

two copies of a disease mutation develop sickle-cell anaemia, but individuals with one copy<sup>8</sup> have partial resistance against malaria. It has also been proposed that the mutations that cause cystic fibrosis<sup>9</sup> and Tay–Sachs<sup>10</sup> disease might reduce susceptibility to typhoid and tuberculosis, respectively, in carriers with only one copy of the mutation. Detailed statistical and evolutionary analyses of Hinds and colleagues' data may suggest which other genomic regions, or genes, are most likely to harbour disease-associated mutations, even without the help of any data on the consequences of the mutations.

The availability of large-scale SNP data therefore brings together the disciplines of medical and evolutionary genetics. Evolutionary thinking underlies many of the common methods used for identifying associations between genetic types and observable traits or diseases using population genetic data, and has led to major advances in genetic epidemiology<sup>11,12</sup>. And with the increased emphasis

on identifying the determinants of genetic diseases comes the awareness that human genomic variation can only be understood fully in light of the evolutionary forces that have shaped it. The data published by Hinds *et al.*<sup>1</sup> will provide a unique resource for explorations in both disciplines. ■

Rasmus Nielsen is at the Centre for Bioinformatics, University of Copenhagen, Universitetsparken 15, 2100 Kbh Ø, Copenhagen, Denmark.  
e-mail: rasmus@binf.ku.dk

1. Hinds, D. A. *et al.* *Science* **307**, 1072–1079 (2005).
2. Cann, R. A., Stoneking, M. & Wilson, A. C. *Nature* **325**, 31–36 (1987).
3. Nielsen, R., Todd, M. J. & Clark, A. G. *Genetics* **168**, 2373–2382 (2004).
4. Gibbs, R. A. *et al.* *Nature* **426**, 789–796 (2003).
5. Sawyer, S. L. *et al.* *PLoS Biol.* **2**, e275 (2004).
6. Enard, W. *et al.* *Nature* **418**, 869–872 (2002).
7. Halushka, M. K. *et al.* *Nature Genet.* **22**, 239–247 (1999).
8. Friedman, M. J. & Trager, W. *Sci. Am.* **244**, 154–164 (1981).
9. Pier, G. B. *et al.* *Nature* **393**, 79–82 (1998).
10. Rotter, J. I. & Diamond, J. M. *Nature* **329**, 289–290 (1987).
11. Slatkin, M. *Genetics* **137**, 331–336 (1994).
12. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. *Am. J. Hum. Genet.* **67**, 170–181 (2000).

## Language

# Syntax for free?

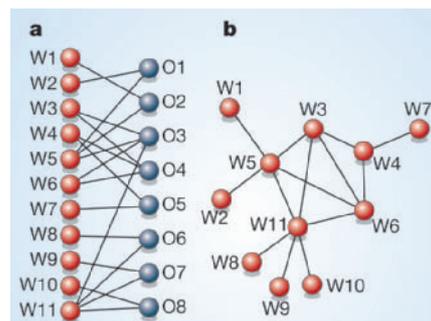
Ricard Solé

Human language is based on syntax, a complex set of rules about how words can be combined. In theory, the emergence of syntactic communication might have been a comparatively straightforward process.

Language does not fossilize — for all that it was one of the great transitions in evolution<sup>1</sup>, the advent of language has left no obvious equivalent to fossil teeth and bones, and seems inaccessible to enquiry. But it is not hard to imagine the emergence of a set of signals to label objects, the combinatorial nature of which allowed an infinite repertoire of sentences to be constructed from a finite set of words. An essential part of the process must have been the acknowledgement of a set of rules to combine words in such a way as to make sentences meaningful. These rules are the syntax that we all easily learn as children, but students of language evolution have a tough time explaining its origins.

So, how did syntax originate? Some authors have conjectured that it resulted from some pre-adaptation of the human brain<sup>2</sup> — that the first step was the building of a rich lexicon upon a cognitive system predisposed to formulate rules able to exploit the underlying combinatorial features of word associations. According to a study by Ramon Ferrer i Cancho and colleagues<sup>3</sup>, however, a simple word–object association matrix can provide the basis for syntax almost for free.

The authors use a simple model in which signals (words) are associated with objects. Such association can be referential (such as *meat* referring to 'edible organic matter') or



**Figure 1 Building a protolanguage network.** a, A bipartite set of connections can be built by linking signals (words, red) to objects (blue). Most words are specific, referring to only one or two objects, whereas a few of them have many links. b, A new network is obtained by linking words that share at least one object of reference. The resulting network has many words with few links, but some acting as hubs. Ferrer i Cancho *et al.*<sup>3</sup> believe that syntactic rules might have emerged from such a scale-free network architecture.

non-referential (*eat* is associated with 'action of eating'). Some words display both associations: *eat*, for example, is also linked to 'edible organic matter'. A bipartite set of connections is obtained linking the two sets (Fig. 1a).

Ferrer i Cancho *et al.* show that most of the architecture of this network stems from a

universal feature common to all languages, Zipf's law. Briefly stated, Zipf's law establishes that if we take all the words in a text and order them by rank, from the most common to the rarest, the frequency (number of appearances) decays inversely with their rank<sup>4</sup>. Most words are rare, whereas a few (such as *the*, *of*, *and*, *to*, *I*) are very common. Common words are less specific and can be linked to many objects, whereas rare ones are more specific and have one or a few links. The number of links of a given word, not surprisingly, follows the same law.

This set of linkages does not provide any obvious clue to the origins of syntax. But it allows another network to be built that incorporates a primitive form of word–word association. The new network naturally links names and actions, because pairs such as *eat* and *meat* will be connected. This is simply done: two words will be linked if they share at least one object of reference (Fig. 1b). This is of course a very rough way of associating symbols, but, as Ferrer i Cancho *et al.* show, the architecture of the resulting network seems to be surprisingly close to many features exhibited by linguistic networks.

Previous studies<sup>5</sup> have demonstrated that networks obtained by linking words exhibit seemingly universal patterns of organization, provided the words are syntactically related: two words are linked if they have been syntactically combined in a collection of sentences. Different languages share the same scale-free structure, with most words having few syntactic links and a few of them being connected to many others.

Ferrer i Cancho and colleagues' network displays the same structure as its real counterparts. Exactly the same distribution of links is found, suggesting the possibility that early 'protolanguage' might have been ready-made for the development of a full syntax. If this is so, the sometimes illogical and quirky appearance of syntactic rules might be nothing but a by-product of scale-free network architecture. The study also suggests that Zipf's law could have been a precondition for syntax and symbolic communication. Once such a condition was met, the basis for the combinatorial explosion characteristic of human language was ready for selection to shape it. The new theory will be subject to debate, but the remnants of the communicative Big Bang are evidently hiding somewhere inside modern language networks. ■

Ricard Solé is in the Complex Systems Laboratory, ICREA-Universitat Pompeu Fabra, Barcelona 08003, Spain.  
e-mail: ricard.sole@upf.edu

1. Maynard-Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Oxford Univ. Press, 1995).
2. Bickerton, D. *Language and Species* (Univ. Chicago Press, 1990).
3. Ferrer i Cancho, R., Riordan, O. & Bollobás, B. *Proc. R. Soc. Lond. B* doi:10.1098/rspb.2004.2957 (2005).
4. Zipf, G. K. *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Cambridge, Massachusetts, 1949).
5. Ferrer i Cancho, R. *et al.* *Phys. Rev. E* **69**, 051915 (2003).