Weighted lexical semantic maps for areal lexical typology

Verbs of perception and cognition as a case study
Outline of the talk

- Semantic maps
  - What are semantic maps?
  - Why use semantic maps for areal lexical typology?

- Case study: verbs of perception and cognition
  - Why is this semantic field interesting?
  - The datasets
    - CLICS
    - Vanhove (2008)
    - WORDNET

- Areal patterns and general discussion
Introduction

Semantic maps
Semantic maps

Figure 1. A semantic map of typical dative functions / the boundaries of English to and French à
(based on Haspelmath 2003: 213, 215)
Semantic maps

Weighted semantic maps

Figure 2a. A simple semantic map of person marking (Cysouw 2007: 231)

Figure 2b. A weighted semantic map of person marking (Cysouw 2007: 233)
Figure 3. Overlapping polysemies: Eng. straight vs. Fr. droit (François 2008: 167)

Colexification = multifunctionality

Languages differ as to which senses they colexify
Figure 4. Semantic map for time-related senses based on the CLICS data
Semantic maps

Why use semantic maps for areal lexical typology?

- The map makes universal claims
  - Frequency: attested vs. non-attested and frequent vs. rare
  - Types of polysemy: possible vs. impossible
  - Implicational hierarchies (unlike other colexification networks)

- The mapping of language specific items allows
  - studying genetic, areal, and culture specific patterns,
  - but it also shows how hard it is to reach statistically significant results
Verbs of perception and cognition
A case study
Why perception & cognition?

- Perception and cognition are among the basic concepts that are lexicalized in the languages of the world (e.g. Swadesh 1952)
- Well-studied domain: our results can be compared (e.g. Sweetser 1990; Evans & Wilkins 2000; Vanhove 2008)
- The relevant literature has revealed both universal and culture-specific patterns
Verbs of perception & cognition

Semantic extensions

Intrafield ( = Intradomain)
(senses: same semantic field)

Interfield ( = Interdomain / Transfield)
(senses: different semantic field)

(based on Wilkins 1996: 274; cf. Matisoff 1978)
Verbs of perception & cognition

Figure 4. Viberg’s sense modality hierarchy for semantic extensions and polysemies of perception verbs (Viberg 1984: 136)

Table. Inventories of the verbs of perception (Viberg 1984: 140)
Verbs of perception & cognition

Interfield extensions

Mind-as-body-Metaphor:

The internal self is understood in terms of the bodily external self (Sweetser 1990: 45)

- Common cross-linguistically (if not universal):
  the connection between VISION and KNOWLEDGE (Sweetser 1990: 45)

Figure 6. The structure of our metaphors of perception (Sweetser 1990: 38)
Verbs of perception & cognition

In Australian languages:
(Evans & Wilkins 2000)

• Cognitive verbs > ‘hear’
  (cf. intrafield extensions confirm the prevalence of vision)

Extensions in cognitive verbs:

• A foot in culture: a relativistic aspect
  (cf. Sweetser 1990)
• A foot in nature: a universal aspect
  (Evans & Wilkins 2000)

The culture sieve:

• “filters” those elements that are in accordance with the premises of a given culture
• “impregnates” the mapping with touches of a culture in contrast with other cultural and social systems
  (Ibarretxe-Antuñano 2013: 324)

Figure 7. The culture sieve (Ibarretxe-Antuñano 2013)
More recent accounts

Vanhove 2008

- Sample of 25 languages (8 phyla); mostly African
  - Intrafield: vision prevails
  - Transfield: the auditory modality prevails
    - Stronger semantic association of hearing and mental perception
  - Implicational universal:
    - Hearing > vision > prehension

[no distinction between controlled activity (e.g. listen) vs. non-controlled experience (e.g. hear); cf. Viberg 1984; 2001]
Wälchli 2016

- **Convenience sample**: Central, East and North European languages
- **Case study**: Auditory and visual perception
  - *Explorative perception verbs* = controlled activity (e.g. listen)
  - *Opportunistic perception verbs* = non-controlled experience (e.g. hear)
  - *Specific perception verbs*: subtype of opportunistic perception verbs

- **Goal**: how the encoding of a specificity distinction may differ cross-linguistically.
  - Particular areal feature for Baltic languages

- **Method**: probabilistic semantic maps based on parallel corpora
More recent accounts

\[ \text{■} = \text{specific ‘hear’,} \quad \text{▲} = \text{non-specific ‘hear’,} \quad \text{○} = \text{‘listen’} \]

Figure 8. Probabilistic semantic map of 44 auditory contexts in Mark based on 64 doculects in English (leb), Lithuanian (1998), Latgalian and Latvian (2012) (Wälchli 2016: 77)
Datasets for building lexical semantic maps
Perception and cognition
**CLICS**

- **N of lgs**: 221
- **N of lg families**: 64
- **N of concepts**: 1280

<table>
<thead>
<tr>
<th>Meaning 1</th>
<th>Meaning 2</th>
<th>N of language</th>
<th>N of forms</th>
<th>language:form</th>
</tr>
</thead>
<tbody>
<tr>
<td>see</td>
<td>know</td>
<td>5</td>
<td>6</td>
<td>aro_std:[ba]//ayo_std:[i’mo?]//haw_std:[ʔike]//mcq_std:[ɓanahe]//mri_std:[kitea]//tel_std:[aarayu]//tel_std:[arayu]</td>
</tr>
<tr>
<td>see</td>
<td>find</td>
<td>15</td>
<td>23</td>
<td>agr_std:[wainat]//arn_std:[pe]//con_std:[’atʰeye]//cwg_std:[yow]//emp_std:[u’nu]//kgp_std:[we]//kpv_std:[addzm]///kyh_std:[mah]//mca_std:[wen]//mri_std:[kitea]//oym_std:[esa]//pbb_std:[uy]//plt_std:[mahita]//pui_std:[duk]//ray_std:[tikeʔa]//rtm_std:[ræe]//sap_Enlhet:[nenjwetay’]//sei_std:[aʔo]//shb_std:[taa]//sja_std:[unu]//swh_std:[ona]//tbc_std:[le]//yag_std:[tiki]</td>
</tr>
<tr>
<td>see</td>
<td>get, obtain</td>
<td>6</td>
<td>6</td>
<td>kgp_std:[we]//mbc_std:[eraʔma]//pbb_std:[uy]//sap_Standard:[akwitayi]//srq_std:[tea]//udi_std:[aƙъсн]</td>
</tr>
</tbody>
</table>

Polysemy data from CLiCs ([http://clics.lingpy.org/download.php](http://clics.lingpy.org/download.php))

(List et al. 2014)
Figure 9. Complete network in CLICS of which SEE is part.
Figure 10. Complete network in CLICS with SEE as a pivot
Waiting for CLICS 2.0 …

(List 2017)

- Increased quantity of data
- Increased quality of data (e.g., links to the Concepticon)
- Include partial colexifications
- Normalize the data which is analysed by CLICS
From CLICS to a more economical map

The economy principle

Given three meanings ($\text{Meaning}_A$, $\text{Meaning}_B$, $\text{Meaning}_C$), if the linguistic items expressing $\text{Meaning}_A$ and $\text{Meaning}_C$ always express $\text{Meaning}_B$, there is no need to draw an edge between $\text{Meaning}_A$ and $\text{Meaning}_C$ (the resulting map will not be triangular, i.e. a vacuous semantic map, with all the meanings connected).

(Georgakopoulos & Polis forthcoming)

**Figure 11.** An abstract semantic map
From CLICS to a more economical map

- The synchronic polysemy patterns are converted into a lexical matrix

```python
Tmap = [Tsenses]
for t in Tclean:
    split_langWord = t[2].split('/\')
    for couple in split_langWord:
        langWord = couple.split(':')
        line = [langWord[0], langWord[1]]
        for i in range (2, len(Tsenses)):
            line.append('0')
        line[Tsenses.index(t[0])] = '1'
        line[Tsenses.index(t[1])] = '1'
        Tmap.append(line)
```

<table>
<thead>
<tr>
<th>Source of constraint</th>
<th>Constraint name</th>
<th>Meaning 1 'SEE'</th>
<th>Meaning 2 'KNOW'</th>
<th>Meaning 3 'GET, OBTAIN'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Araona</td>
<td>ba</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ayoreo</td>
<td>i'mo?</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hawaiian</td>
<td>?ike</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ese</td>
<td>banahe</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Maori</td>
<td>kitea</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Telugu</td>
<td>aarayu</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Kaingang</td>
<td>we</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Macushi</td>
<td>era?ma</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Páez</td>
<td>uy</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sanapaná (Standard)</td>
<td>akwitayi</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sirionó</td>
<td>tea</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Udi</td>
<td>акъсун</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
From CLICS to a more economical map

<table>
<thead>
<tr>
<th>Source of constraint</th>
<th>Constraint name</th>
<th>Meaning 1 ‘SEE’</th>
<th>Meaning 2 ‘KNOW’</th>
<th>Meaning 3 ‘GET, OBTAIN’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Araona</td>
<td>ba</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ayoreo</td>
<td>i‘mo?</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hawai‘ian</td>
<td>?ike</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ese</td>
<td>banaha</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Maori</td>
<td>kitea</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Telugu</td>
<td>aarayu</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Kaingang</td>
<td>we</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Macushi</td>
<td>era?ma</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Páez</td>
<td>uy</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sanapaná (Standard)</td>
<td>akwitayi</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sirionó</td>
<td>tea</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Udi</td>
<td>akъсун</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Weighted semantic map** based on an adapted version of the algorithm suggested by Regier et al. (2013)

Python script β
From CLICS to a more economical map

Figure 12. Full semantic map for the cognition-perception domain, visualized with modularity analysis* (Blondel et al. 2008) in Gephi

* A method to extract the community structure of large networks. Here, the different colors point to modules (also called clusters or communities) with dense connections between the nodes within the network.
From CLICS to a more economical map

- Direct edge between perception verbs denoting *non-controlled experience* (e.g., HEAR, SEE) and cognitive verbs (e.g., UNDERSTAND)

**Figure 12.** Full semantic map for the cognition-perception domain, visualized with modularity analysis* (Blondel et al. 2008) in Gephi

**Figure 13.** Snapshot from CLICS with SEE as a pivot
From CLICS to a more economical map

- Direct edge between perception verbs denoting *non-controlled experience* (e.g., HEAR, SEE) and cognitive verbs (e.g., UNDERSTAND)

**Figure 12.** Full semantic map for the cognition-perception domain, visualized with modularity analysis* (Blondel et al. 2008) in Gephi

**Figure 14.** Snapshot from CLICS with LOOK as a pivot
From CLICS to a more economical map

- Direct edge between perception verbs denoting non-controlled experience (e.g., HEAR, SEE) and cognitive verbs (e.g., UNDERSTAND)
- No intrafield extension between SEE and HEAR, without going through interfield meanings

Figure 12. Full semantic map for the cognition-perception domain, visualized with modularity analysis* (Blondel et al. 2008) in Gephi
Mapping Vanhove’s data

- Visualization of frequency of polysemy patterns

- Implicational hierarchies:
  - If THINK and SEE, then KNOW
  - If HEAR and LEARN, then KNOW

- The map predicts more than the attested data
  - If REMEMBER and SEE, then UNDERSTAND
  - “[A] good model always predicts a few things not yet encountered” (Cysouw 2007: 233)

- HEAR, KNOW, and UNDERSTAND are the most important nodes in the map (articulation points)

Figure 15. Semantic map for the cognition-perception domain based on Vanhove’s (2008) data
Mapping Vanhove’s data

- Visualization of frequency of polysemy patterns

- Implicational hierarchies:
  - If THINK and SEE, then KNOW
  - If HEAR and LEARN, then KNOW

- The map predicts more than the attested data
  - If REMEMBER and SEE, then UNDERSTAND
  - “[A] good model always predicts a few things not yet encountered” (Cysouw 2007: 233)

- HEAR, KNOW, AND UNDERSTAND are the most important nodes in the map (articulation points)

- Again, no intrafield polysemy is allowed without the intervention of an interfield polysemy
  - If SEE and HEAR, then either KNOW or UNDERSTAND

Figure 15. Semantic map for the cognition-perception domain based on Vanhove’s (2008) data
Wordnet

A database of words that are linked together by their semantic relationships

• **N of lgs**: 25

Core concept

Words are grouped together as sets of synonyms (Fellbaum 1998: 72ff.)

**Synset**: A synonym set; a set of words that are roughly synonymous in a given context

A prerequisite for the representation of meanings in a lexical matrix (Miller et al. 1993)
Wordnet

A database of words that are linked together by their semantic relationships

- N of lgs: 25

Core concept

Words are grouped together as sets of synonyms (Fellbaum 1998: 72ff.)

Synset: A synonym set; a set of words that are roughly synonymous in a given context

A prerequisite for the representation of meanings in a lexical matrix (Miller et al. 1993)

Method

1. Choose four basic senses:
   a) SEE, HEAR (non-controlled experience / opportunistic perception verbs)
   b) LOOK, LISTEN (controlled activity / explorative perception verbs)
2. Collect the forms that lexicalize these 4 senses
3. Retrieve the list of all the senses of these forms (the total of the synsets)
Method

4. For each form, check whether the senses collected are among its senses
5. Generate a polysemy matrix

These five steps are implemented in a Python script that uses the Wordnet module of the Natural Language Toolkit (NLTK)

Image 1. A snapshot of Wordnet’s synsets of the verb see in English

<table>
<thead>
<tr>
<th>Language</th>
<th>Form</th>
<th>Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SEE</td>
</tr>
<tr>
<td>English</td>
<td>see</td>
<td>1</td>
</tr>
<tr>
<td>French</td>
<td>regarder</td>
<td>0</td>
</tr>
<tr>
<td>Spanish</td>
<td>mirar</td>
<td>0</td>
</tr>
<tr>
<td>Spanish</td>
<td>observar</td>
<td>0</td>
</tr>
<tr>
<td>Spanish</td>
<td>ver</td>
<td>1</td>
</tr>
</tbody>
</table>
Direct edges between perception verbs denoting *non-controlled experience* (e.g., HEAR) and cognitive verbs (e.g., UNDERSTAND)

**Figure 16.** Semantic map for the perception-cognition domain based on the Wordnet dataset
Discussion
I. Areal patterns in CLICS
II. General discussion
Figure 17. **HEAR/ LISTEN vs. SEE/ LOOK**: A 2D t-SNE projection (van der Maaten & Hinton 2008) of CLICS polysemy data
Discussion: Areal patterns in CLICS

Macro-areas in CLICS

Africa
Australia
Eurasia
North America
Papua
South America

- Coverage of the world’s languages in CLICS is biased towards certain regions of the world (South American languages, languages of the Caucasus, languages of Europe figure particularly prominently).

(List et al. 2014)
Discussion: Areal patterns in CLICS

Figure 18. A 2D t-SNE projection of the polysemy patterns of verbs with meanings HEAR or LISTEN and SEE or LOOK from the CLICS dataset

Verbs with meanings UNDERSTAND
Figure 19. Correlations between different meanings in Eurasia
Figure 20. Correlations between different meanings in Papua
Discussion: Areal patterns in CLICS

Figure 21. Correlations between different meanings in South America
General discussion

- The colexification patterns presented here are typical Greenbergian implicational universals.

- The three samples show some stable patterns
  - The indirect connection between SEE and HEAR that are mediated by cognition verbs
  - The direct connection between perception verbs denoting non-controlled experience and cognitive verbs

- The areal impact is difficult to establish besides some limited cases (cf. SEE)
  - Smaller areas might provide more insightful results (provided that we have an adequate sample).
  - Statistical significance is difficult to reach with the ‘small’ samples at our disposal
General discussion

- A sample of areally related, but genetically diverse languages (with enough languages in each family in order to reach statistical significance) would be the way to go in order to investigate further these questions (i.e., beyond semantic factors).


Acknowledgments

http://web.philo.ulg.ac.be/lediasema/