Lexical Diachronic Semantic Maps

Thanasis Georgakopoulos (athanasios.georgakopoulos@ulg.ac.be)
Stéphane Polis (s.polis@ulg.ac.be)

University of Liège – F.R.S.-FNRS

International Summer School on Typology and Lexicon (TyLex), Moscow, 2-7 September, 2017

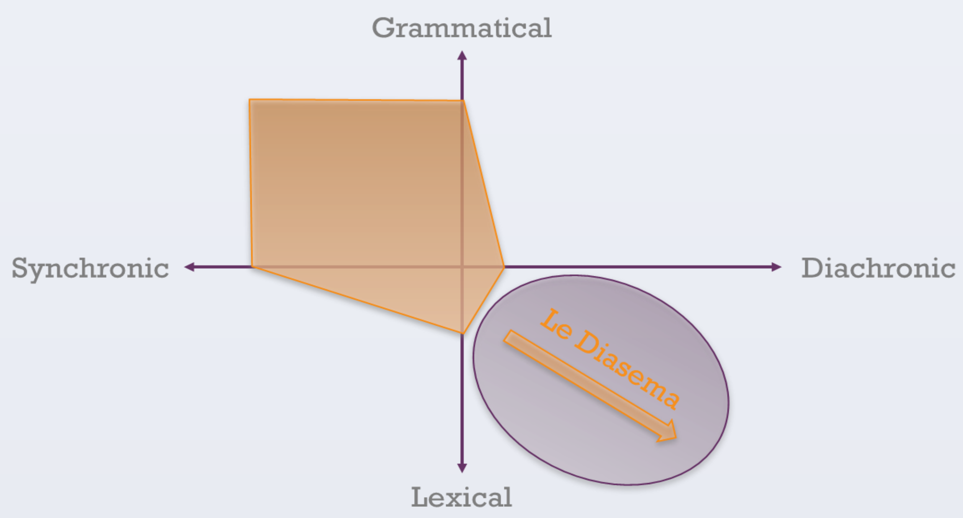
Introduction

The Project: Le Diasema

Main objectives

- To incorporate the **diachronic dimension** into semantic maps of **content words**
- To create an **online platform** for **automatically plotting** diachronic semantic maps based on polysemy data from the languages of the world
- To **extend the method** so as to also include information about the cognitive and cultural factors behind the development of the various meanings

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



What are semantic maps?

'A semantic map is a geometrical representation of functions (…) that are linked by connecting lines and thus constitute a network' (Haspelmath, 2003). It constitutes a 'model of attested variation' (Cysouw, 2007).

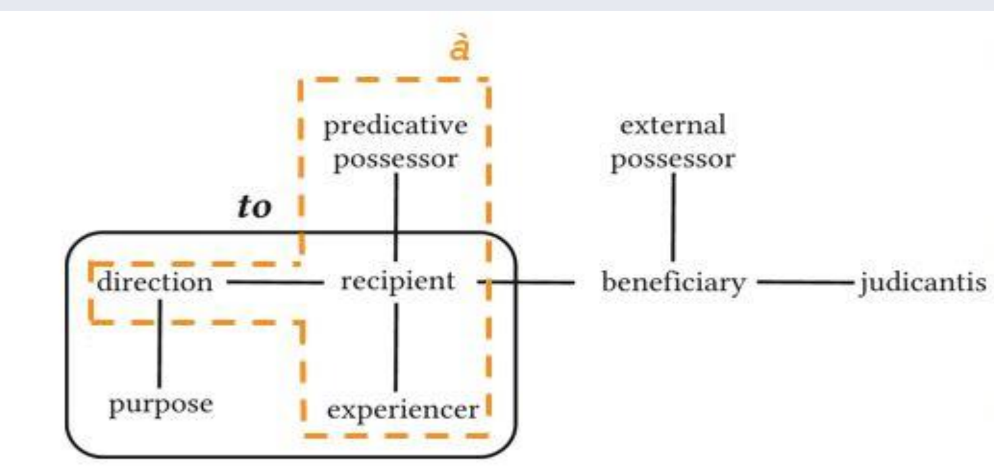


Figure 1. A semantic map of typical dative functions, with the boundaries of English *to* and French *à* (based on Haspelmath, 2003: 213, 215)

- Meaning distinctions are based on **cross-linguistic evidence** and designed to have **cross-linguistic validity**
- They combine the **onomasiological** and the **semasiological** perspective
- **Multifunctionality**. No commitment to a particular claim about conventionalization of senses

Automatic plotting

'[I]deally (…) it should be possible to generate semantic maps automatically on the basis of a given set of data' (Narrog & Ito, 2007: 280)

```
objfn = C(G,T)
while (objfn < 0):
    print ("objective fn is currently", objfn,)
    max_score = 0
    # choose next edge greedily: the one that increases
    for e in PossE:
        # temporarily add e to graph G
        G.add_edge(*e)
        score = C(G,T) - objfn
        G.remove_edge(*e)
        if (score > max_score):
            max_score = score
            max_edge = e
    print ("adding", max_edge, "with score", max_score)
    G.add_edge(*max_edge)
    PossE.remove(max_edge)    # remove max_edge fr PossE
    objfn = C(G,T)
```

Figure 2. Main loop of the inference algorithm

Formally, this is similar to other 'network inference problems,' for instance in biology or epidemiology where one tries to infer networks based on observed features.

Regier et al. (2013) successfully applied the algorithm of Angluin et al. (2010) to linguistic data. They conclude that this algorithm produces equal or better results than the manually plotted maps.

Visualization tools

Visualization techniques and actual semantic analysis will be inseparable in the future of the semantic map model (cf. Malchukov, 2010: 177)

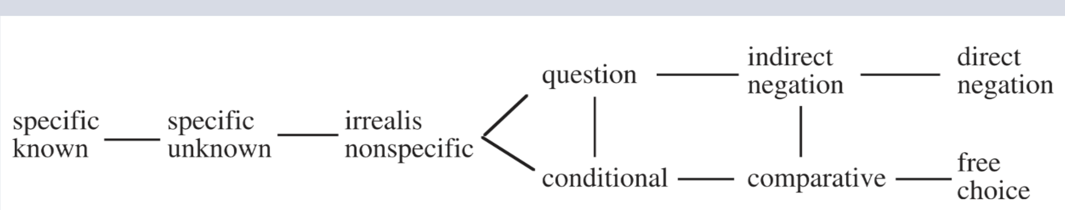


Figure 3. Haspelmath's (1997: 4) original semantic map of the indefinite pronouns functions

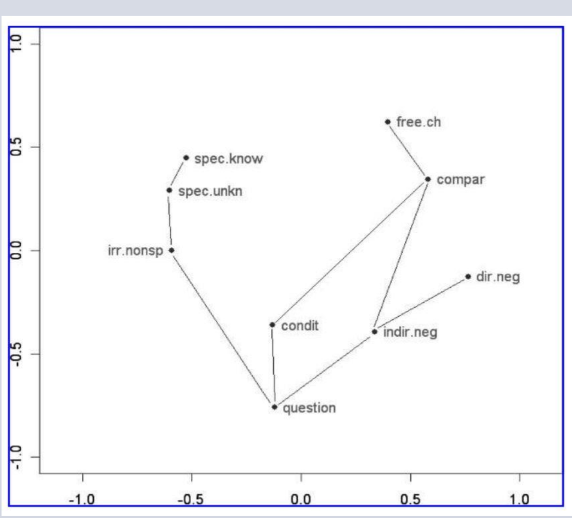


Figure 4. MDS analysis of Haspelmath's (1997) data with the superimposed graph structure (Croft & Poole, 2008: 17, Fig. 6)

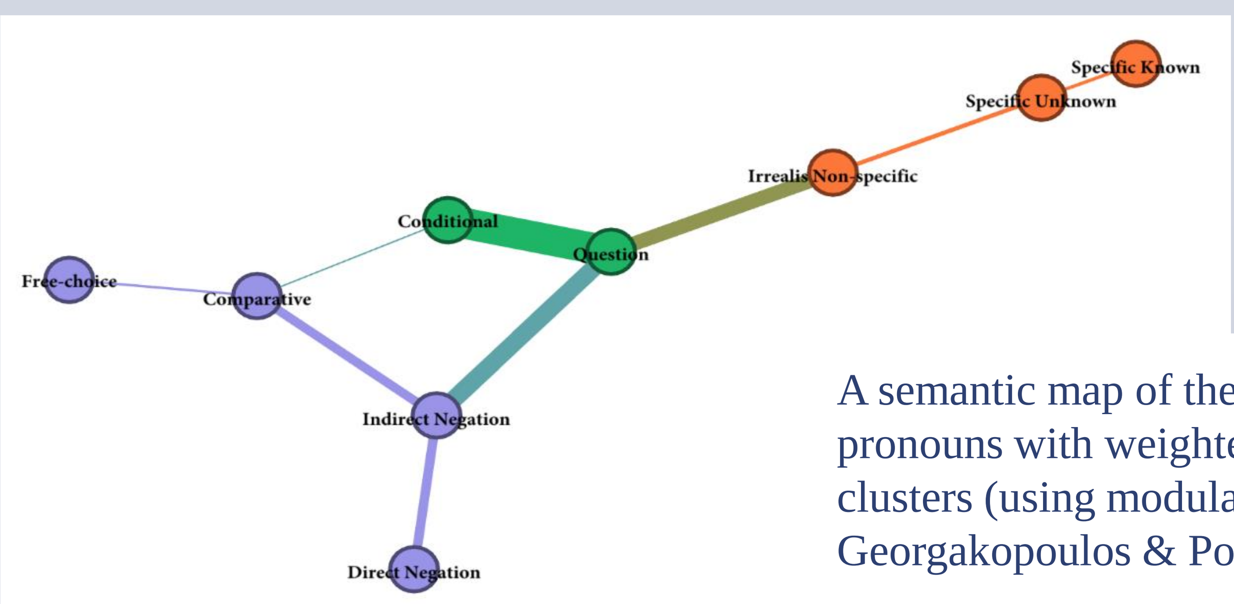


Figure 5. Visualization of Haspelmath's 1997 data in Gephi (https://gephi.org) with the Force Atlas algorithm

A semantic map of the functions of indefinite pronouns with weighted edges and semantic clusters (using modularity analysis); Georgakopoulos & Polis (under review)

Case study: the semantic extension of time-related lexemes in Ancient Egyptian and Ancient Greek

Starting point

1. Choose the concepts: time-related concepts in the Swadesh 200-word list (Swadesh, 1952: 456-457):
A. **DAY/DAYTIME**; B. **NIGHT**; C. **YEAR**

Why?
- Universality
- Stability
- Comparability

Crosslinguistic polysemy patterns

2. Identify in CLICS (List et al., 2014) **the main polysemy patterns** attested for these three meanings
- E.g., **DAY/DAYTIME**: CLOCK/TIMEPIECE, HOUR, SEASON, SUN, TIME, WEATHER

381 polysemy patterns

	A	B	C
119	day	afternoon	hau_std:rana//ket_std:i?//plj_std:piidi//rus_std:den//tli_std:yakyee
120	day	again	kha_std:sngi
121	day	age	gui_std:'ara//yad_std:hnda
122	day	anger	tzz_std:k'ak'al
123	day	bright	tzz_std:k'ak'al
124	day	clock, timepiece	gue_std:wurlngarn//sei_std:šã?

Inferring a weighted semantic map

3. Convert the polysemy patterns into a lexical matrix (Python script)

```
Tmap = [Tsenses]
for t in Tclean:
        split_longWord = t[2].split('//')
        for couple in split_longWord:
            longWord = couple.split(':')
            line = [longWord[0], longWord[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)
```

Python script α

	A	B	C	D	E	F
1			age	acid, sour	city, town	day
2	yad_std	hnda	1	1	0	1
3	vec_std	edat	1	0	0	0
4	jpn_std	toshi	1	0	1	0
5	gui_std	'ara	1	0	0	1

Lexical matrix

4. Plot a weighted semantic map with an adapted version of the algorithm suggested by Regier et al. (2013) that computes weighted edges (Python script β)

```
# CREATE INITIAL GRAPH
# graph G: add each term's nodes, no edges in graph yet.
G = nx.Graph()  # create empty graph (undirected)
PossE = []      # list of possible edges, filled below
for t in T:
    # add all nodes in t, if not already in graph
    for n in t:
        if (not G.has_node(n)):
            G.add_node(n)
    # add to PossE a link between each pair of nodes in t
    # adding a link between every node in G is needless and slower
    for pair in allpairs(t):
        u = pair[0]
        v = pair[1]
        if (not ((u,v) in PossE) or ((v,u) in PossE)):
            PossE.append((u,v))
```

Python script β

5. Visualize the crosslinguistic semantic map of time-related senses

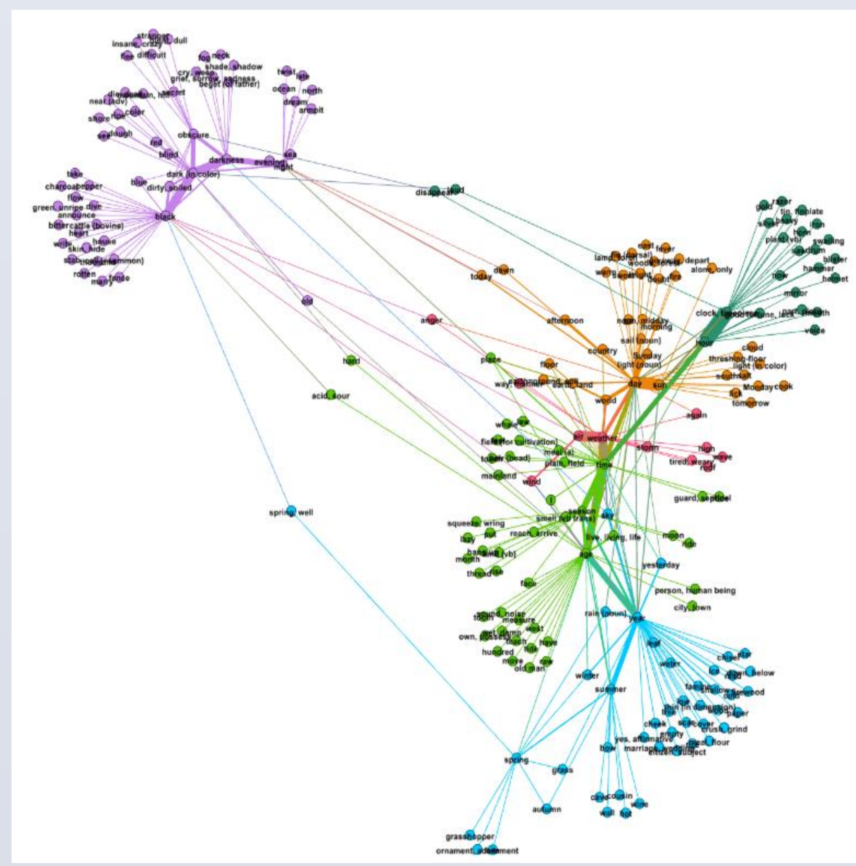
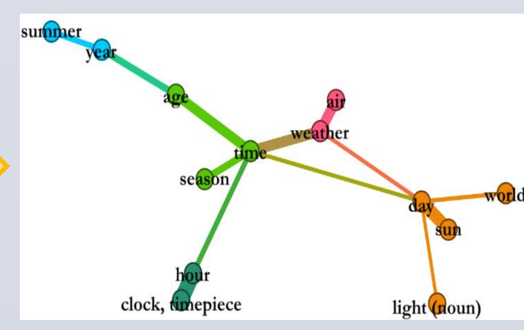


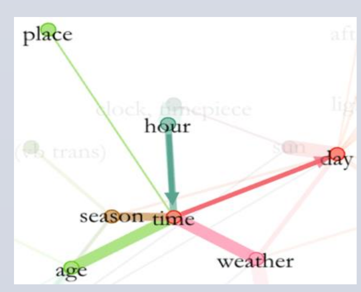
Figure 6. Full semantic map for time-related meanings, visualized with modularity analysis (Blondel et al., 2008) in Gephi

Dynamicizing the map

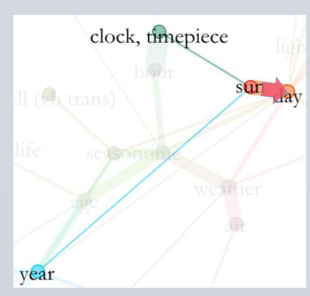
6. For a diachronic investigation
- **Remove** articulation points in the DARK/NIGHT semantic domain
- **Get rid of** poorly attested patterns of polysemy,
- **Keep** meanings that are connected in at least 8 different languages



7. Select languages with diachronic data
- **Ancient Greek** (8th – 1st c. BC)
  Sources: Perseus digital library (http://www.perseus.tufts.edu/hopper/); Cunliffe (A lexicon of the Homeric Dialect), LSJ
- **Ancient Egyptian** (26th c. BC – 10th c. AD)
  Sources: Thesaurus Linguae Aegyptiae (http://aaew.bbaw.de/tla/); The Ramses corpus (http://ramses.ulg.ac.be), Lexical resources (Coptic etymological dictionaries)



8. Add diachronic information based on the attested material



AG: *hēlíos* 'sun' ⇒ 'day'

AEg. *nw* 'hour/moment'

Coptic ⲛⲁⲩ (Crum, 1959)
- hour
- time

+ day(light) [rare]

Selected references

Croft, W., & Poole, K. T. (2008). Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics*, *34*(1), 1–37.
Haspelmath, M. (1997). *Indefinite Pronouns*. Oxford: Oxford University Press.
Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In: M. Tomasello (Ed.), *The new psychology of language*, Vol. 2 (pp. 211–242). Mahwah/ New Jersey: Lawrence Erlbaum Associates.
List, J.-M., Mayer, Th., Terhalle, A., & Urban, M. (2014). *CLICS: Database of Cross-Linguistic Colexifications*. Marburg: Forschungs-zentrum Deutscher Sprachatlas (Version 1.0, online available at http://CLICS.lingpy.org, accessed on 2016-27-10).
Narrog, H., & Ito, S. (2007). Re-constructing semantic maps: the comitative-instrumental area. *Language Typology and Universals*, *60*(4), 273–292.
Regier, T., Khetarpal, N., & Majid, A. (2013). Inferring semantic maps. *Linguistic Typology*, *17*(1), 89–105.
Swadesh, M. (1952). Lexicostatistic Dating of Prehistoric Ethnic Contacts. *Proceedings of the American Philosophical Society*, *96*, 452–463.

LIÈGE université · HR EXCELLENCE IN RESEARCH · BeIPD - COFUND · euraxess RESEARCHERS IN MOTION