# A PRELIMINARY CASE FOR EXPLORATORY NETWORKS IN BIOLOGY AND LINGUISTICS: THE PHONETIC NETWORK OF CHINESE WORDS AS A CASE-STUDY

*Philippe Lopez, Johann-Mattis List, Eric Bapteste*

## THE RAISE OF NETWORKS AS COMPARATIVE METHODS IN EVOLUTIONARY BIOLOGY

Linguists, as well as biologists, study historical objects that form lineages, undergoing transformations over time. Biologists, as well as linguists, therefore, are very dependent on comparative analyses to structure and analyze their data. Thus, it seems intuitive that conceptual and methodological researches in both fields could inform each other, and benefit to both fields. In particular, the comparative approaches elaborated in biology are experiencing massive developments that could be explored in linguistic studies.

In biology, the general picture of evolution is becoming increasingly complex. Evolutionary innovations and changes are effected both by processes of vertical descent and introgressive (or combinatory) processes (recombination, lateral gene transfer, symbioses) (Bapteste et al. 2012; Bapteste et al. 2009; Dagan et al. 2008; Dagan and Martin 2009; Huang and Gogarten 2007; Kloesges et al. 2011; Marin et al. 2005; Wu et al. 2011). Vertical descent processes are usually modeled and studied using a common tree (e.g. a gene or a species tree) (O'Malley 2011). By contrast, combinatory processes reassort, regroup or merge evolutionary objects. Examples include mosaic genes, genomes and intricate symbiotic associations, and coalitions based on multiple lineages, persisting via the tight co-evolution of evolutionary players from distinct lineages (Bapteste et al. 2012) (e.g. cells and mobile genetic elements such as plasmids and phages in multispecies biofilms (Ghigo 2001; Hall-Stoodley et al. 2004; Periasamy 2009; Wintermute 2010) or in the gut microbiomes (Jones 2010; Lozupone et al. 2008; Qu et al. 2008; Martin et al. 2007)). Thus original genetic associations from multiple sources, sustained by a diversity of evolutionary processes, can be cemented into novel evolutionary units, i.e. when the transfer of domains produces new genes, or the transfer of genes produces new gene clusters, pathways and mosaic genomes. Likewise genetic associations between distantly related entities can evolve into novel symbiotic organisms and microbial coalitions (Dagan et al. 2008; Martin et al. 2007; Moustafa et al. 2009).

Consequently, the usual framework of a single tree fails to represent the evolution of many biological entities, at different biological scales, in particular when these entities are mergers from multiple lineages. Problems also arise when the in-

vestigated entities are too divergent from reference entities already put in a reference tree to be simultaneously analyzed on the same tree. Highly divergent entities will typically not be readily comparable with reference entities in a single analysis, as the homology between divergent entities becomes too distant to be effectively detected. This problem does not constitute a major limit for network analyses, which can handle higher levels of divergence (Alvarez-Ponce et al. 2013; Beauregard-Racin et al. 2011). Thus, in biology networks are increasingly used as alternative models to describe more of the complexity of biological evolution (Bapteste et al. 2012; Dagan et al. 2008; Dagan 2009; Alvarez-Ponce et al. 2013; Beauregard-Racine et al. 2011; Fondi and Fani 2010; Halary et al. 2010; Lima-Mendez et al. 2008); Skippington and Ragan 2011. These methods are an invaluable complement to the construction of common trees of single lineages of objects from a given level of organisation (e.g. gene trees focusing on genes, organismal trees focusing on organisms, species trees focusing on species, etc.). Moreover, their potential to provide a novel analytical framework for exploratory evolutionary studies is also increasingly acknowledged.

Indeed, networks are more flexible graphs than trees. They are less constrained in their representation of the data and the relationships between objects, and can support different levels of abstraction. Typically, a network G, noted G = (V, E), comprises a set V of vertices or nodes associated with a set E of edges. The nature of the nodes (e.g. a domain, a gene, a gene cluster, a genome, an environment) as well as their rules of connection can be used as parameters that vary in exploratory analyses (Burian 2011). Thus, using networks of genes, or of genomes, or of lineages, or of environments, biological diversity can be observed at many levels, e.g. within one (or many) gene families, genomes, lineages, communities, or environments (Zhaxybayeva and Doolittle 2011), by simply varying the nature of the investigated nodes. Moreover, each level of biological diversity can be structured in different informative ways by changing the types of edges represented in these graphs. For instance, for nodes corresponding to the same set of protein sequences, a graph could either only show connections retracing functional interactions between these proteins (Martha et al. 2011; Vinayagam et al. 2011; Wang et al. 2011), or connections reflecting only genealogical relationships between these proteins (Beauregard-Racine et al. 2011; Alvarez-Ponce and McInernex 2011), etc. When these variations in the type of edges represented in these networks induce changes in the graph topology between the nodes, networks comparisons can identify robust/transient patterns of connections, appearing over a large/limited range of conditions / biological levels, i.e. transient functional interactions between unrelated proteins.

Interestingly, these remarkable patterns need not necessarily be a priori expected. Network-based studies of genetic diversity typically foster the discovery of many unrecognized patterns, and thus contribute to actively generate novel hypotheses about the evolution of genetic diversity. For example, in a gene network (Beauregard-Racine et al. 2011; Bittner et al. 2010), nodes are gene sequences, connected by weighted edges when they share a relationship of homology/identity, as assessed by a BLAST score. Each gene family is easily characterized as it falls in a separated

connected component. When environmental sequences are included in such analyses along with sequences from cultured organisms, novel environmental gene families can be discovered. Moreover, such gene networks can be used to detect evolutionary units that a tree of sequences alone cannot detect. For instance, the study of $10^4$–$10^6$ sequences allows to detect groups of genes families with complex evolutionary patterns (expansions, high evolutionary rates, combinations (Beauregard-Racine et al. 2011), etc.), fused genes (Gallagher and Eliassi-Rad 2008; Jachiet et al. 2013), and very distantly related gene forms (Alvarez-Ponce et al. 2013), branching off rather than within the sequences from known gene families. Hence, an exploratory approach of genetic diversity can unravel unsuspected highly divergent gene forms, and questions what their biological function might be.

The exploratory use of network can be formalized in even more general terms. While most evolutionary studies are mainly concerned by the justification of theories about how entities diverge or by the test of genealogical hypotheses seeking to establish sister-group relationships, exploratory sciences try to develop new concepts to 'fix any evolutionary phenomenon' calling for explanation (Burian 2011; Franklin-Hall 2005). It uses networks to establish and classify relevant patterns that had not yet been well characterized, such as the patterns that a tree-based approach would fail to represent. Networks can not only quickly sort massive amount of data with limited *a priori* on the connections between the objects analyzed in these data, but also rapidly expose their potential underlying (intriguing) patterns/structures.

Finally, networks offer a precious mathematical framework for comparative and exploratory analyses, because the topological properties of their nodes and edges (Koschützki 2008) can be computed and compared. Topological indices, such as the conductance (Leskovec et al. 2008) of a group of nodes can be estimated. For a given group of nodes (e.g. nodes corresponding to words from a given cognate, or to genes from a given gene family), the conductance C is computed as: $C = N_{ext} / (N_{ext} + 2*N_{int})$, where $N_{int}$ is the number of internal edges (e.g. linking members of that cognate or gene family) and $N_{ext}$ is the number of external edges (e.g. linking a member from that cognate/gene family and a member from another cognate/gene family). Clustered and/or isolated groups (e.g. of words from the same cognate or of genes from the same gene family) have a conductance close to 0, while spread out or fragmented groups have a conductance close to 1. Thus, the conductance measures whether nodes with a given label cluster in the networks (i.e. whether words from the same cognate, grammatical class or dialect are more similar to one another than to any other words; or whether genes with the same function, or from the same genus, or from the same environment, are more similar to one another than to any other gene).

Importantly, the increasingly recognized diversity of biological evolutionary processes and patterns observed in biological studies may also find some echo in the field of linguistics. This latter discipline also inquires history and evolution of numerous evolving entities, such as word families and languages, which may very well be effected by vertical and combinatory processes (Nelson-Sathi et al. 2011) (see Table 1, for a possible analogy between the evolution of biological and linguistics objects). Therefore, we wanted to test here whether the study of some linguistic

objects using networks could foster novel hypotheses about their evolution, and offer a test-case for the relevance of some of the analogies between biological and linguistic objects. More precisely, we used a dataset of 48 semantic glosses translated into 40 Chinese dialects to reconstruct a word network based on phonetic similarity. We classified these words by their meaning, dialect of origin and grammatical categories, and estimated the conductance (e.g. the phonetic consistency) of each meaning, dialect, and grammatical category. We observed that different selective (sociological and linguistic) pressures are acting on how a word sounds, introducing phonetic variability and structure in Chinese languages according to different rules, influenced by the grammatical category to which the words belong. Yet, cognate sets and gene families present rather different levels of diversity (phonetic and genetic, respectively), encouraging the innovative development of specific network methods in linguistics rather than the simple import of comparative methods of evolutionary biology that are currently better suited for biological objects.

*Table 1. Some possible correspondence for an analogy between evolutionary biology and linguistics*

| **Evolutionary Biology** | **Evolutionary Linguistics** |
|---|---|
| Gene (particular function) | Word (meaning) |
| Gene family | Cognate set |
| Gene functional ontology | Grammatical category |
| Genome | Dialect, Language |
| Lateral Gene Transfer | Lexical borrowing |
| Genetic diversity | Phonetic diversity |
| Distant homology | Hidden cognacy |
| Selective pressures | Sociological, linguistic constraints |

## THEORETICAL POWER OF EXPLORATORY NETWORKS IN LINGUISTICS

Network-based analyses allow relaxing some *a priori* constraints generally imposed by tree-based analyses. Although disquieting in the first place for practitioners more trained to work within a tree-based framework, this reduction of constraints in data display offers a novel way to capture more of the evolutionary processes and patterns in addition to the process and pattern of 'vertical descent with modification'. This general observation, we believe, probably holds true for both evolutionary biology and linguistics, assuming that in both fields several processes cannot be properly represented and modeled with a tree-based approach, which inexorably constrains the analyses to be only expressed in terms of divergence and dichotomies, as well as the type of data suited for an evolutionary analysis. In molecular phylogenetics for instance, the suitable material are homologous sequences

that align well with one another, since they belong to a single sequence family derived from an ancestral gene copy. This practice considerably restricts the amount of molecular data amenable for analysis, and thus the scope of the analysis (Bapteste et al. 2012; Bapteste et al. 2008; Dagan and Martin 2006; Leigh et al. 2011). Sequences undergoing more complex evolutionary processes are not included in the analyses, because they would blur the reconstruction of the gene (or species) genealogy that tree-based analysis generally aims for. Gene networks overcome this issue of massive *a priori* data exclusion, by allowing the display of more processes and relations between gene forms (although these are not only the usual relations of homology) than is permitted by a tree. Similarly, we argue that word networks may extend the scope of linguistic analysis beyond inferences focused on predefined cognate sets by recovering more distant cases of cognacy or by introducing novel measures of similarity (here phonetic distances) between words.

Various types of distances could be used to reconstruct a word network. In the present analysis, we focus on phonetic word networks, as a mean to display (and then later to analyze) the phonetic diversity of words within several dialects (Figure 1).

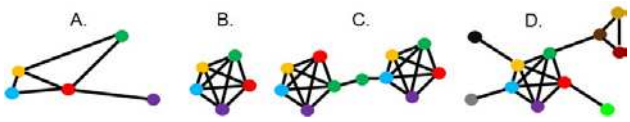


*Figure 1: Virtual weighted cognate networks.*

**A**. A component corresponding to a set of words that may, or may not, belong to an accepted cognate family. Nodes are words, color-coded based on their dialect of use. Edges are weighted according to any distance metric, i.e. a phonetic distance between pairs of words. **B-D.** Virtual component topologies that would support distinct interpretations on words evolution. B. A phonetically conserved word family, the typical pattern in a word network for a bona fide cognate set. C: The bridging node is an emerging word resulting from borrowing and fusion events of words from distinct dialects. D: A family of words with two strongly connected communities and peripheral nodes (light green, grey, and black), indicating distantly related versions of these words in several dialects, suggesting different evolutionary rates (of the sounds) of these words in the dialects, while showing some 'regional' conservation in the brown dialects.

Such word networks are disconnected, because each set of words using a common pool of phonemes will create its own connected component in the graph. Indeed, when two words are phonetically different (e.g. presenting less than a minimal phonetic similarity with one another), no edge is drawn between them, and they fall into distinct subgraphs within the word network. Otherwise, when two words display some phonetic similarity (e.g. when their phonetic distance is lower than a given threshold), these words are connected by weighted edges, with an edge weight that is inversely proportional to the phonetic distance, so that words closer in phonetic distance have edges with higher weights on the graph.

A notable consequence of the great inclusiveness of such phonetic word networks is that not all their components have to be cliques, i.e. maximally connected components in which each and every node directly connects to each and every other

node in the "word family". In biological networks (gene networks), the clique pattern is typical for ubiquitous and conserved gene families, in which all sequences are highly similar because inherited from a last common ancestor and affected by a relatively limited amount of mutations so that their homology can still be successfully assessed. However, it is an empirical question whether, in word networks, the analogs of gene families, the cognates, will also produce cliques, with related words sounding sufficiently alike to be all directly connected together to the exclusion of other unrelated words. More or less structured connections can emerge in these graphs showing phonetic distances between words. In particular, cognates may not produce cliques, and belong to components that are either the result of more complex evolutionary processes than vertical descent from an ancestral word alone (e.g. the evolution of these words and word families may involve combinatory processes); or they could belong to components joining groups of words affected by phonetic convergences (a phenomenon that is expected because a typical word is short and the phoneme diversity is limited). These latter components may mix together words belonging to different cognate sets, however connected because they exploit overlapping pools of "phonemes". Finally, members of the same cognate set may also be highly disconnected in the network, if those cognates are word families in which sounds evolve very fast, to a point that it becomes impossible to detect the common historical origin of these words based on phonetic distances alone. A further investigation of the classes of words (organized by dialects, grammar and meanings) may unravel some rules of "phoneme" associations and the constraints that may affect how words sound. Here, we investigated phonetic diversity from multiple perspectives to test whether and how the dialect of origin of a word, or its grammatical function, or its meaning affected its phonetic consistency.

## APPLICATION TO THE NETWORK OF CHINESE COGNATES

We used a subpart of Hóu's collection of Chinese dialect data (Hóu 2004), consisting of 48 semantic glosses translated into 40 Chinese dialects. The whole data comprised 2,999 different words. Following cognate judgments provided in the original data, these words were grouped into 337 different cognate sets. We further calculated phonetic distances between all words using the SCA method (List 2012) to derive alignment scores and the formula by Downey et al. (Downey et al. 2008) to convert similarity into distance scores. Mean distance between any two words was estimated to be 1.17, but only 0.35 between two cognates (Figure 2).

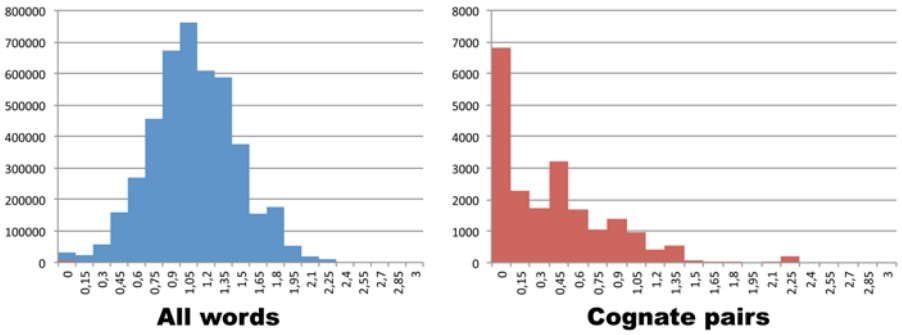


*Figure 2: Phonetic distance between pairs of words estimated by SCA.*

Left panel: Phonetic distance between all pairs of words; Right panel: Phonetic distance between pairs of cognates. This graph shows that words belonging to the same cognate set sound much more alike than random pairs of words, however they can still show important phonetic variations.

This simple observation indicated that words belonging to the same cognate set sound much more alike than random pairs of words, a property that is also observed in evolutionary biology. Extant sequences diverging from the same ancestral sequence (homologous sequences) are expected to be much more similar to each other than to any unrelated sequence and given the extremely low probability that two random sequences will show some similarity by chance alone, molecular evolutionists usually consider that analogy implies homology. We thus investigated whether similarity networks applied to linguistic data would perform in a similar way as they do in molecular evolution. Just as a threshold is needed to determine whether two sequences should be considered homologous, a maximum phonetic distance has to be used to determine if two words show significant phonetic proximity. Since the average distance between words from the same cognate was 0.35, we used that threshold to build our network and linked pair of words that showed a phonetic distance lower or equal to 0.35. This protocol allowed us to filter the 4.5 million potential edges (for 2,999 words) of the most inclusive word network to reduce it to its most pronounced relationships, summarized by about 60,000 edges encompassing 97 % of the word dataset. This phonetic diversity was then refined by distinguishing two kinds of edges in the reduced network: on the one hand, cognate edges, connecting two members of the same cognate set, and on the other hand, similarity edges connecting members from different cognate sets. First, for representation purposes, we used only cognate edges (Figure 3). The resulting sub-network (i) very neatly split some cognates into different graph components (indicating that groups of words from the same cognate set can sound very differently) and (ii) showed components that were not cliques, demonstrating the importance of phonetic variation within cognate sets.

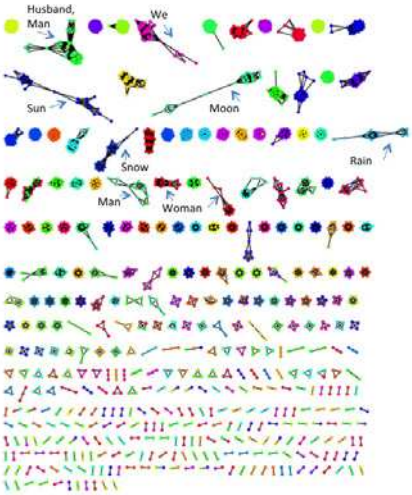


*Figure 3: Network of close phonetic proximity between words belonging to the same cognate set.*

Nodes correspond to words, colored by meanings, connected by edges indicating a close phonetic proximity (distance < 0.35) between pairs of words from the same cognate set. Some meanings are indicated along this subnetwork. Some connected components are not cliques, indicating strong divergence.

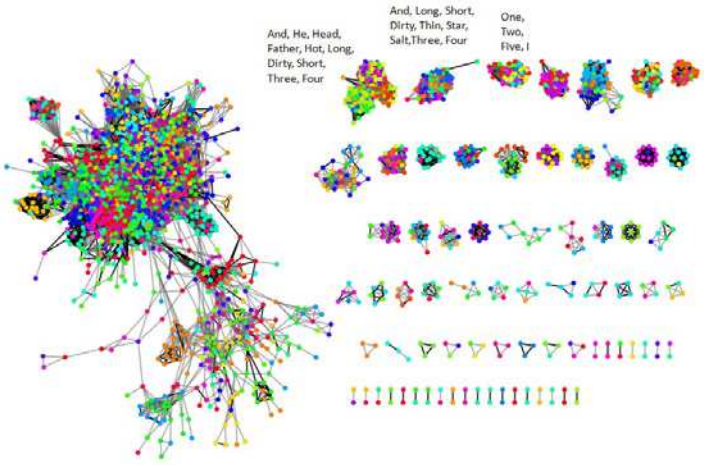


*Figure 4: Network of close phonetic proximity with both cognate and similarity edges.*

Nodes correspond to words, colored by meanings, connected by cognate edges (in black) and similarity edges (in grey) indicating a close phonetic proximity (distance < 0.35) between pairs of words. Colors are extremely scrambled, showing that the phonetic consistency (clustering and isolation) of most Chinese meanings is low.

However, one should not be mistaken by the exclusive focus on cognate edges, because when all phonetic comparisons (hence when both cognate and similarity edges) are considered, it is equally obvious that words from different cognate sets also sound very close (Figure 4). In other words, there are strong phonetic convergences between historically unrelated words. This observation brings forward a fundamental difference between cognate sets and their presumed analogs in the biological sciences: gene families. While gene families can be identified based on their genetic distances, cognate sets cannot be successfully identified based on their phonetic distances. In linguistics, this is known as the problem of *phenotypic* as opposed to *genotypic* similarity (Lass 1997). Phenotypic similarity refers to word similarities based on language-independent surface criteria by which the similarity of phonetic segments is determined. Genotypic similarity refers to similarity that is language-specific.That means that it can be only defined for distinct pairs of languages that are known to be genetically related. For a given language pair, genotypic similarity is determined in form of sound correspondences, that is sounds (phonemes) that are known to be homologous. As an example for such correspondences, compare the cognate words English *token* [təʊkən] and German *Zeichen* [ʦaɪçən] 'sign'. Although these words sound very different, it is easy to show that the sounds regularly correspond to each other, as can be seen from English *weak* [wiːk] vs. German *weich* [vaɪç] 'soft' for the correspondence of [k] with [ç], and English *tongue* [tʌŋ] vs. German *Zunge* [ʦʊŋə] 'tongue' for the correspondence of [t] with [ʦ]. Genotypic similarity is quite similar to the relation between a source text and its encryption, where all characters may refer regarding their substance, although they are related by an underlying distinct mapping.

Thus, while the alignment problem in biology can be stated under the assumption that two sequences are both drawn from the same alphabet (e.g. proteins), the alignment problem in linguistics is essentially the problem of aligning two sequences drawn from two *different* alphabets. Although from a general perspective cognate sets and gene families are the same kind of classes of objects, practically they cannot be detected, hence studied alike. Indeed, members of both cognate sets and gene families share the extrinsic, relational, property of originating from the same common ancestor; yet this historical essence of cognates and of gene families does not translate into the definition of sets of objects with intrinsic exclusive properties. There is no obligate (nor strong) correlation between 'having the same origin' and 'sounding alike' for words, while there is a stronger correlation between 'having the same origin' and 'having closer genetic sequences than anyone else' for genes. This difference implies that while the classification of genes into gene families allows for some generalizations about the members of the gene family, the classification of words into cognate sets allows for less generalization regarding their phonetic similarity. This difference may not come as a surprise, since the limited amount of phonemes and the limited size of words (as opposed to the high combinatorials of DNA bases in relatively longer sequences) makes such convergences expected. However, this high level of phonetic convergence means that word network based on phonetic distances are unlikely to be as discriminating as gene networks based on genetic distances. While the latter can be used to infer

classes of undetected (hidden) homology, it seems more problematic to use the former to infer undetected (hidden) cognacy, without the development of specific, genotypic, distances, adapted to the objects of linguistics. Moreover, the fact that gene families can be defined both based on correlated extrinsic and intrinsic properties, while cognate sets seem to be mostly characterized by extrinsic properties raises questions on whether these objects can be used alike for explanations, descriptions and inductive inferences in the fields of biology and linguistics, and if not, whether the analogy between cognate and gene families should not be also refined.

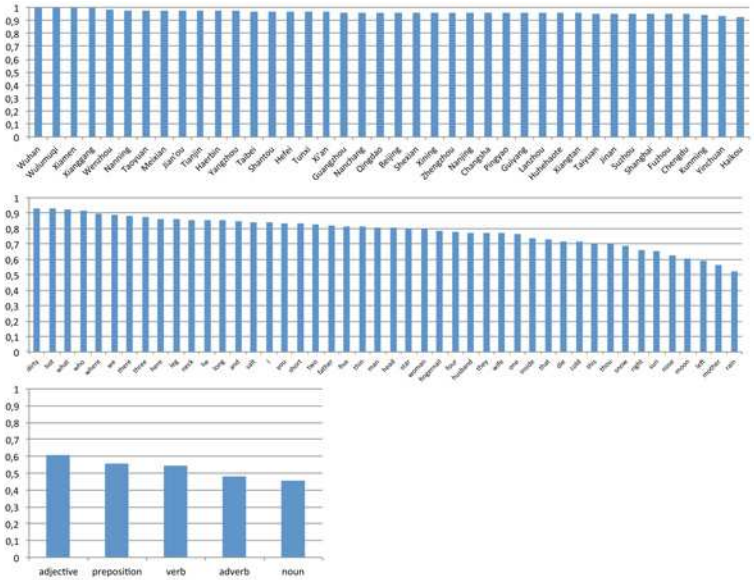


*Figure 5: Conductance of the meanings, dialects and grammatical types in the Chinese word network.*

Conductance for each item (x-axis) is indicated on the y-axis, and computed as described in the text. Significance was assessed by shuffling the labels of the original network, then computing the various conductances (Top: dialect, Middle: meaning, Bottom: grammar) on this randomized network. The procedure was repeated 1,000 times to obtain a normal distribution of conductances for random classes of the same size as the tested classes. Except for the 11 leftmost dialects (Wuhan to Haerbin), all observed conductances were less than 2 sigma below the mean of their corresponding normal distribution, meaning that the estimated conductance was not a mere effect of the sample size.

Importantly however, the fact that the Chinese word network based on phonetic distances is not strongly structured by cognate sets does not mean that this network does not show another type of informative structure. We classified the words into three functional categories to test whether, in spite of this high amount of phonetic convergence, phonetic properties of the words were not random, suggesting some rules and selective pressures on phoneme combinations. To this end, each word was labeled based on its meaning, dialect of origin, and grammatical type (Adjective,

Adverb, Noun, Preposition, Verb), and the conductance of each of these labels in the network was assessed (Figure 5). All dialects presented a very high conductance (close to 1) indicating that they cannot easily be distinguished based on the phonetic similarity between their words: words from different dialects can sound very close, or to put it differently, no dialect shows strong phonetic consistency. This is not surprising, since high phonetic similarity between the words of a single language would make it difficult for the speakers to communicate. Meanings also had high conductances, however lower than those of dialects, and some meanings ("right", "sun", "nose", "moon", "left", "mother" and "rain") had even relatively low conductances, testifying that some combinations of phonemes were preferentially associated with these meanings. Given that words denoting these meanings are usually highly preserved, often going back to the same ancestor form in all Chinese dialects, this result is also not unexpected. Strikingly, grammatical types present the lowest conductances of all the tested classes and structure the similarity network more strongly than all the other types we checked (Figure 6). A mechanical explanation for this might lie in the distance measure that we used. Although normalized for word length, the distance measure is still rather sensitive to the comparison of words that have an equal length, yielding lower distance scores for words with a similar length. Since average word length tends to be very similar for parts of speech in Chinese (prepositions usually consist only of one syllable, nouns usually have two syllables), this may also be a reason for the low conductance of words corresponding to the same part of speech.

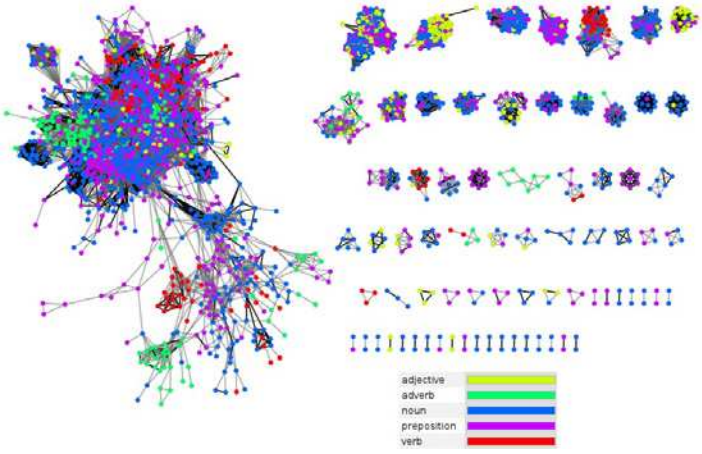


*Figure 6: Network of close phonetic proximity with both cognate and similarity edges colored by grammatical types.*

Nodes correspond to words, colored by grammatical types, connected by cognate edges (in black) and similarity edges (in grey) indicating a close phonetic proximity (distance < 0.35) between pairs of words. This figure can be contrasted to Figure 4 to verify that meanings are less phonetically structured than grammatical types.

## CONCLUSION & PERSPECTIVES

In both the fields of evolutionary biology and linguistics, graphs appear as excellent tools for the exploratory analysis of evolving objects, be they words or genes. However, because the evolution of words is affected by more convergences than the evolution of genes, specific adjustments seem still to be required to integrate networks in the toolkit of linguistic studies. In particular, unsupervised automatic cognate detection might prove much harder than gene family detection. The investigation of the phonetic diversity of cognate words from Chinese dialects with relatively simple networks however was already powerful enough to identify different phonetic structures at different levels of linguistic organization. Dialects, meanings and grammatical categories seem subjected to distinct intensities of selective pressures, affecting the diversity of phonemes used in their making, in ways that now deserved to be explained.

## MATERIAL AND METHODS

### Dataset

The data that we used for our analysis is taken from the Hànyǔ Fāngyán Yīnkù (Hóu 2004), a CD-ROM that offers many different resources on Chinese dialects, including phonological descriptions, phonetic transcriptions, and sound recordings for 40 different dialect varieties. From the CD-ROM, we extracted a part of the lexical subset, consisting of 48 glosses ("concepts") translated into the 40 varieties. These 48 glosses belong to the basic vocabulary in the strict sense of Swadesh (Swadesh 1952; Swadesh 1955). Chinese dialects often have a lot of synonyms for the very same concept; therefore the resulting dataset is made of 2,999 words in total. The source material was given in a format not tractable for computational analyses. Therefore, the extraction procedure was carried out semi-automatically, applying additional manual cleaning. All entries were double-checked by comparing the phonetic transcription for each word with its corresponding sound recording. The data was further enriched by looking up the grammatical categories of the glosses, translating the glosses into English, adapting the phonetic transcriptions to plain IPA, and applying a rough procedure for automatic cognate detection that is described in the following section.

### Cognate Judgments

In Chinese dialectology it is common to give not only the pronunciation of a given dialect word, but also an assessment regarding its homology. Homology assessments are usually coded by giving the Chinese characters corresponding to a given word. Since for most Chinese characters the Middle Chinese readings (spoken around the 6th century) can be reconstructed from old rhyme books, a character is

somewhat similar to a proto-form. Thus, Táoyuán [ŋit$^{22}$t$^{h}$eu$^{11}$] and Hǎikǒu [zit$^{3}$hau$^{31}$] "sun" are both written as 日头, and the proto-form would have been pronounced as *ȵit$^{4}$duw$^{1}$ in Middle Chinese times (if the compound was already present during that time). Note that the character assignments in Chinese dialectology are homologs in the strict sense, since no distinction between borrowing and vertical inheritance is drawn. Using this procedure, the 2,999 words could be grouped into 337 cognate sets.

## Phonetic distances

Phonetic distances between all words were calculated using the SCA method (List 2012) to derive alignment scores, and the formula by Downey et al. (Downey et al. 2008) to convert similarity into distance scores. The resulting distance measure is "phenotypic" in the sense of Lass (Lass 1997) in so far as it is language-independent, neglecting the presence or absence of previously established sound-correspondence patterns. However, it is based on an enhanced function for the scoring of phonetic segments, and previous studies (List 2012) could show that it outperforms alternative distances measures, such as the normalized Levenshtein distance (Levenshtein et al. 2010), or the measure underlying the cognate detection method by Turchin et al. (Turchin et al. 2010). Therefore, this distance measure seems to be a more reliable basis for network applications than alternative ones.

## Network visualization and analyses

The network layouts were produced by Cytoscape software (Smoot et al. 2011), using force directed layouts. Conductances were computed as: $C = N_{ext} / (N_{ext} + 2*N_{int})$, where $N_{int}$ is the number of internal edges (e.g. between members of that cognate or gene family) and $N_{ext}$ is the number of external edges (e.g. between a member from that cognate/gene family and a member from another cognate/gene family). Significance of these conductances was assessed by shuffling the labels of the original network, then computing the various conductances (dialect, meaning, grammar) on this randomized network. The procedure was repeated 1,000 times to obtain a normal distribution of conductances for random classes of the same size than the tested classes. Unless specified otherwise, most observed conductances were more than 2 sigma lower than the mean of their corresponding normal distribution, meaning that the conductance values are not a mere effect of the sample size.

## REFERENCES

Alvarez-Ponce D. & J.O. McInerney (2011) 'The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences', Genome biology and evolution 3: 782–90.

Alvarez-Ponce D., P. Lopez, E. Bapteste & J.O. McInerney (2013) 'Gene similarity networks provide new tools for understanding eukaryote origins and evolution.', Proceedings of the National Academy of Sciences of the United States of America 110(17): E1594–603.

Bapteste, E. et al. (2008) 'Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny.', Mol Biol Evol 25(1): 83–91.

Bapteste, E. et al. (2012) 'Evolutionary analyses of non-geological bonds produced by introgressive descent', Proceedings of the National Academy of Sciences of the United States of America 109(45): 18266–72.

Bapteste, E., F. Bouchard & R.M. Burian (2012) 'Philosophy and Evolution : Minding the Gap Between Evolutionary Patterns and Tree-Like Patterns', in M. Anisimova (ed), Evolutionary Genomics: Statistical and Computational Methods, Vol.2 (New York, Humana Press: Springer): 81–112

Bapteste, E., M. O'Malley & R. Beiko (2009) Prokaryotic evolution and the tree of life are two different things, Biol Direct. 4:34.

Beauregard-Racine, J. et al. (2011) 'Of woods and webs: possible alternatives to the tree of life for studying genomic fluidity in E. coli', Biol. Direct 6(1): 39.

Bittner, L. et al. (2010) 'Some considerations for analyzing biodiversity using integrative metagenomics and gene networks', Biol Direct 5:47.

Burian, R.M. (2011) 'Experimentation, Exploratory', in W. Dubitzky, O. Wolkenhauer, K.-H. Cho & H. Yokota, Encyclopedia of Systems Biology (in press).

Dagan, T. & W. Martin (2009) 'Getting a better picture of microbial evolution en route to a network of genomes', Philos Trans R Soc Lond B Biol Sci 364(1527): 2187–96.

Dagan, T., W. Martin (2006) 'The tree of one percent', Genome Biol 7(10): 118.

Dagan, T., Y. Artzy-Randrup & W. Martin (2008) 'Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution', Proceedings of the National Academy of Sciences of the United States of America 105(29): 10039–44.

Downey, S.S., B. Hallmark, M.P. Cox, P. Norquest & S. Lansing (2008) 'Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction', Journal of Quantitative Linguistics 15(4): 340–69.

Fondi, M. & R. Fani (2010) 'The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks', Environ Microbiol 12(12): 3228–42.

Franklin-Hall, L. (2005) 'Exploratory Experiments', Philosophy of Science 72: 888–99.

Gallagher B. & T. Eliassi-Rad (2008) 'Leveraging Label-Independent Features for Classification in Sparsely Labeled Networks: An Empirical Study.', SNA-KDD '08.

Ghigo, J.M. (2001) 'Natural conjugative plasmids induce bacterial biofilm development', Nature 412(6845): 442–5.

Halary, S., J.W. Leigh, B. Cheaib, P. Lopez & E. Bapteste (2010) 'Network analyses structure genetic diversity in independent genetic worlds', Proceedings of the National Academy of Sciences of the United States of America 107(1): 127–31.

Hall-Stoodley, L., J.W. Costerton & P. Stoodley (2004) 'Bacterial biofilms : from the natural environment to infectious diseases', Nat Rev Microbiol 2(2): 95–108.

Hóu, J. (2004) *Xiàndài Hànyu* fangyán yinkù [Phonological database of Chinese dialects] (Shanghai: Shànghai Jiàoyu).

Huang, J. & J.P. Gogarten (2007) 'Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids ?' Genome Biol 8(6): R99.

Jachiet, P.A., R. Pogorelcnik, A. Berry, P. Lopez & E. Bapteste (2013) 'MosaicFinder: identification of fused gene families in sequence similarity networks', Bioinformatics.

Jones, B.V. (2010) 'The human gut mobile metagenome: a metazoan perspective', Gut Microbes 1(6): 415–31.

Kloesges, T., O. Popa, W. Martin & T. Dagan (2011) 'Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths', Mol Biol Evol 28(2): 1057–74.

Koschützki, D. (2008) 'Network Centralities', in B.H. Junker & F. Schreiber, Analysis of Biological Networks (Hoboken, NJ: John Wiley & Sons, Inc.): 65–84.

Lass, R. (1997) Historical linguistics and language change (Cambridge: Cambridge University Press).

Leigh, J.W., F.J. Lapointe, P. Lopez & E. Bapteste (2011) 'Evaluating phylogenetic congruence in the post-genomic era', Genomic biology and evolution 3: 571–87.

Leskovec, J., K.J. Lang, A. Dasgupta & M.W. Mahoney (2008) 'Statistical properties of community structure in large social and information networks.', Proc. 17-th International WWW: 695–704.

Levenshtein, V.I. (1966) 'Binary codes capable of correcting deletions, insertions, and reversals', Soviet Physics Doklady 10(8): 707–10.

Lima-Mendez, G., J. Van Helden, A. Toussaint & R. Leplae (2008) 'Reticulate representation of evolutionary and functional relationships between phage genomes', Mol Biol Evol 25(4):762–77.

List, J.-M. (2012) 'LexStat. Automatic Detection of Cognates in Multilingual Wordlists.', Joint Workshop of LINGVIS & UNCLH: 117–25.

List, J.-M. (2012) 'SCA. Phonetic alignment based on sound classes', in M. Slavkovik & D. Lassiter (eds), New directions in logic, language and computation (Berlin/Heidelberg: Springer): 32–51.

Lozupone, C.A. et al. (2008) 'The convergence of carbohydrate active gene repertoires in human gut microbes', Proceedings of the National Academy of Sciences of the United States of America 105(39): 15076–81.

Marin, B., E.C. Nowack & M. Melkonian (2005) 'A plastid in the making: evidence for a second primary endosymbiosis', Protist 156(4): 425–32.

Martha, V.S. et al. (2011) 'Constructing a robust protein-protein interaction network by integrating multiple public databases', BMC Bioinformatics 12 Suppl. 10:S7.

Martin, W. et al. (2007) 'The evolution of eurakyotes', Science 316(5824): 542–3.

Moustafa, A. et al. (2009) 'Genomic footprints of a cryptic plastid endosymbiosis in diatoms', Science 324(5935): 1724–6.

Nelson-Sathi, S. et al. (2011) 'Networks uncover hidden lexical borrowing in Indo-European language evolution', Proceedings. Biological sciences / The Royal Society 278(1): 1794–803.

O'Malley, M.A. & E.V. Koonin (2011) 'How stands the Tree of Life a century and a half after The Origin?' Biol Direct 6: 32.

Periasamy, S. & P.E. Kolenbrander (2009) 'Aggregatibacter actinomycetemcomitans builds mutualistic biofilm communities with Fusobacterium nucleatum and Veillonella species in salvia', Infect Immun 77(9): 3542–51.

Qu, A. et al. (2008) 'Comparative metagenomics reveals host sprecific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome', PloS One 3(8): e2945.

Skippington, E. & M.A. Ragan (2011) 'Lateral genetic transfer and the construction of genetic exchange communities', FEMS Microbiol Rev 35(5): 707–35.

Smoot, M.E., K. Ono, J. Ruscheinski, P.L. Wang & T. Ideker (2011) 'Cytoscape 2.8: new features for data integration and network visualization', Bioinformatics 27(3): 431–2.

Swadesh, M. (1952) 'Lexicostatistic dating of prehistoric ethnic contacts', Proceedings American Philosophical Society 96: 452–63.

Swadesh, M. (1955) 'Towards greater accuracy in lexicostatistic dating', International Journal of American Linguistics 21: 121–37.

Turchin P., I. Peiros & M. Gell-Mann (2010) 'Analyzing genetic connections between languages by matching consonant classes', Journal of Language Relationship 3: 117–26.

Vinayagam A. et al. (2011) 'A directed protein interaction network for investigating intracellular signal transduction.', SciSignal 4(189): rs8.
Wang, T.Y., F. He, Q.W. Hu & Z. Zhang (2011) 'A predicted protein-protein interaction network of the filamentous fungus Neurospora crasse', Mol Biosyst 7(7): 2278–85.
Wintermute, E.H. & P.A. Silver (2010) 'Dynamics in the mixed microbial concourse', Genes Dev 24(23): 2603–14.
Wu, D. et al. (2011) 'Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees', PloS One 6(3): e18011.
Zhaxybayeva. O. & F. Doolittle (2010) 'Metagenomics and the Units of Biological Organization', BioScience 60(2): 102–12.

## ACKNOWLEDGEMENTS

We thank Thorsten Halling for initiating collaboration between the three of us, and Bill Martin and Tal Dagan for their pioneering work in linguistics using language networks.