Lexico-semantic networks on various levels of resolution

Matthias Urban
Leiden University Centre for Linguistics
Today’s talk

• Part 1: lexico-semantic networks based on large-scale lexical databases using CLICS

• Part 2: fine-grained analyses of lexico-semantic associations within language families. Example: Quechuan internal parts of the torso in diachrony
Part 1

CLICS

CLICS is an online database of synchronic lexical associations [READ MORE]

Database of Cross-Linguistic Colexifications
The CLICS team

Mattis List  

Thomas Mayer  

Anselm Terhalle  

Matthias Urban
Sources

Currently CLICS utilizes four different sources, all of which are freely available online themselves.

- The intercontinental dictionary series ([IDS, Key & Comrie 2007 eds.]) features lexical data for 233 world languages. IDS data were provided mostly by experts on the respective languages, although in some cases published written sources have been used. There are 1,310 entries to be filled for each language, though, of course, there are gaps in coverage for individual languages. The list of concepts is inspired by Buck (1949). Of all 233 languages in the IDS, 178 were automatically cleaned and included in CLICS.

- The IDS list, in turn, provides the basis for the choice of meanings in the World Loanword Database ([WOLD, Haspelmath & Tadmor 2009 eds.]). The principal aim of this source is to provide a basis for generalizations on the borrowability of items in different parts of the lexicon. The WOLD data consist of vocabularies of between 1,000-2,000 items for 41 languages, with annotations about the borrowing history of particular items where applicable. WOLD data was coded by experts on the respective languages, in some cases also with the aid of extant sources. Of all 41 languages in WOLD, 33 are included in CLICS.

- The Logos Dictionary is a freely accessible multilingual online dictionary that is regularly updated online by a network of professional translators. It offers lexical data for more than 60 different languages. We manually extracted lexical data for 4 languages that were neither present in IDS nor in WOLD.

- The Språkbanken project (University of Gothenburg) offers a couple of wordlists for South Asian and Himalayan languages. The wordlists mirror the IDS format closely, and we included 6 of currently 8 wordlists in CLICS.
The situation before CLICS: IDS
The situation before CLICS: WOLD

Vocabulary Imbabura Quechua

The vocabulary contains 1297 meaning-word pairs ("entries") corresponding to core LWT meanings from the recipient language Imbabura Quechua. The corresponding text chapter was published in the book Loanwords in the World's Languages. The language page Imbabura Quechua contains a list of all loanwords arranged by donor languoid.
Colecification

In the context of CLICS, we use the term *colexification* (coined by François 2006) to refer to the situation when two or more of the meanings in our lexical sources are covered in a language by the same lexical item. For instance, we would say that Russian *pyka* colexifies ‘hand’ and ‘arm’, that is, concepts that are semantically related to each other. Roughly spoken, colexification can correspond either to polysemy or semantic vagueness in lexical semantic analyses. Since we have not performed such analyses that would allow us to further discriminate between the two, we chose *colexification* as a label that deliberately does not make a commitment with regard to this distinction. However, we offer measures to rule out effects of accidental homonymy.

Conceptual links, which constitute our main interest in colexification, are clearly not the only possible reason why two concepts can be expressed by the same form in a given language. The most prominent among the others is *homonymy*, which is in very simple terms an accidental similarity in form, often arising from sound changes causing originally distinct lexical items to collapse phonologically. Such cases may lead to spurious links in the database (compare, for example, the links between the concepts ‘arm’ and ‘poor’ which are due to homonymy in some Germanic languages). To deal with this issue in a consistent and generally applicable way (even if the history of the lexical items in question is not known) we recommend to employ a typological criterion to distinguish between homonymy and polysemy (see Croft 1990): for semantic connections to be accepted as genuine rather than accidental, the connection should be detectable in more than one language family. We should, however, point out that this criterion has been developed and originally applied by the aforementioned author in the realm of polyfunctionality of grammatical markers, that is, items belonging to a paradigmatically relatively well-structured set of items with a manageable semantic range. When applied to lexical meanings, there is a danger that the criterion rules out a set of genuine, but simply rare, semantic associations. Still, we feel that our approach is justified, methodologically because it offers a simple and non-subjective decision criterion, and conceptually because our approach relies on cross-linguistic data in the first place.
Some characteristics of the data (inherited from major sources)

• synchronous
• predefined meanings
• resolution cannot be changed
• no balancing for language families/areas
Some (imagined) characteristics of research done on the basis of CLICS

- Worldwide scope (resolution for areal studies probably too low)
- Broad-scale typological-comparative, no expertise of research on individual languages and their history
Part II: Quechuan terminology for internal organs of the torso
The Quechuan language family

http://www.quechua.org.uk/Eng/Main/i_DIALS.HTM
Major Andean ecological zones

Figure 6.2. Cross-section of the Central Andes showing the major environmental zones according to Pulgar Vidal (1987). (Adapted from Richardson 1994).

Quilter (2014)
Traditional Central Andean economical and sociological organization

- The Qollahuaya ayllu (traditional Andean kin-group) of Kaata consists of settlements at high, mid, and low ecological zones of the Andes.

- Triple mapping between environment, community structure, and body: in analogy to the body, the upper level has a “head,” a “mouth” and “eyes,” the central one a “stomach” and a “heart,” and the lower level “feet” and “toenails.”

- The central settlements consist of several hamlets which “are joined together as are the vital organs surrounding the heart”, and “Kaatapata, the oldest and highest of these hamlets, forms the liver.”
Turning to lexical data...
• The Quechua term corresponding to ‘heart’ at Kataa is sonco.

• “Although the sonco is translated as “heart,” Qollahuayas refer to it as a compression-distillation center, which preforms circulatory, respiratory, and digestive processes (The sonco is also associated with thought and emotions, which are qualities of the fluids) ...Within the sonco, fat is separated from food by the force of the blood coming together in an inward spiral of centripetal movement. This force, then, reverses itself in outward movement and disperses the fat and blood to the parts of the body.” (Bastien 1985: 598-599)

• Intension vs. extension
• in colonial texts and also in the dictionary of Santo Tomás (1560), the <shonqo> is associated with the indigenous concept of *camac*, the life force permeating all beings classified as animate in Quechua ontology
Some different configurations of the entire field in different languages...
Junín-Huanca Quechua (based on Cerrón-Palomino 1976)
Imbabura Quichua (based on Gómez Rendón 2009)
Heart of humans and heart of animals

- Junín-Huancapuywan ‘heart of big animals’ (Cerrón-Palomino 1976)

- Ayacucho (Cangallo)puywan ‘heart of big animals’ (Parker 1969)

- Ancash-Huailaspuywan ‘heart of big animals’ (Parker and Chávez 1976)

- 16th century Cuzco Quechua<puyhuá> ~ <puyhuan> is ‘heart of animals’ (González Holguin 1608)

- <puyhuan soncoyok> “a man like a beast, without understanding”
Huallaga (based on Weber et al. 1998)
(similar: Tarma, Adelaar 1977)
What was the meaning of *puywan?

- Wilkins’s (1996: 274) “natural tendency” of semantic change no. 4:

- “It is a natural tendency for the term for an animal part to shift to refer to a person part”
Conclusions

• lexico-semantic organization in a given language or family stems from the complex interplay
  • culture-specific conceptualizations,
  • the cognitive-perceptual apparatus shared by all humans,
  • linguistic history (cf. Koptjevskaja-Tamm 2016: 2),
Conclusions

• To access and understand different aspects of lexico-semantic associations, we need different and complementary study designs at various levels of resolution
References


- González Holguín, Diego. 1608. Vocabulario de la lengua general de todo el Perv llamada lengua Qquichua, o del Inca. Ciudad de los Reyes: Francisco del Canto.


References


• Santo Tomás, Domingo de. 1560. Lexicon, o vocabulario de la lengua general del Perv. Valladolid: Francisco Fernandez de Cordoua.
