is to contain phonemes with a low intrinsic value  $V$ ; However, the recruited phonemes have a tendency to positively interact with each other (high  $I$ ).

Figure 9 represents the global measure of cohesion  $e$ . At first sight, the results are similar to those of intrinsic values  $V$ . The main reason is that the range of variation of  $I$  is much lower than  $V$  variation (this is visible by comparing scales of axes of ordinates from Figures 7 and 8). There are however differences to be highlighted with respect to Figure 7:

The ranking of the systems with the strongest cohesions for given sizes leads to the following order:  $e_{S(5)} > e_{S(7)} > e_{S(3)} > e_{S(6)}$ . Like for the intrinsic value, the max is obtained for a 5-vowel system (/i, e, a, o, u/), but the following hierarchy is different, suggesting that interactions play a role in the global cohesion of a system.

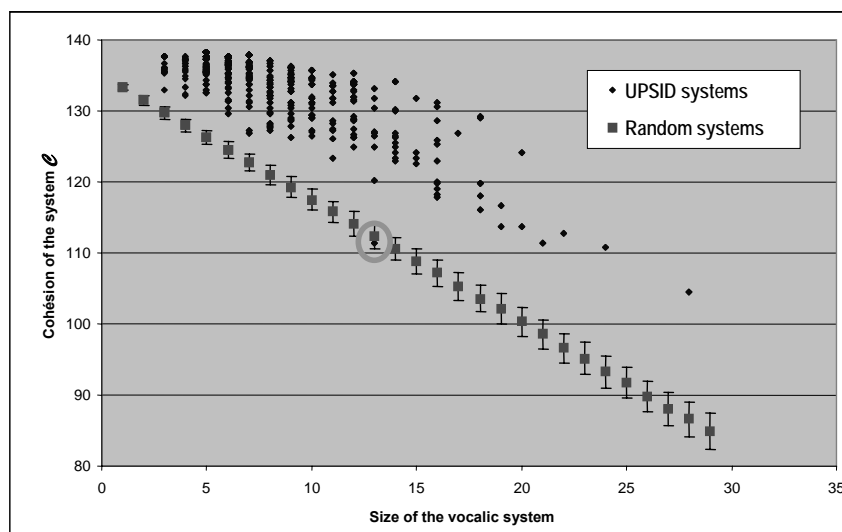


Figure 9. Global cohesion value for UPSID vs. random vocalic systems. Dinka is circled.

It is worth noticing that Dinka (Nilotic family, circled on the graph) with its 13 vowels doesn't follow the general trend of UPSID PIs but falls into the variability of random systems. This system is indeed extremely uncommon since it contains 7 breathy-voiced and 6 creaky-voiced vowels with no modal voiced vowels (although we should probably take such a description of Dinka with caution). This fact decreases the intrinsic value of the system even though its interactional strength is not especially low (3.40).

As a last remark, we would mention that none of the most cohesive systems violates principles such as symmetry, gradual filling of the vocalic space, etc. without these principles being directly specified in the calculations we've presented. Among the most cohesive systems, the first to use a secondary feature is a ten vowel system for which the five vowels /i, e, a, o, u/ are contrasted in terms of nasality.

### 3.4. From static measures of cohesion to evolutionary dynamics: the notion of stability

Using the measure of cohesion of the PIs as a fitness measure, we can now build a relatively simple model of stochastic evolution where various possible evolutionary trajectories are implemented and evaluated. The main driving force (and hypothesis) of this model is that a change is more likely to happen if it increases the global cohesion of the system where it takes place. This does not imply that changes decreasing the cohesion of a PI are impossible, for example under the influence of social constraints, but such changes are less likely to happen, and consequently, they are rarer in the simulations.

The evolutionary algorithm processes as follow:

- \* For a given PI  $S$ , 100 new systems are built differing by 0, 1 or more segments. These systems represent possible evolutions from  $S$  to a neighbour system.
- \* The probability of each potential evolution is calculated by comparing its global cohesion with that of  $S$ , and then normalizing the differences in order to have a set of probabilities ranging from 0 to 1. The changes leading to an increase of cohesion will have the highest probabilities.
- \* A system among the 100 is randomly chosen with respect to this distribution of probabilities. This system is considered as the new state of system  $S$ .

Several mechanisms have been tested to explore the energetic landscape of the given PI  $S$ , as well as for the normalization of the differences between initial and final cohesion. They all lead to comparable results.

Our model can test several evolutionary routes and then estimate the stability of a system as a function of its cohesion compared to that of its neighbouring systems.

The stability is evaluated as follows: given a particular system, let us consider 500 independent evolutionary hypotheses (as the one described above) and evaluate the percentage of evolutions that maintain the system in its initial state (no change). The more cohesive the system  $S$  compared to its neighbours, the more likely the continuation of this state and thus the higher the stability. Vice-versa, a system surrounded by more cohesive systems is instable and very likely to change.

Figure 10 presents several indicators derived from the stability simulation for UPSID vowel systems. For a given size of the systems (X axis), the graph displays the stability of the less stable system (diamond shape), the most stable one (triangle shape) and the average stability of all the systems of that given size (square shape).

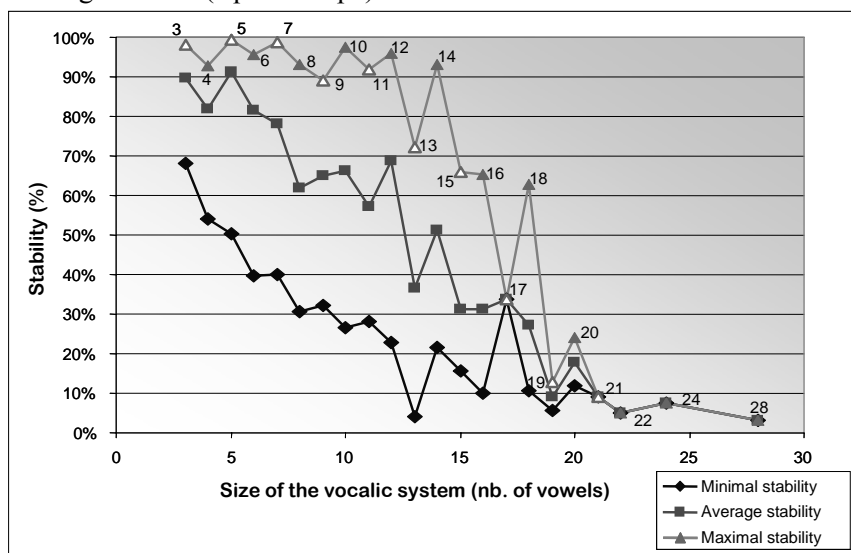


Figure 10. Stability of UPSID vowel systems sorted by increasing size. Numbers along the top curve give the size of the corresponding system.

Interestingly, whereas the maximal global cohesion decreased as soon as systems reached 7 or 8 segments, we notice that maximal stability is still high even for 12 or 14 vowels. Thus, the simulated evolutions reveal that large systems can play the role of stable attractors even if their cohesion is not really high; what matters is that it is higher than their neighbours.

Another interesting result is the change of mode that operates at sizes close to 9 vowels: for smaller systems, odd numbers of vowels are the most stable (empty triangles in the graph) whereas for larger systems, stability comes with an even number of segments (full triangles). This particularly salient effect is worth being linked to the change of organization observable in the contents of phonological systems precisely around 9 vowels (Vallée, 1994). Below this threshold, we mostly find primary vowels in systems when above, systems tend to reorganize in series, contrasting a primary set of vowels by/with a secondary feature. To this regard, Kolokuma Ijo (spoken in Nigeria) is a good example as it has 18 vowels: 9 different qualities



and their 9 nasal counterparts; it turns out to be the most stable 18-vowel system in UPSID (around 63%).

#### 4. Conclusions and perspectives

Phonological systems, because of their variety and their structure, constitute an archetype of complex organized systems. They are the reflection of physiological, cognitive and linguistic constraints together, as well as socio-linguistic ones linked to the interactions between speakers. Our understanding of these constraints, of their interactions and their impact on the evolution of systems themselves is still limited due to their complexity. In this picture, the science of complexity provides particularly powerful tools to shed new light on the issues at hand, especially to understand better the connections between the microscopic level (the phonological constituents) and the macroscopic level (each system, considered as a whole). However, their development and their adaptation to linguistics is not straightforward, and even if the first results are promising we must keep this difficulty in mind.

The different approaches developed in this paper aim at extracting the intrinsic information hidden in a typological database of phonological inventories, avoiding as much as possible traditional *a priori*s in linguistic theories. More precisely, we paid attention to two factors of the complexity of PIs: the structural complexity and the interactional complexity. The issue of the hierarchy of phoneme complexity has been partially addressed in a previous paper (Marsico et al., 2004), by correlating the frequency of occurrence of phonemes in PIs to their capacity to generate new phonemes derived by addition of secondary features.

The methodology used to evaluate the structural complexity comes from graph theory; it tries to take into account some regular patterns of organisation potentially important in phonological systems (like principles of economy or symmetry for example) and to dissociate the effects of the topology of the system from those of its size. These metrics shed light on the fact that the systems of the languages of the world are globally more structured than randomly constituted ones, and that significant topological differences exist between different linguistic areas. Furthermore, the complexity of consonantal systems is higher than that of vocalic systems.

This last result raises a recurrent question relative to phonological systems: can we apply the same analysis to consonantal and vocalic spaces?

For a while, the structural differences seemed irreducible (discontinuity-continuity, different acoustic cues and articulatory gestures). However, we think, following Lindblom and Maddieson (1988), that a common theory is possible, at least concerning the main internal principles structuring these spaces with a balance between a perceptual principle of sufficient contrast and an articulatory one of least effort or economy.

Regarding the interactional complexity of PIs – and their intrinsic complexity as well - we used a methodology directly inspired by the interactional forces within physical systems and by the calculation of the resulting energy of this system. We have thus calculated indices of intrinsic value (linked to the identity of the different phonemes of a PI), and of interactional force (linked to the reciprocal influences of phonemes among them), for all the vocalic systems of our database. Here again, this approach has shown that a significant difference exists between real and random systems. Furthermore, these measures confirm that when the size of the vocalic system increases, the existing phonemes have a strong reciprocal positive influence (i.e. the interactional force increases), in order to partially compensate for the fact that these phonemes may have a smaller intrinsic value. This compensation, similar to a positive retroaction loop typical of numerous complex systems, can furthermore be a determining factor in the mechanisms of evolution of phonological systems.

To test this hypothesis, we modelled a stochastic evolution of PI based on real systems, taking into account the global cohesion reached by the systems at each step of evolution. This led us to estimate a stability value for each system, and it showed that even systems with a relatively small cohesion (most often because of a large number of vowels) are judged stable by the model. Moreover, stable systems with more than 9 vowels use two distinct series of vowels (oral vs. nasal for example), thus illustrating a principle of feature economy or parsimony. We see here an emergent regularity which, if not predictable on the basis of cohesion alone, is nevertheless compatible with the fact that systems with a relatively large number of vowels are not rare (45% of UPSID languages have 8 or more vowels).

More than the results themselves, our intention is to validate the fact that an interdisciplinary approach coming from the science of complexity allows the effective extraction of relevant information from PIs. This should not hide the fact that various issues are still at stake and require further study. One of the most important points is to define an approach which better validates the predictions of the evolutionary model, despite the limited size of UPSID. As a matter of fact, the frequency of distribution of

systems (which can be linked to some extent to the quality – or fitness – of their response to synchronic and diachronic constraints) cannot be properly estimated with this database. If this major problem is solved, we will be able to evaluate more precisely if the second order complexity (consideration of 2-2 interactions at the microscopic level) gives a good approximation of the fitness of systems or if the relations existing between micro- and macroscopic levels are even more complex.

Finally, another interesting aspect deals with the study of the evolutionary routes themselves, so as to discover potential attractors and cyclic attested trajectories in the history of languages. In the long term, the instantiation of these elements within a multi-agents model will allow us to address the external factors of evolution (socio-linguistic ones) as well, and to confront this approach with rich theoretical frameworks, such as the one proposed in Mufwene (2001).

### Notes

1. All our data come from a slightly modified version of the UPSID database (Maddieson, 1984, Maddieson & Precoda, 1990) which contains 451 languages balanced regarding geographical distribution and genetic affiliation.
2. This approach has been selected from among several potential methods because it preserved interesting properties in terms of structure (see below).
3. The sets of features describing vowels and consonants being disjoint, we applied the algorithm separately on the two sub-systems.
4. The only language in UPSID presenting that situation is Kashmiri, with an opposition between /ɒ/ and /ã/.

### References

- Abraham R.  
2001 *The genesis of complexity*, unpublished ms available at: <http://www.ralph-abraham.org/articles/MS%23108.Complex/complex.pdf> (consulted in December 2007).
- Bonchev, D. and Buck, G. A.  
2005 Quantitative measures of network complexity. In *Complexity in chemistry biology and ecology*. Bonchev, D. and Rouvray, D. (Eds.). Springer Verlag. New York.

- Cancho, R. F. i. and Solé, R. V.  
 2001 The small-world of human language. *Santa Fe Institute Working Paper* 01: 03-016.
- Cancho, R. F. i., Solé, R. V. and Köhler, R.  
 2004 Patterns in syntactic dependency networks. *Physical Review E*. 69: 051915-051911-051919.
- Claussen, J. C.,  
 2004 Off-diagonal Complexity: A computationally quick complexity measure for graphs and networks. *q-bio.MN/0410024*.
- Clements, G. N.  
 2003 Feature economy as a phonological universal. *Proceedings of the 15<sup>th</sup> International Congress of Phonetic Sciences*. Barcelona. pp. 371-374.
- Dorogovtsev, S. N. and Mendes, J. F. F.  
 2001 Language as an evolving word web. *Proceedings of The Royal Society of London Series B, Biological Sciences*, 268: 2603-2606.
- Erdős, P. and Rényi, A.  
 1960 The Evolution of Random Graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* 5: 17-61.
- Ferguson, Charles H.  
 1978 Historical backgrounds of universal research. In J. H. Greenberg, C. A. Ferguson & E. A. Moravcsik (eds.). *Universals of human language*, vol. 1. pp. 61-93. Stanford, CA: Stanford University Press.
- Jensen, P.  
 2006 Network-based predictions of retail store commercial categories and optimal locations. *Phys. Rev. E* 74: 035101.
- Jukna, S.  
 2006 On graph complexity. *Combinatorics, Probability & Computing* 15: 1-22.
- Liljencrants, J., Lindblom, B.  
 1972 Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* 48: 839-862.
- Lindblom, B.  
 1986 Phonetic universals in vowel systems. In *Experimental phonology*. Ohala, J., Jaeger, J. (eds). Orlando: Academic Press. pp. 13-44.  
 1998 Systemic constraints and adaptive change in the formation of sound structure. In *Approaches to the evolution of language*. Hurford J.R., Studdert-Kennedy M. & Knight C. (eds). Cambridge University Press: Cambridge. pp. 242-264.
- Lindblom, B. and Maddieson, I.  
 1988 Phonetic universals in consonant systems. In *Language, Speech and mind*. Li, C., Hyman, L. (eds). London: Routledge. pp. 62-78.
- Maddieson, I.  
 1984 *Patterns of sounds*. Cambridge: Cambridge University Press.

- 2006 Correlating phonological complexity: data and validation. *Linguistic Typology* 10.1: 106-123
- Maddieson, I., Precoda, K.  
1990 Updating UPSID. *UCLA Working Papers in Phonetics* 74: 104-111
- Marsico, E., Maddieson, I., Coupé, C., Pellegrino, F.  
2004 Investigating the hidden structure of phonological systems. *Proceedings of the 30th Meeting of the Berkeley Linguistic Society*. Berkeley. pp. 256-267.
- Mufwene, S. S.  
2001 *The ecology of language evolution*. Cambridge: Cambridge University Press.
- Neel, D. L. and Orrison, M. E.  
2006 The Linear Complexity of a Graph. *The Electronic Journal of Combinatorics* 13.
- Ohala, J. J.  
1980 Moderator's summary of symposium on 'Phonetic universals in phonological systems and their explanation', *Proceedings of the 9th International Congress of Phonetic Sciences*, Vol. 3. Copenhagen: Institute of Phonetics, pp. 181-194.
- Pastor-Satorras, R. and Vespignani, A.  
2001 Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86: 3200-3203.
- Schwartz, J. L., Boč, L. J., Vallée, N., and Abry, C.  
1997 The Dispersion-Focalization Theory of Vowel Systems. *Journal of Phonetics* 25.3: 255-286
- Solé, R. V.  
2004 Scaling laws in language evolution. In *Power Laws in the Social Sciences*. Cioffi, C. (Ed.), Cambridge University Press, Cambridge, MA.
- Steels, L.  
1997 The synthetic modeling of language origin. *Evolution of Communication Journal* 1: 1-34
- Vallée, N.  
1994 *Systèmes vocaliques : de la typologie aux prédictions*. Thèse de doctorat, Grenoble: Université Stendhal.
- Watts, D.J. and Strogatz, S. H.  
1998 Small world. *Nature*. 393: 440-442
- Weaver, W.  
1948 Science and Complexity. *American Scientist* 36: 536