Reconstructing Concept Networks on the Basis of Cross-Linguistic Polysemy

JOHANN-MATTIS LIST, ANSELM TERHALLE
(Heinrich-Heine-University Düsseldorf)

What do "milk" and "udder" have to do with each other? Conceptually, they are closely related, since the former is the product and the content of the latter. Linguistically, they may even look the same, being referred to by identical word forms in many different languages, such as, e.g., by [nax] in Judeo-Tat (an Indo-European language), by [ukun] in Oroqen (an Altaic language), or by [mis] in Miao (a Hmong-Mien language).[1] Historically, the conceptual relation between "milk" and "udders" may show up in the form of *semantic shifts* where a word which was formerly used to express one of the concepts in a given language is henceforth used to express the other one. Thus, in Standard Chinese, the word for [niou$_{35}$nai$_{214}$] "milk" is a compound of [niou$_{35}$] "cow" and [nai$_{213}$] "milk" which originally meant "breast".

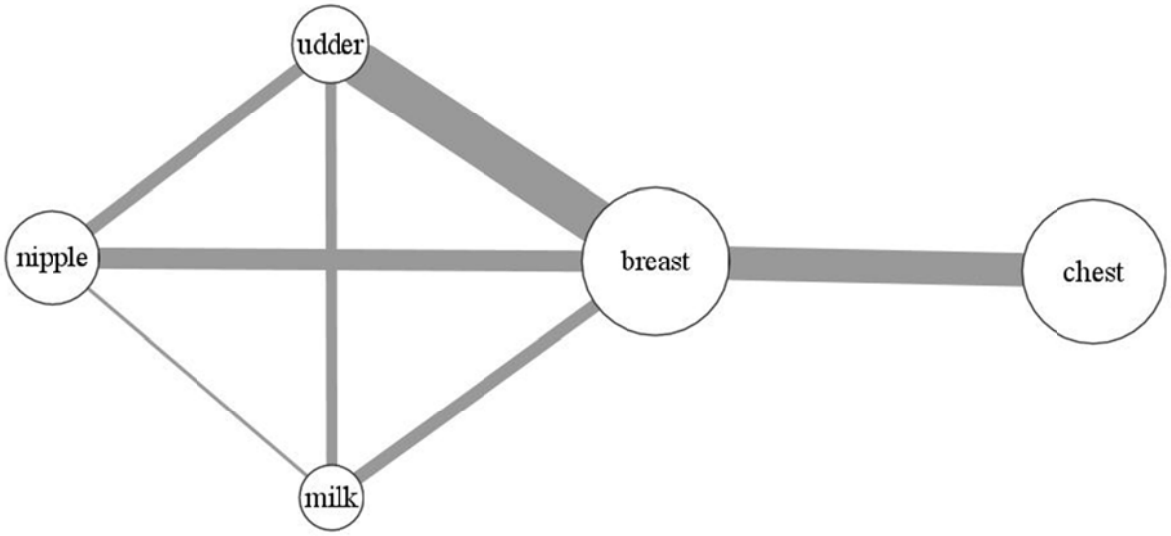

Figure 1: Milk and Udder

We investigate conceptual structures by analyzing cases of polysemy in the lexicon of a large number of the world's languages. The underlying idea, which is largely accepted in cognitive historical semantics, is that polysemy is an important synchronic trace of semantic development and that semantic development in turn reveals information about the human conceptual system (Wilkins, 1996; Blank, 1997). While in traditional historical linguistics the study of concept relations and semantic change is usually based on detailed manual analyses of individual cases, we propose an automatic, quantitative approach which draws upon an idea initially proposed by Steiner, Stadler, and Cysouw (2011). The authors use information on cross-linguistic polysemy in order to establish a metric for conceptual distance. However, in contrast to our approach, they only take immediately related concepts into account, neglecting indirect conceptual connections which can only be modeled by means of network approaches.

[1] The genetic classification of the languages follows Lewis (2009), the data for Oroquen and Judeo-Tat is taken from Key and Comrie (2007).

The concept networks are reconstructed as follows: Based on the lexical data consisting of 1310 concepts translated into 175 languages,[2] we extract all lexical entries which are polysemous in so far as they stand for two or more concepts in each language. In a further step, we count how many times all possible concept pairs are expressed by the same word in all languages. As a result, we obtain a matrix which can be directly translated into a weighted network, as illustrated in Figure 1, where the edge width reflects the number of lexical links and the node size reflects the average number of forms per concept.

Cross-linguistic polysemy networks provide several kinds of interesting information both for linguistics and cognitive science. From a descriptive point of view they offer a new perspective on conceptual structures, reflecting the degree to which concepts are associated. Apart from presenting these major findings, we will address the following questions:

- Is there a correlation between the average number of forms per concept and the number of links to other concepts (i.e., a correlation between node size and node degree)?
- Is there a correlation between the degree of the connectivity of certain regions and the number of languages exhibiting the respective links (i.e., a correlation between the number and the weight of the edges in a given region)?
- How close is the association of the most distant concepts (diameter) of the network? How closely are the concepts associated on average (average shortest path)?
- How neatly are existing linguistic models of semantic change, such as the concepts of *form expansion* or *concept attraction* (Sperber 1923), reflected in our network?

Blank, A. (1997). Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen. Beihefte zur Zeitschrift für romanische Philologie 285. Tübingen: Niemeyer. Haspelmath, M. and U. Tadmor, eds. (2009). Worl Loanword Database. URL: http://wol d.livingsources.org.

Key, M. R. and B. Comrie, eds. (2007). IDS – The Intercontinental Dictionary Series. URL: http://lingweb.eva.mpg.de/ids/.

Lewis, M. P., ed. (2009). Ethnologue. Languages of the World. 16th ed. Dallas: SIL International. Online Version: http://www.ethnologue.com.

Logos Group, ed. (2008). Logos Dictionary. URL: http://www.logosdictionary.or g/index.php.

Sperber, H. (1923). Einführung in die Bedeutungslehre. Bonn and Leipzig: Kurt Schroeder.

Steiner, L., P. F. Stadler, and M. Cysouw (2011). "A pipeline for computational historical linguistics". In: Language Dynamics and Change 1.1, 89–127.

Wilkins, D. P. (1996). "Natural tendencies of semantic change and the search for cognates". In: The comparative method reviewed. Regularity and irregularity in language change. Ed. by M. Durie. With an intro. by M. D. Ross and M. Durie. New York: Oxford University Press, 264–304.

[2] Our data is taken from Key and Comrie (2007), Logos Group (2008), Haspelmath and Tadmor (2009).