Morpheme networks reveal language dynamics

Daniela Keller[1], Jörg Schultz[1]

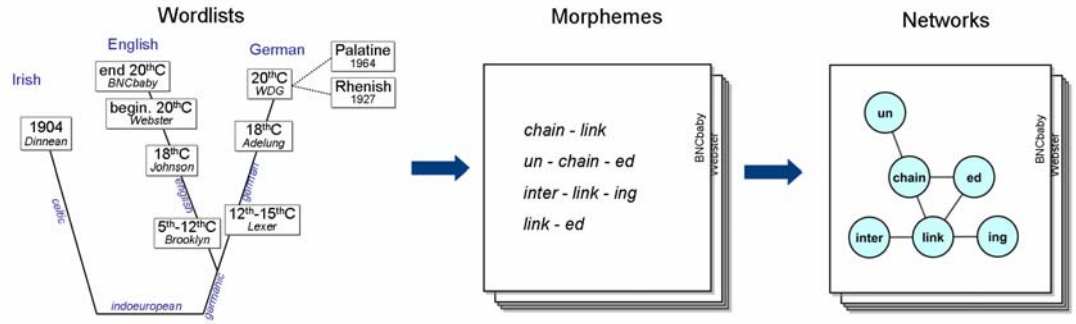
[1]Department of Bioinformatics, Biocenter
University of Würzburg,
97074 Wuerzburg, Germany.

Abstract

The existence of analogous structures between language and biology has been fiercely debated and still is. Anyhow, when looking at morphemes—the smallest meaningful parts of words—and how they are combined to build words, the analogy to domains—the structural, functional and evolutionary units of proteins—suggests itself.
Here, we analyzed three different indo-european languages and compared variants of these languages differing in time and region (Figure 1). In contrast to biology, we were able to look back in time and investigated the evolution of English and German over about 1000 years. For this purpose, we adopted a network based approach frequently used for domains[1,2] with morphemes as nodes and an edge between two morphemes if thy appear side by side in a word (Figure1).
We found global similarity of the networks with all networks being small world, scale free, hierarchical and disassortative. However, we identified a classifier based on network properties to distinguish between languages according to their complexity of word formation (Figure 2). Considering the topological features of the morphemes, differences between evolutionary steps showed cultural change. Upcoming hubs—for example car (‘Auto’), sport (‘Sport’) and movie (‘Film’) in present German—indicate the emergence of new concepts in a language. Apart from these exceptions, word evolution happens primarily at the border of the morpheme networks. Furthermore we identified the rewiring of common morphemes as a main feature of word evolution with 54 % to 75 % changed edges between two evolutionary steps. Comparisons of the networks again revealed that the fate of a morpheme is mainly determined by its connectivity (Figure 3). This contrasts words, whose evolution is driven by usage[3,4].
These results show that the analysis of morphemes will be fundamental not only in the understanding of the evolution of words, but also of meaning, just as the analysis of domains gave fundamental insights into the evolution of proteins.

Figures

Figure 1



Generating morpheme networks. **Left**: Dictionaries and corpora data represented as a genealogical tree of languages. **Middle**: Wordlists with decomposed words. **Right**: Resulting network form decomposed wordlist.

Figure 2

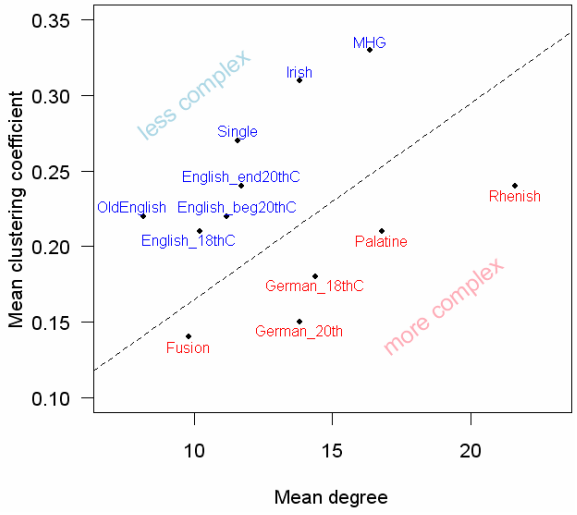
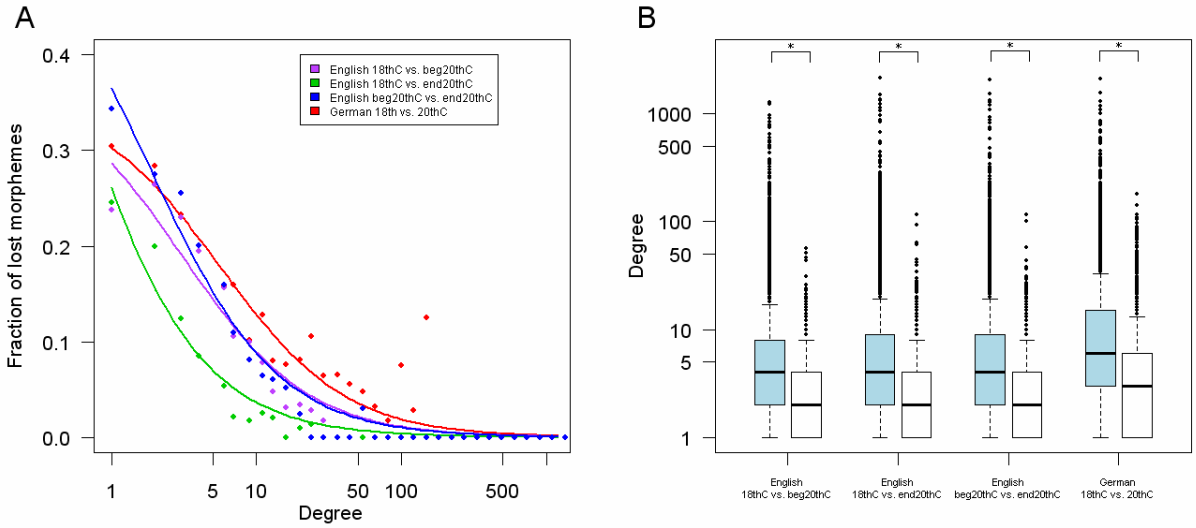


Figure 3



Connectivity of lossed and gained morphemes**. A** Dependence of loss of morpheme on degree. Fit of the function y=a/(x+b) with least squares. **B** Comparison of degree of common (blue) and gained (white) morphemes. * show highly significant difference.

References

1. Wuchty, S. Scale-free behavior in protein domain networks. *Mol Biol Evol* **18**, 1694-702 (2001).
2. Wuchty, S. & Almaas, E. Evolutionary cores of domain co-occurrence networks. *BMC Evol Biol* **5**, 24 (2005).
3. Pagel, M., Atkinson, Q. D. & Meade, A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717-20 (2007).
4. Lieberman, E., Michel, J. B., Jackson, J., Tang, T. & Nowak, M. A. Quantifying the evolutionary dynamics of language. *Nature* **449**, 713-6 (2007).