**Article**

Thanasis Georgakopoulos, Eitan Grossman, Dmitry Nikolaev*
and Stéphane Polis

# Universal and macro-areal patterns in the lexicon

## A case-study in the perception-cognition domain

https://doi.org/10.1515/lingty-2021-2088
Received October 12, 2018; accepted May 17, 2020; published online September 28, 2021

**Abstract:** This paper investigates universal and areal structures in the lexicon as manifested by colexification patterns in the semantic domains of perception and cognition, based on data from both small and large datasets. Using several methods, including weighted semantic maps, formal concept lattices, correlation analysis, and dimensionality reduction, we identify colexification patterns in the domains in question and evaluate the extent to which these patterns are specific to particular areas. This paper contributes to the methodology of investigating areal patterns in the lexicon, and identifies a number of cross-linguistic regularities and of area-specific properties in the structuring of lexicons.

**Keywords:** areal semantics; coexpression; colefixication networks; dimensionality reduction techniques; formal concept lattices; lexical typology; linguistic universals; perception and cognition; semantic maps

## 1 Introduction

This paper investigates universal and areal structures in the lexicon as manifested by coexpression patterns in the semantic domains of perception and cognition. The present study focuses on two related questions. First, to what extent do

***Corresponding author: Dmitry Nikolaev [dmʲitrʲɪj nʲɪkɐˈlaɪf]**, Stockholm University, Stockholm, Sweden, E-mail: dmitry.nikolaev@ling.su.se
**Thanasis Georgakopoulos [θaˈnasis ʝeorɣaˈkopulos],** National Research University, Higher School of Economics, Moscow, Russia; and Aristotle University of Thessaloniki, Thessaloniki, Greece
**Eitan Grossman [ˈeɪtɐn ˈgʁosmɐn],** Hebrew University of Jerusalem, Jerusalem, Israel
**Stéphane Polis  [steˈfan poˈlis],** F.R.S.-FNRS/University of Liège, Liège, Belgium

bottom-up methods using language samples of different sizes match (or challenge) the results of case-studies conducted by experts on individual languages? Second, to what extent do these methods reveal new universal or area-specific generalizations about the organization of lexicons? In order to operationalize these questions, we use different exploratory strategies (including weighted semantic maps, colexification networks, formal concept lattices, correlation plots, and dimensionality reduction techniques) in order (a) to evaluate the validity and limits of proposed universal generalizations in lexical-typological work in the domains of perception and cognition, (b) to test proposed claims concerning language- or culture-specific associations, and (c) to identify coexpression patterns that have not been discussed in the literature and to analyze their distribution in the world's languages. This study is based on data from three different datasets, i.e. Vanhove's (2008) study of verbs of perception and cognition, the Open Multilingual WordNet (Bond and Paik 2012) and the Database of Cross-Linguistic Colexifications (List et al. 2018a). We chose these datasets because of their accessibility and their broad and diverse cross-linguistic coverage. The use of multiple datasets also allows us to see to what extent results from one sample are replicated or not across other samples.[1]

We focus on the type of coexpression called 'colexification'. This concept has been used in typological studies to refer to "the capacity, for two senses, to be lexified by the same lexeme in synchrony" (François 2008: 171). Consider for example (1)–(2):

(1) Can you *see* the bird in that tree?
(2) I just can't *see* your point.

(Examples from Princeton WordNet of English)

In these two examples, two senses, SEE and UNDERSTAND, are lexified by the same word, namely *see*, in English.[2] A basic assumption of our method is that recurrently colexified meanings[3] are semantically related in some way (Haspelmath 2003: 217; see also Wälchli and Cysouw 2012). In this case, the perception meaning SEE would be somehow linked to the cognition meaning UNDERSTAND. As such, it might be assumed that colexification reflects natural semantic connections in a straightforward way. Indeed, it is often relatively simple to posit semantic links between

[1] All scripts and supplementary data can be found at https://osf.io/2huz6/.

[2] In this paper, we use the following conventions: small capitals for meanings; italics for language-specific forms, e.g., for a verb in a particular language; angle brackets for colexification patterns. We also follow the convention in Conceptual Metaphor Theory of writing conceptual metaphors in small capitals.

[3] The term 'meaning' is used here interchangeably with the term 'sense'.

different senses colexified by one and the same word. However, things are not that simple (Cristofaro 2010). Cross-linguistic distributions – such as a recurrent colexification pattern of the sort described above – are thought to result from two main types of causal factor, which Bickel, in a number of publications, has called 'functional' and 'event-based.' Functional factors are

> grounded in the biological/cognitive or social/communicative conditions of language, such as specific processing preferences […], or specific sociolinguistic constellations […] that systematically bias the way linguistic structures evolve. The defining property of functional triggers is that they affect transition probabilities universally, independent of concrete historical events (Bickel 2017: 42).

In the present context, functional triggers are any universally available semantic (or other) factors that bias the way that lexical items either extend their meanings by developing additional colexified senses or restrict their meanings by losing colexified senses. A possible candidate for such a functional trigger is Traugott and Dasher's (2002) pragmatically-based account of meaning change, and numerous hypotheses about functional triggers can easily be derived from the semantics literature.

Event-based factors, on the other hand, are "tied to single historical events, leading to idiosyncratic, one-off changes" (Bickel 2017: 43) and are often tied to language contact. The crucial point here is that cross-linguistic distributions are potentially always the result of the interaction between functional and event-based triggers. For example, 'have' perfects and relative pronouns, prominent in Europe but rare elsewhere, have been proposed to be mainly the result of event-based factors. And typological studies have shown (Bickel et al. 2014; Sinnemäki 2014) that other grammatical patterns, such as the prevalence of animacy vis-à-vis definiteness and differential argument marking, do not simply follow a single worldwide distribution of occurrence governed by some language-internal factors, but also display clear macro-areal dependencies. Much earlier, Dryer (1989) argued that Greenbergian word order correlations can be understood as resulting purely from functional factors only to the extent that they do not show clear areal signals. Of course, this is not an either-or issue: there may be universal functional factors that make certain colexification patterns inherently more or less likely, while language contact can change the real probabilities of such colexification patterns to develop or be lost.

We would therefore like to test if there is an interaction between universal semantic factors and diffusion inside macro-areas governing the frequency of occurrence of certain colexification patterns. At this point, we do not model this directly in the sense of deriving the frequency of a colexification pattern from its 'semantic naturalness' and macro-area. However, we aim to broadly classify colexifications into universal and areally-restricted, under the assumption that the universal ones provide information about the 'natural' organization of the perception-cognition semantic domain.

In the present context, we assume that verbs with meanings associated with a basic modality of perception have currently-unknown base probabilities of developing colexification patterns including meanings from both within the domain of perception and outside of it. The actual documented colexification patterns are the result of the interaction of functional factors and event-based factors. We will be interested in two main empirical facts: whether the semantic structure of the domains of perception and cognition is similar across macro-areas or not and whether there are correlations, positive or negative, between meanings both at a global scale and at a smaller areal scale.

Following an early idea articulated by Greenberg,[4] we hypothesize that strong global associations between meanings plausibly point to a high base probability of the development or stability of a particular colexification pattern. Similarly, weak global associations point to a lower base probability of a particular colexification pattern. On the other hand, strong areal signals are likely evidence of low base probabilities of spontaneous colexification, i.e., a reduced role for inherent semantics, but a high degree of diffusibility. Similarly, weak areal signals combined with an overall low frequency probably point to a low base probability and a low degree of diffusibility. We stress that these interpretations are very tentative, and may be wrong. First of all, the sample used for testing our hypothesis is not phylogenetically balanced across macro-areas, which is a major weakness of our study. As a result, it may be that strong areal correlations are the result of common inheritance in large phyla dominating their respective macro-areas, which may in turn point to a high degree of stability (Nichols 2003; Wichmann and Holman 2009) of a colexification pattern. Furthermore, strong global correlations might not in fact point to a prominent role of inherent semantic factors; rather, it might be the case that such colexification patterns are so diffusible that they spread over several whole areas, giving the impression of a global preference. It also might be the case that there is a constant base probability of the diffusion of colexification patterns—which is ultimately probably mostly due to calquing—which is itself enhanced or inhibited by the sociolinguistic features of contact situations. Despite these drawbacks, however, these interpretations provide hypotheses that can be explored in future research that directly targets the problems of this study.

[4] "If a particular phenomenon can arise very frequently and is highly stable once it occurs, it should be universal or near universal (…). If it tends to come into existence often and in various ways, but its stability is low, it should be found fairly often but distributed relatively evenly among genetic linguistic stocks. […] If a particular property rarely arises but is highly stable when it arises, it should be fairly frequent on a global scale but be largely confined to a few genetic stocks. If it occurs only rarely and is unstable when it occurs, it should be highly infrequent or non-existent or sporadic in its genealogical and genetic distribution" (Greenberg 1978: 76).

In its search for areal (or universal) colexification patterns, our study continues and extends the work of Gast and Koptjevskaja-Tamm (2018), which identifies colexification patterns that show areal signals in two lexical databases, CLICS and the database of the Automated Similarity Judgment Programme (ASJP). The protocol of Gast and Koptjevskaja-Tamm's study involves three steps: first, it identifies clusters of colexification patterns that show an areal bias (by applying the Join Count test, which is used for determining spatial autocorrelations in the data); second, for the patterns that show positive autocorrelation, it looks for cluster areas that are characterized by a given colexification pattern (using hierarchical cluster analysis); and third, it controls for genealogical relationships (using Bayesian logistic regression).

Similarly to Gast and Koptjevskaja-Tamm (2018), we test the possibility of drawing meaningful generalizations regarding areal patterns of co-expression in the lexicon. However, we do not start with the identification of colexification patterns showing areal biases, but consider a lexical field in its entirety, first showing how this field is structured and unveiling recurrent cross-linguistic colexifications, before turning to patterns that are specific to particular macro-areas and, to a lesser extent, to smaller regions within macro-areas. The final step of this process consists of detecting good candidates for colexifications that result from diffusion events rather than inheritance (cf. Koptjevskaja-Tamm and Liljegren 2017).

The present paper has primarily a methodological focus. Using a variety of tools, it contributes to our knowledge of investigating both cross-linguistic regularities and area-specific properties in the structuring of lexicons. On the one hand, our study makes use of automatically plotted weighted semantic maps for unveiling cross-linguistically recurrent semantic structures in the semantic domains of perception and cognition. On the other hand, it uses additional methods (colexification networks, correlation plots and dimensionality reduction techniques) in order to uncover patterns that are specific to particular macro-areas (and, potentially) to micro-areas. Beyond the methodological issues highlighted here, the findings regarding the areal distributions of colexification patterns in the domains of perception and cognition are an empirical contribution to lexical typology, on the one hand, and to areal typology, on the other.

The paper is structured as follows. In Section 2, we summarize the main findings of previous typological studies on the domains of perception and cognition. In Section 3, we make use of automatically plotted weighted semantic maps for unveiling cross-linguistically recurrent semantic structures in the domains under study based on three different datasets, Vanhove's (2008) dataset (Section 3.1), WordNet (Section 3.2), and the Database of Cross-Linguistic Colexifications (Section 3.3). The semantic structures revealed using the semantic map methodology have the form of typical Greenbergian implicational universals of the type: if a lexical item colexifies meanings A and C, then it should also colexify meaning B. In order to study the impact of areality on meaning associations, in Section 4 we use network comparison, correlational plots, and dimensionality-reduction methods. Section 5 is devoted to a general discussion of the results and to the presentation of the conclusions.

## 2 Perception and cognition in typological studies

We focus on the domains of perception and cognition for several reasons. First, both domains are central to human experience: "Every language has a way of talking about seeing, hearing, smelling, tasting and touching" and "[e]very language has a way of speaking about how one knows" (Aikhenvald and Storch 2013: 1). As a result, meanings belonging to sense perception and cognition appear in the main collections of basic concepts, such as the Swadesh list (Swadesh 1952), the Leipzig-Jakarta meaning list (Haspelmath and Tadmor 2009), and the meaning list of the Intercontinental Dictionary Series (IDS). Consequently, information about lexicalization patterns is readily available for many languages. Second, an extensive literature, which has revealed both universal and culture-specific patterns, reports on meaning extensions within these domains, on the one hand, and on the semantic connections between the domains, on the other (e.g., Evans and Wilkins 2000; Sweetser 1990; Vanhove 2008; Viberg 1983). As such, the results of the present study can be assessed against an existing body of typological evidence.

We take the two main sense modalities, namely sight and hearing, as our starting point. This is not to say that other sense modalities are beyond the scope of our research, as will become clear from Sections 2–3 below, since colexification patterns reveal connections between sight and hearing, on the one hand, and touch, taste, and/or smell, on the other.[5] For each modality, we consider both lexemes expressing controlled activities (e.g., *to look*, *to listen*) and non-controlled experiences (e.g., *to see*, *to hear*), cf. Viberg (1983), Evans and Wilkins (2000) and Wälchli (2016).[6] Here we focus only on verbs and only on cases in which two senses are colexified by the same verb in synchrony, namely on what François (2008: 171)

[5] In the last decade, there has been a growing interest in these three 'lower' sense modalities (for this term see Classen 1997), especially the olfactory modality. This is reflected in a series of publications by Asifa Majid and colleagues (see, e.g., Burenhult and Majid 2011; Majid and Burenhult 2014; Wnuk and Majid 2014).

[6] Note that we do not distinguish between experiencer-based (both activities and experiences are experiencer-based) and source-based verbs (on this distinction see Evans and Wilkins 2000; Vanhove 2008; Viberg 1983: 123–124; see also Section 3.1).

terms ‘strict colexification.’ An example of strict colexification is the case of the form *wum* in the language Dwot (Afro-Asiatic, Africa), which colexifies TASTE and HEAR (see also examples 1–2). Although instances of loose colexification, such as etymologically related forms or derivationally-related forms, can reveal interesting semantic associations, these were not taken into consideration (see Georgakopoulos et al. 2016 for the ramifications of either including or excluding loose colexification from the analysis), mostly because they are difficult to identify automatically in large lexical datasets.

Our choice to investigate sight and hearing is mainly motivated by two facts. First, these two sense modalities seem to be universally more prominent within the domain of sensory modalities (see, e.g., Evans and Wilkins 2000; Levinson and Majid 2014; San Roque et al. 2015; Viberg 1983; Vanhove 2008; Winter et al. 2018). Second, they are cross-linguistically more closely connected to mental perception than smell, taste, and touch are (see, e.g., Evans and Wilkins 2000; Ibarretxe-Antuñano 2008; Sweetser 1990).

The primacy of sight and hearing is represented in the unidirectional hierarchies proposed in Viberg (1983: 136; 2001). Figure 1 presents Viberg’s (2001) lexicalization chain for perception verbs with sight at the top and touch, taste, smell at the bottom. This hierarchy is based on three markedness criteria (see Croft 2003: 91ff): (a) structural coding, which pertains to the number of morphemes the linguistic elements in question have; (b) behavioral potential, which refers to the number of formal distinctions in an inflectional paradigm as well as to the number of syntactic environments in which an element can occur; and (c) textual frequency (how often a token occurs in a given text sample in individual languages) and cross-linguistic frequency (e.g., in how many languages the meaning SEE is lexicalized distinctively from other meanings associated with perception).

sight > hearing > touch/smell/taste

**Figure 1:** The sense-modality hierarchy for perception verbs (Viberg 2001: 1297).

The hierarchy reads as follows: a verb with a prototypical meaning denoting a certain modality, e.g., sight, may extend its meaning to cover lower modalities (to the right) in the hierarchy (e.g., hearing, touch, taste, smell). Viberg (2001: 1297) mentions the example of Djaru (Pama-Nyungan, Australia), in which HEAR is realized using the same root as SEE (i.e., *ɲaŋ-*) with the extension *-an* (‘hear’: *ɲaŋ-an*; ‘see’: *ɲaŋ-*).[7] Given the structural coding criterion, since the number of

[7] This is a case of loose colexification. In CLICS$^2$, SEE and HEAR are strictly colexified only four times, two of which are found in Pama-Nyungan languages.

morphemes in *ɲaŋ-an* is greater than in *ɲaŋ-*, the former is considered more marked than the latter. Since unmarked concepts should appear higher up in the hierarchy, such an example supports the priority of sight against hearing. Numerous studies support Viberg's (1983, 2001) proposal (see, e.g., Evans and Wilkins 2000; Vanhove 2008). There are studies, however, which only partly confirm the proposed hierarchy. For instance, San Roque et al. (2015) corroborated the view that the visual modality dominates (see also Winter et al. 2018), but their data show that the ranking of the other modalities varies in the languages of their sample. As for hearing, its dominance over touch, taste, and smell is a tendency rather than a hard-and-fast rule. Lastly, there are a few counterexamples that challenge the universalist hypothesis that vision is always the dominant modality. Such a counterexample comes from Kolyma Yukaghir (Yukaghir, Eurasia), in which a diachronic extension of an auditory construction to a general meaning of perception that encompasses visual perception is in progress (Maslova 2004; see also Brenzinger and Fehn 2013; Nakagawa 2012).

Beyond intrafield semantic extensions, i.e., extensions that occur within the same semantic domain, we are also interested in transfield semantic extensions, i.e., mappings from one domain (in our case perception) onto another (in our case cognition) (on the distinction between intrafield and transfield changes, see Matisoff 1978: 176–179). Although with respect to intrafield extensions there is a general consensus in the literature concerning the universal prevalence of certain concepts over others (see, e.g., Vanhove 2008), when it comes to transfield extensions, research has led to contradictory results. Sweetser (1990) suggested that cognition is linked first and foremost to VISION. She advocates a general MIND-AS-BODY conceptual metaphor (Sweetser 1990: 28–32), which is motivated by correlations between the bodily external self and the internal self and includes such metaphors as SEEING IS KNOWING and HEARING IS UNDERSTANDING. But Sweetser adds a caveat: "[i]t would be a novelty for a verb meaning 'hear' to develop a usage meaning 'know' rather than 'understand,' whereas such a usage is common for verbs meanings 'see'" (Sweetser 1990: 43). In their study on 69 Australian languages, Evans and Wilkins (2000) challenged this claim, showing that the basic source for cognition in Australian languages is auditory perception rather than visual perception. Vanhove (2008) strengthens the view that the intellectual side of our mental life is more frequently connected to the HEARING sense with data from 25 languages belonging to six language families. Guerrero (2010) further supports the connection of HEARING with cognition in a study in a large sample of languages from the Uto-Aztecan family (see also various articles in Aikhenvald and Storch 2013). What is common in the aforementioned studies is that all mappings discussed include one sensory modality (hearing or sight) and one aspect of cognition (knowledge or understanding). This led Ibarretxe-Antuñano (2013: 324) to propose a more general

COGNITION IS PERCEPTION metaphor (which resembles Sweetser's MIND-AS-BODY metaphor), which manifests itself differently depending on the culture: as COGNITION IS SEEING in English (Indo-European, Eurasia), COGNITION IS HEARING in Warluwarra (Pama-Nyungan, Australia) and Nunggubuyu (Gunwinyguan, Australia), COGNITION IS SMELLING in Jahai (Austroasiatic, Papunesia) (see Caballero and Ibarretxe-Antuñano 2014: 277–278; Evans and Wilkins 2000: 572).

## 3 Lexical semantic maps for the sight-hearing domains

In order to evaluate the generalizations suggested in the literature (Section 2), we compare classical semantic maps covering the domains of sight and hearing, which were plotted based on different datasets. A classical semantic map is a graph consisting of nodes—which stand for meanings—and edges connecting the nodes —which stand for the relationships between the meanings. Such a graph is shown in Figure 2. The nodes on the semantic map can be thought of as elements of a comparative methodology that are used by typologists to formulate cross-linguistic generalizations, and edges are posited based on patterns of co-expression: meanings that are expressed by the same linguistic item should map onto a connected region of the graph. It should be clear, however, that researchers do not agree as to whether semantic maps reflect the global geography of the human mind (compare Croft 2010 and Cristofaro 2010).

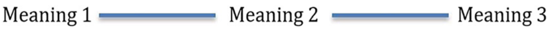

**Figure 2:** Abstract classical semantic map.

Employing semantic maps as a heuristic representational device in typology has a long tradition. The method was initially created in order to describe patterns of polysemy (or 'polyfunctionality') of grammatical morphemes (Croft 2001; Cysouw et al. 2010; Georgakopoulos and Polis 2018; Haspelmath 2003; van der Auwera and Plungian 1998), but recent studies—among others—by François (2008), Perrin (2010), Urban (2012), Wälchli and Cysouw (2012), Rakhilina and Reznikova (2016), Georgakopoulos et al. (2016) and Youn et al. (2016) have shown that it can fruitfully be extended to lexical items.

The three datasets on which the lexical semantic maps of the present paper are based are: a list of crosslinguistic semantic associations in the field of perception (Section 3.1), the Open Multilingual WordNet (Section 3.2), and the Database of

Cross-Linguistic Colexifications (Section 3.3). We used the same methodology for all datasets. First, the data were converted into a lexical matrix (see the example of Table 1) with the meanings on the *X*-axis, the (polysemic) lexical items on the *Y*-axis, and values 1 and 0 indicating whether a lexical item expresses a meaning (1) or not (0).

**Table 1:** Head of the lexical matrix based on Vanhove (2008: 355, 361).

| | | SEE | HEED | UNDERSTAND | KNOW | LEARN | THINK | HEAR | OBEY | REMEMBER |
|---|---|---|---|---|---|---|---|---|---|---|
| English | *see* | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| German | *sehen* | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| French | *voir* | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Italian | *vedere* | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

Second, from these lexical matrices, we infer weighted lexical semantic maps, using an adapted version of the algorithm introduced by Regier et al. (2013).[8] The resulting maps respect the ‘connectivity hypothesis’ (Croft 2001) and the ‘economy principle’ (Georgakopoulos and Polis 2018), and as such generate testable implicational universals. The connectivity hypothesis assumes that “any relevant language-specific and construction-specific category should map onto a CONNECTED REGION in conceptual space” (Croft 2001: 96). This hypothesis determines the relative position of the various meanings on the map: when two meanings are coded by the same linguistic item in a particular language, they should appear connected on the map. The economy principle states that two meanings are connected by an edge if and only if they are not already part of a subgraph of meanings expressed by a single polysemic item in a given language of the sample. In other words, given three meanings ($M_1$, $M_2$, $M_3$), if the items expressing $M_1$ and $M_3$ always express $M_2$, there is no need to draw an edge between $M_1$ and $M_3$ (see Figure 2). The application of the economy principle is what makes the semantic maps interesting: they generate implicational universals that can be tested. Such maps crucially differ from colexification networks (compare the visualization in Figures 12–15) in which no such implicational universals can be inferred from the graph.

Third, we visualize the weighted semantic maps with Gephi[9] (cf. Bastian et al. 2009), a software solution that allows us (1) to filter out rare patterns of co-expression based on the weight of the edges so as to generate stronger hypotheses

[8] Details are provided in Georgakopoulos and Polis (2021).

[9] https://gephi.org.

about the structure of the semantic field of perception and cognition, and (2) to detect groups or communities in the network by using the measure known as modularity. A community is a cluster of nodes with high density of connections (i.e. with many connecting edges) within the community and low density of connections (i.e. with fewer edges) outside the community (Newman 2006). In the context of the visualizations discussed in the present study, a community should be thought of as a cluster of senses that are closely linked and as a whole are only weakly linked to other clusters.

## 3.1 Semantic associations in the semantic field of perception

In order to study the semantic associations between vision, hearing, and prehension (i.e., meanings referring to taking or grasping) on the one hand and mental perception on the other, Vanhove (2008) collected data for 25 languages belonging to eight phyla. For the verbs of perception strictly speaking (i.e., vision and hearing), the dataset consists of 46 lexical items colexifying at least two of the nine following meanings: SEE, HEAR, HEED, UNDERSTAND, KNOW, LEARN, THINK, OBEY, REMEMBER. Note that Vanhove's study does not differentiate between controlled activities (e.g., LISTEN), non-controlled experiences (e.g., HEAR) and experience-based constructions (e.g., SOUND). Consequently, we use the non-controlled experiences SEE and HEAR as cover label.

The map in Figure 3 visualizes both individual colexification patterns and their frequencies, reflected by the relative thickness of the edges connecting the nodes. The use of modularity analysis reveals that there are two communities of senses, as evidenced by the two different colors on the nodes.[10] It shows, as already noted by Vanhove (2008), that from a crosslinguistic point of view the auditory modality prevails in terms of frequency over the visual modality as far as the transfield associations between perception and cognition is concerned ($N_{\text{<SEE,KNOW>}}$: 7; $N_{\text{<HEAR,KNOW>}}$: 10; $N_{\text{<SEE,UNDERSTAND>}}$: 9; $N_{\text{<HEAR,UNDERSTAND>}}$: 16). The most important difference, as clearly indicated by the heavy-weight edge of the <HEAR, UNDERSTAND> colexification, comes from the association of perception verbs with MENTAL MANIPULATION (i.e., with UNDERSTANDING), rather than with KNOWLEDGE (cf. the results from CLICS$^2$ in Sections 3.4 and 4).

[10] On modularity as a measure of the strength of division of a graph into communities, see https://en.wikipedia.org/wiki/Modularity_(networks). Note that the relevant unit for community detection is the node, not the edge. As a consequence, the color of the edges between communities is a blend of the colors of the two connected communities.

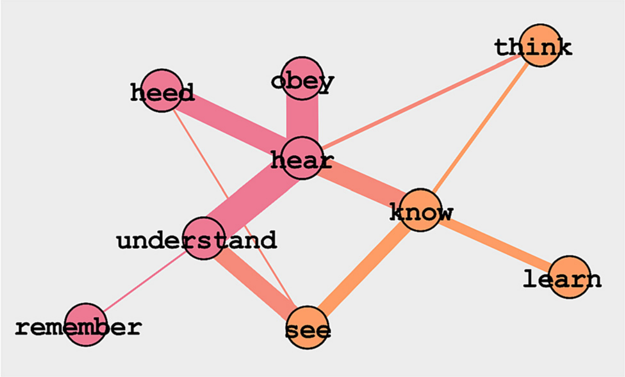

**Figure 3:** Semantic map of the associations between the verbs of seeing, hearing and cognition.

Additionally, Figure 3 illustrates how the economy principle works (see Section 3). Given that KNOW and UNDERSTAND are colexified in eight languages of the dataset, the map could include an edge between these two meanings. However, in all eight languages, the form lexifying the two meanings also lexifies HEAR or SEE. Since <KNOW, UNDERSTAND> is always attested in the presence of HEAR or SEE, and given that a form expressing <HEAR, UNDERSTAND> or <SEE, UNDERSTAND> does not necessarily express KNOW as well, one can remove the edge <KNOW, UNDERSTAND>. The economy principle respects the empirical data, but generates a stronger, hence typologically more interesting, implicational universal.

In this respect, the semantic map in Figure 3 reveals interesting implicational hierarchies. For instance, it tells us that, if the meanings LEARN and HEAR are co-expressed, KNOW is also a meaning of the lexical item (e.g., *sentire* in Italian [Indo-European, Eurasia], *entendre* in French [Indo-European, Eurasia], *hören* in German [Indo-European, Eurasia]), or if THINK and SEE are colexified, then KNOW is also colexified (e.g., *raʔa* in Arabic [Afro-Asiatic, Eurasia]). However, as can happen with models in general (see Cysouw 2007: 233), the map makes a number of predictions about possible patterns that are not attested in the data. For example, it predicts that, if a lexical item colexifies UNDERSTAND and LEARN, this item should also express SEE and KNOW. Such a colexification pattern is however not attested in Vanhove's dataset.

A possible answer to this shortcoming of the classical approach is to systematically map the forms onto the meanings, using formal concept lattices, as introduced by Ryzhova and Obiedkov (2017). In the context of lexical typology, a formal concept lattice can be understood as a set of words, a set of meanings and

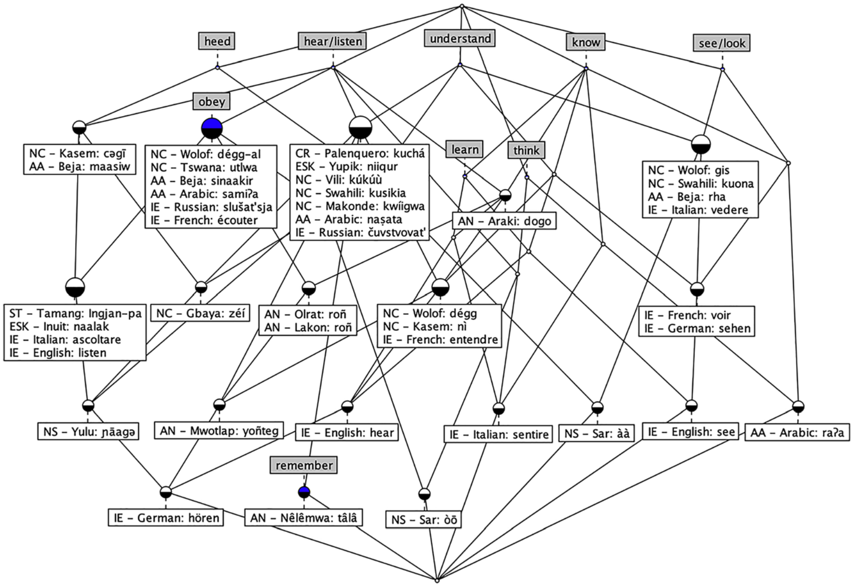

**Figure 4:** Formal concept analysis of Vanhove's (2008) dataset.

binary relations which specify which words have which meanings. Figure 4 shows a formal concept analysis of Vanhove's (2008) dataset.[11]

This kind of lattice[12] fits the underlying data better than the standard graph-based maps, since no information is lost here in the process of building the lattice. In such a lattice, the meanings represented as gray labels are hierarchically structured, and the lexical items (in white boxes) are mapped onto these concepts: one can, for instance, observe that OBEY is hierarchically strictly subordinated to HEAR (since it is lower in the lattice and only linked to HEAR), which means that the five lexical items associated with the node OBEY all also have the meaning HEAR. In terms of visual conventions, a black lower-half means that a lexical item is associated with the node, and a blue upper-half means that this node is a labeled

[11] The representational complexity does not allow for an easy visual exploration of large-scale lexical datasets, which is why we do not repeat the formal concept analysis for the other datasets. Note that the lattice does not include etymological colexifications and morphological derivations mentioned by Vanhove (2008). We must point out that the lattice contains two questionable Russian lexical items. First, the verb *čuvstvovat'* 'to feel' never expresses the meaning 'to understand'. Secondly, the verb *slušat'sja* 'to obey' does not express the meaning 'to listen', although the verb *slušat'* indeed expresses both these meanings.

[12] The lattice is visualized with Concept Explorer (https://sourceforge.net/projects/conexp/).

concept in the original matrix. Finally, the size of the node is proportional to the number of lexicalizations of a particular concept (see ConExp Project 2006).

Especially interesting are the facts that (a) the dependencies between meanings in the dataset are directly visible, e.g., if obey, then hear; if learn, then know; if remember, then both hear and understand; and (b) the meaning combinations attested in this semantic field are explicitly displayed, with the size of the nodes of the lattice reflecting the number of lexical items. For instance, seven verbs (15%) strictly express the meanings hear and understand, and 16 lexical items in total (35%) include these two meanings among their polysemy patterns; the meanings see and understand on the other hand are less colexified: nine polysemic verbs (20%), among which four (9%) have only these two meanings. It is worth noting that, unlike understand, the meaning know is rarely colexified uniquely with hear (one case: *dogo* in Araki [Austronesian, Papunesia]) or see (two cases in Russian [Indo-European, Eurasia] and Yulu [Nilo-Saharan, Africa]). In most cases, if perception verbs express the meaning know, then they also include among their polysemy patterns other meanings such as think, understand, learn, or obey. Finally, the more complex colexification patterns (i.e., patterns involving three or more meanings, at the bottom of the lattice) appear to be limited to no more than one language in the sample. This is illustrated, for example, by German *hören*, English *see*, and Sar (Nilo-Saharan, Africa) *áá*, which display patterns of polysemy that are not attested in other languages.

Finally, Figure 3 suggests a generalization that has not been properly acknowledged in the literature so far, namely, the fact that the intrafield connection between verbs of vision and hearing—see for instance Viberg's hierarchy discussed in Section 2—is mediated by interfield meanings, specifically know and understand (for a similar observation, cf. San Roque et al. 2018: 397–398). In other words, the two sensory modalities are connected only through mental perception. This point is discussed further based on the data of the Open Multilingual WordNet in the next section (Section 3.2).

## 3.2 Open Multilingual WordNet

The Open Multilingual WordNet (henceforth OMW; Bond and Paik 2012) gives access to WordNets in 29 language varieties,[13] all of which are linked to the original Princeton

[13] The coverage of individual WordNets is fairly limited for some languages (http://compling.hss.ntu.edu.sg/omw/). The language varieties in the Open Multilingual WordNet are: Albanian, Arabic, Basque, Bulgarian, Catalan, Chinese (Mandarin), Chinese (Taiwan), Croatian, Danish, Dutch, English, Finnish, French, Galician, Greek, Hebrew, Indonesian, Italian, Japanese, Malay, Norwegian, Norwegian Bokmål, Persian, Polish, Portuguese, Slovenian, Spanish, Swedish, Thai. In effect, this means that this dataset is basically a sample of languages from the Eurasian landmass and some surrounding islands.

WordNet of English (PWN). The basic units of WordNet are sets of cognitive synonyms (called synsets), each expressing a distinct sense. For example, the English verbs *understand*, *realize*, and *see* belong to a synset—called understand.v.02 (meaning that this is the second sense associated primarily with the verb *understand* in English)—which is defined as 'perceive (an idea or situation)'. Synsets are interlinked through lexical relations (any word can occur in several synsets) and conceptual/semantic relations (hyperonymy, hyponymy, meronymy, etc.). In the context of this paper, we use lexical relations in order to posit relationships between meanings. The verb *see*, for instance, is part of another synset—called witness.v.02—defined as 'to perceive or be contemporaneous with' that further includes the verbs *witness* and *find*. Based on the occurrence of *see* in both understand.v.02 and witness.v.02, one can posit a semantic relationship between the two senses, i.e., UNDERSTAND and WITNESS. A caveat is due at this point. The synsets of the OMW are based on the original English WordNet. As such, the meanings are heavily Anglocentric, i.e., very rich where English lexicalizes fine-grained distinctions (consider for instance the cluster around SEE in Figure 5) and rather poor in other cases (see the cluster around HEAR). In addition, the languages included in the OMW are mostly Eurasian and socio-politically dominant modern languages (see fn. 13). Despite these shortcomings, we used the OMW dataset for unveiling cross-linguistically polysemy structures, because *(a)* the language sample is different from the other samples in the current study and *(b)* its rich inventory of meanings allows for certain patterns to emerge.

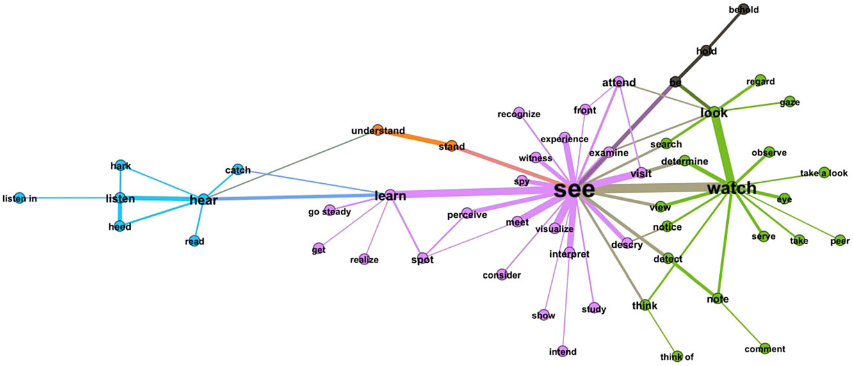

Figure 5: Semantic map for the main polysemy patterns of SEE, LOOK, HEAR, and LISTEN in OMW.

As the Open Multilingual WordNet can be queried using the WordNet corpus reader of the Natural Language ToolKit[14] (NLTK), we use Python to generate a

[14] http://www.nltk.org.

lexical matrix akin to Table 1. The general procedure is as follows. Four basic senses for controlled perception (LOOK and LISTEN) and non-controlled experience (SEE and HEAR) are taken as point of departure (namely, look.v.01, listen.v.01, see.v.01, hear.v.01). For the 29 language varieties of the Open Multilingual WordNet, we collect all the verbs expressing these four meanings, which amounts to 231 verbal lexemes. We then look for all the synsets in which these lexemes occur ($N = 431$ synsets). Finally, since sense distinctions are very fine-grained in WordNets, we merged the synsets that are glossed by the same English verb under a single meaning before generating the lexical matrix, ending up with 274 different meanings. For instance, the synsets hear.v.01 ('perceive (sound) via the auditory sense'), hear.v.02 ('examine or hear (evidence or a case) by judicial process', e.g., 'the jury had heard all the evidence'), and hear.v.03 ('receive a communication from someone', e.g., 'we heard nothing from our son for five years') are merged under a single meaning HEAR.

The original graph inferred from the lexical matrix comprises 274 nodes (i.e., the 274 meanings) connected by 322 edges. We filter out the edges that are required less than five times by the sample, and end up with a graph of 54 nodes connected by 70 edges[15] (Figure 5). This is done in order to build a stronger model; in essence, it is needed in order to identify generalizations without weakening them by rare counterexamples that might result from homonymy or similar phenomena. Similarly to what we observed in Vanhove's (2008) dataset in Section 3.1, the vision senses, namely, SEE/LOOK are not directly connected to the hearing senses, namely HEAR/LISTEN. Again, the field of mental perception mediates between vision and hearing. Crucially, unlike Vanhove's (2008) dataset, WordNet distinguishes between controlled activities (LOOK, LISTEN) and non-controlled experience (SEE, HEAR), and the semantic map of Figure 5 reveals one highly interesting finding: controlled activity senses in both visual and auditory modalities are not directly linked to cognition. Rather, these need to first visit nodes that contain senses of non-controlled experience, namely SEE and HEAR, respectively. This appears to suggest that cognition is pervasively conceptualized as a non-controlled experience in the languages of the sample, and perhaps beyond. In the next section, we turn to a conceptually and cross-linguistically more balanced dataset, CLICS$^2$, which supports this observation with additional evidence.

[15] This model accounts for 70% of the complexity of the original datasets (i.e., 1,055 out of 1,516 colexifications).

## 3.3 Database of cross-linguistic colexifications (CLICS$^2$)

CLICS$^2$ is an online database of synchronic lexical associations.[16] It provides information on about 2638[17] distinct colexification patterns that cover 2487 different concepts (called 'Concepticon concept sets') based on 15 lexical datasets in (currently) 1220 language varieties (List et al. 2018a). CLICS$^2$ allows one to explore the different colexifications via its web-based interface. These cross-linguistic colexification data are represented in the form of networks with weighted edges reflecting the different frequencies of individual colexifications. Also interesting for the purposes of the present paper is the fact that it enables users to look for areal patterns. As such, CLICS$^2$ has great potential for being a powerful tool for studies in lexical typology (cf. Gast and Koptjevskaja-Tamm 2018: 77).

In order to build a semantic map based on the colexification patterns of CLICS$^2$, we followed Robert Forkel's cookbook[18] for CLICS$^2$, and extracted in CSV format all the meanings that are attested for lexemes that express at least one of the four concepts SEE, LOOK, HEAR and LISTEN: 4,045 different word forms lexicalize one (or more) of these four concepts and co-express 362 meanings in total. Among the 4,045 word forms, 819 colexify at least two meanings and can be used for inferring a semantic map.

The full semantic map consists of 362 nodes (i.e., the meanings) connected by 433 edges. Having filtered out the nodes that are supported by only one (305 cases) or two (53 cases) colexification patterns in the whole sample, we end up with the graph of Figure 6.[19] Again, this graph respects the economy principle. According to the dataset, the map in Figure 6 could include an edge between LOOK and UNDERSTAND. This pattern is indeed attested in three language varieties, in Siona (Tucanoan, South America), Manchu (Tungusic, Eurasia), and Kaingáng (Nuclear-Macro-Je, South America). However, in all three languages, the form used to lexify

[16] https://clics.clld.org.

[17] This number refers to colexifications that occur in at least three different language families (List et al. 2018b: 288). Since the submission of the present paper, a new version covering 3,156 language varieties, CLICS$^3$, has been published online.

[18] Released as part of the CLICS$^2$ repository: https://github.com/clics/clics2/releases/tag/v1.1.1; cf. List et al. (2018d).

[19] This model accounts for 67% of the complexity of the original datasets (i.e., 817 out of 1,228 links). For the sake of clarity, we have taken out the meanings that do not belong to the ontological category 'action/process' according to the Concepticon (List et al. 2018c; https://concepticon.clld.org), some of which appear semantically motivated while others are likely to be homonyms or cases of rare pathways of semantic shift: <HEAR, BEAUTIFUL>, <HEAR, EAR>, <HEAR, FULL>, <HEAR, MOSQUITO>, <HEAR, SHADE>; <LISTEN, BLUE>, <LISTEN, HAY>, <LISTEN, PAIN>, <LISTEN, SICK>; <LOOK, EYE>, <LOOK, DAY (24 HOURS)>, <LOOK, FROST>, <LOOK, HORSE>; <SEE, SMALL BOWL>, <SEE, HIGH>, <SEE, STRAW SANDAL>.

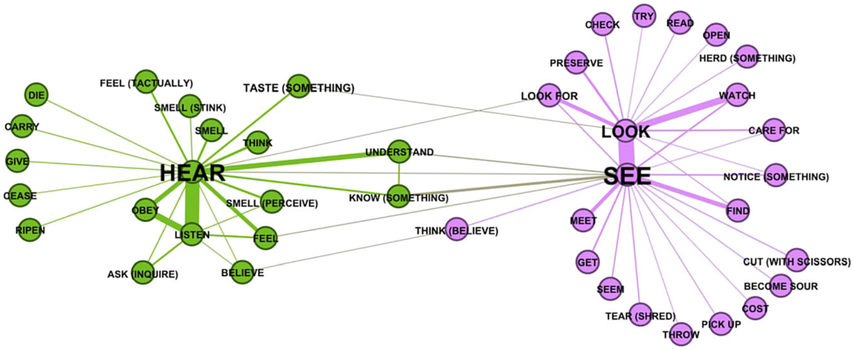

**Figure 6:** Semantic map for the main colexification patterns of SEE, LOOK, HEAR, and LISTEN in CLICS$^2$.

UNDERSTAND and LOOK is also used to lexify SEE. Since <LOOK, UNDERSTAND> is always attested in the presence of SEE and given that a form expressing <SEE, UNDERSTAND> does not necessarily express LOOK as well, the edge <LOOK, UNDERSTAND> is not added to the graph.

Interesting conclusions can be drawn from the map in Figure 6, some of which converge with the information obtained from the other two datasets and some of which are new. To start with, one can observe that cognition senses (THINK, KNOW, UNDERSTAND) again mediate between the domains of VISION and HEARING. Two exceptions are however noteworthy. First, there are a few languages that colexify HEAR and SEE without the support of a cognition sense, as illustrated by their direct connection in Figure 6. For example, this is the case of the verb *nyajil* in the Kuku-Yalanji (Pama-Nyungan Australia). In fact, SEE and HEAR are colexified four times, two of which are found in Pama-Nyungan languages. However, note that for all four varieties in which this colexification is attested, the dataset lacks information about core concepts, such as UNDERSTAND, FEEL, and TASTE. This obviously has an impact on the resulting edge that connects directly SEE and HEAR, which might be an artefact of the lack of some concepts in the dataset for many languages (for the issue of coverage, see the discussion in Section 4). The second exception is a case in which a perception sense, i.e., TASTE, can mediate between HEAR and LOOK, which belong to the domain of perception as well; this connection is limited to some languages and is discussed in the context of areal patterns in Section 4.2 (Africa). The analysis of this dataset further supports the main finding of Section 3.2: non-controlled experiences (SEE and HEAR) are linked directly to cognition, while controlled activities (LOOK and LISTEN) are not.

Additionally, the different weights of certain edges suggest again that knowledge is more frequently linked to sight, whereas mental manipulation (i.e., understanding) is more closely linked to hearing. The <SEE, KNOW> colexification is attested in more languages than <HEAR, KNOW> ($N_{<\text{SEE,KNOW}>}$: 17; $N_{<\text{HEAR,KNOW}>}$: 11), whereas <HEAR, UNDERSTAND> is more robust across languages than <SEE, UNDERSTAND> ($N_{<\text{HEAR,UNDERSTAND}>}$: 43; $N_{<\text{SEE,UNDERSTAND}>}$: 6). However, in terms of modularity, both cognition meanings are more tightly associated with the HEAR cluster (in green) than with the SEE cluster (in purple).

Finally, the map in Figure 6, when approached from the point of view of the colexification patterns in the SIGHT and HEARING domains, clearly shows that the meanings belonging to the other sensory modalities (such as FEEL (TACTUALLY), SMELL, TASTE) form a group with HEAR rather than with SEE. The other two datasets did not provide any insights on this issue.

## 3.4 From sight and hearing to perception and cognition

The previous sections took four basic meanings (SEE, LOOK, HEAR, and LISTEN) as the point of departure, primarily in order to investigate the semantic affinities of the sensory modalities involving SIGHT and HEARING and their relationships to verbs of cognition. Here, we extend the scope of our investigation to the broader domain of perception and cognition so as to produce a structured representation of the semantic domain of perception and cognition as a whole.

Crossing the concepts that are flagged as action/process in the Concepticon (List et al. 2018c)[20] within the semantic fields of sense perception and cognition with the ones that appear on the map of Figure 6 (for the sake of comparability), we selected 22 central concepts belonging to this semantic field: BELIEVE, FEEL, FIND, GET, HEAR, KNOW (SOMEBODY), KNOW (SOMETHING), KNOW OR BE ABLE, LEARN, LISTEN, LOOK, LOOK FOR, MEET, OBEY, READ, SEE, SEEM, SMELL (PERCEIVE), TASTE (SOMETHING), THINK (BELIEVE), THINK (REFLECT), UNDERSTAND.[21] We then extracted from CLICS$^2$ all the verbs that colexify at least two meanings from this set of meanings. This gave us 962 colexification patterns, with 873 unique forms (89 forms being shared between language varieties of the dataset). These colexification patterns were turned into a binarized matrix from which we inferred the semantic map in Figure 7.

This semantic map accounts for 92% of the colexification patterns found in the binarized matrix (1,127 out of 1,227), with edges of weight four and less having been removed.[22] The size of the labels for the nodes is based on betweenness centrality

[20] https://concepticon.clld.org/parameters.

[21] This approach led to the exclusion of meanings such as PINCH, REMEMBER, SNIFF, or TOUCH for instance.

[22] For the semantic maps based on the CLICS (Sections 3.4 and 4), we determined a goodness of fit of at least 90%.

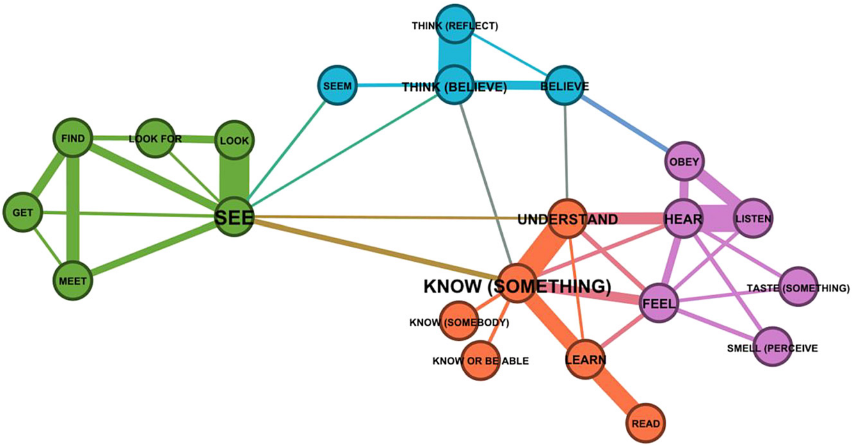

**Figure 7:** Semantic map for the colexification patterns in the domains of perception and cognition based on CLICS$^2$.

and the colors refer to modularity classes computed automatically with the algorithm of Blondel et al. (2008).[23]

The modularity analysis identifies four communities, respectively around SEE, KNOW, HEAR, and THINK/BELIEVE. It is striking that the SEE and the HEAR communities are totally disconnected (the threshold of four cases of colexification resulted in the removal of the rare edge that connected SEE with HEAR), which confirms the results of Section 3.2 and partly those of Section 3.3 with a wider array of meanings taken as a point of departure. The two sensory modalities are only linked via the nodes belonging to two clusters that contain senses belonging to the cognition domain: the KNOW/UNDERSTAND cluster, on the one hand, and the THINK/BELIEVE cluster, on the other.

As far as intrafield associations are concerned, the semantic map of Figure 7 converges with the map in Figure 6: senses belonging to other sensory modalities, i.e., TASTE and SMELL are grouped with HEAR. The fact that all the sensory modalities belong to a single cluster[24] to the exclusion of meanings referring to sight reinforces the finding that perception appears not to be a unitary domain, as generally assumed in linguistics. As far as the cross-linguistic organization of the domain is concerned, sight stands on its own (rather than on the top of a hierarchy of senses; cf. the hierarchy in Viberg 2001), while all the other senses cluster together.

[23] With standard parameters: randomize: ON, edge weights: ON, resolution: 1.0.

[24] Note that TOUCH would belong to the same cluster based on its strong association with FEEL (25) and TASTE (14) in CLICS$^2$.

The behavior of FEEL is also worth commenting on. This concept covers different types of perception (ranging from emotional sensation and particular state of mind to examination by touching). As such, it behaves as a kind of hypernym, and it is unsurprising that it is connected to four meanings in the domain of perception (HEAR, LISTEN, SMELL, TASTE) and three in the domain of cognition (KNOW, LEARN, UNDERSTAND), mediating between the two domains.

One of the advantages of a model of linguistic variation that is visualized in the form of a map such the one in Figure 7 is that it generates testable predictions that can be refuted (or supported) by additional data. Conversely, as already noted in Section 3.1, a disadvantage of such a model is that it over-generates possible constellations of meanings and does not distinguish between (a) patterns that are actually attested—it is for instance predicted that, if in a language a form colexifies LOOK and UNDERSTAND, the form will also lexify SEE, and this prediction is borne out: in three language varieties, in Siona, Manchu, and Kaingáng; (b) patterns that are possible but unattested so far—the prediction that if <FEEL, SEE>, then <FEEL, SEE, KNOW> is not supported by the dataset; and (c) very unlikely patterns—a single form colexifying all the meanings.[25] On average, individual forms of the sample express 2.16 meanings, which means that cases of lexical items with three meanings are infrequent and words with four meanings or more are very rare in the CLICS$^2$ dataset as shown by Figure 8.

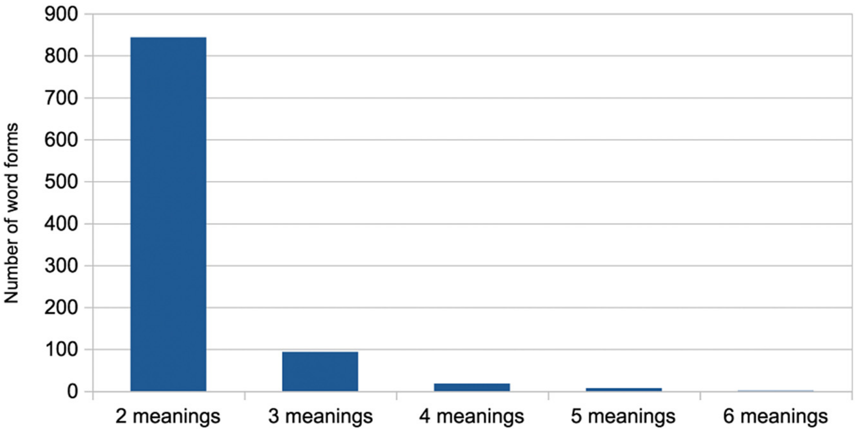

**Figure 8:** Number of meanings per word-forms in the dataset.

We also ask whether verbs expressing meanings from this domain form well-defined groups and whether the structure of these groups recurs across languages.

[25] For further methodological considerations regarding the inclusion of frequency data in semantic maps, see Cysouw (2007: 3.3).

In other words, we try to see how we can group verbs from different languages based on which meanings from the domain of interest they express.

Mathematically, we can think about verbs as inhabiting a 22-dimensional semantic space:[26] they are characterized by 22-place binary vectors (0, 1, 0, … 1) where each coordinate signifies if a verb expresses a particular meaning. Verbs with similar sets of meanings are situated closer in this space and form natural groups. Humans are unable to think efficiently about high-dimensional spaces, and many methods have been devised to make such datasets more manageable by reducing the number of coordinates down to two or three and making the groups visible to the eye. After applying such procedures, it will be possible to see if verbs form similar or different spatial structures when semantic spaces of different macro-areas are compared. We chose UMAP (McInnes and Healy 2018), a very efficient dimensionality-reduction method, to elucidate the structure of verbal semantic space in different macro-areas.[27]

The results of the application of UMAP to the whole dataset are presented in Figure 9. Each dot corresponds to a particular verb in a particular language, and verbs cluster on the 2-D surface according to their colexification profiles.

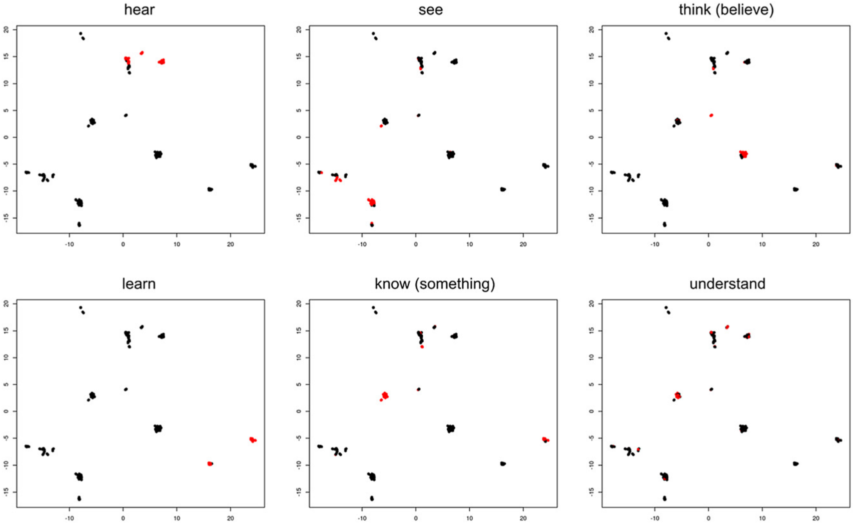

**Figure 9:** UMAP plots reflecting semantic clustering of perception/cognition lexical items in the world's languages. Verbs co-expressing a particular meaning (reflected in the subplot's title) are shown in red.

[26] More precisely, a 22-dimensional cube with 4,194,304 corners.

[27] We used the implementations of UMAP provided in R package uwot (Melville 2018) and used distance matrices computed by the standard R function dist using the Manhattan distance metric.

We repeat the same plot, each time highlighting verbs that have a particular meaning component. The plots indicate that perception/cognition verbal lexicons in the world's languages are structured around several pivotal meanings: verbs almost never colexify HEAR, SEE, THINK (BELIEVE), and LEARN (split into two parts based on other meanings) with each other, and these four meanings together carve up the lion's share of the lexical items. On the other hand, the meanings UNDERSTAND and KNOW (SOMETHING) form a separate cluster only when colexified together. Separately they are predominantly colexified with one of the more 'basic' meanings and may be seen as serving as a kind of glue keeping this chunk of the lexicon semantically connected (cf. their positions in the semantic maps above).

# 4 Macro-areal patterns in the domains of perception and cognition

In this section, we turn to network comparison, correlational plots, and dimensionality-reduction methods for studying the impact of macro-areality on the types of cross-linguistic meaning associations identified above (Section 3.4). The macro-areas under consideration are those commonly used in typology for purposes of creating balanced samples, namely, Africa, Australia, Eurasia, North America, Papunesia, and South America. While these areas might be (or seem to be) too large to show revealing signals, we follow Dryer (1989), Nichols (1992), and Bickel (2020) in assuming that contact-related patterns can scale up to continent-sized areas. More data would allow us to pursue more fine-grained studies of smaller areas with more detailed information on language contact (Muysken 2008). Since the dataset of Vanhove (2008) is too small to conduct a macro-areal investigation (Section 3.1), and given the bias in geographical coverage and typological diversity of the Open Multilingual WordNet (Section 3.2), we use only the data from CLICS$^2$ (Section 3.3–4) for studying areal effects.

As noted by List et al. (2018b: 298–300), however, the data collection in CLICS$^2$ is also unbalanced: for the 22 concepts investigated here, the Average Mutual Coverage (AMC), which is defined as "the average number of concepts shared between all pairs of languages in a given wordlist divided by the number of concepts in total", is 0.278. Furthermore, the areal distribution of the 962 colexification patterns in the domain of perception and cognition, as summarized in Figure 10, shows that the large macro-area of Eurasia is massively over-represented, while Australia and North America together amount to less than 5% of the colexification patterns.

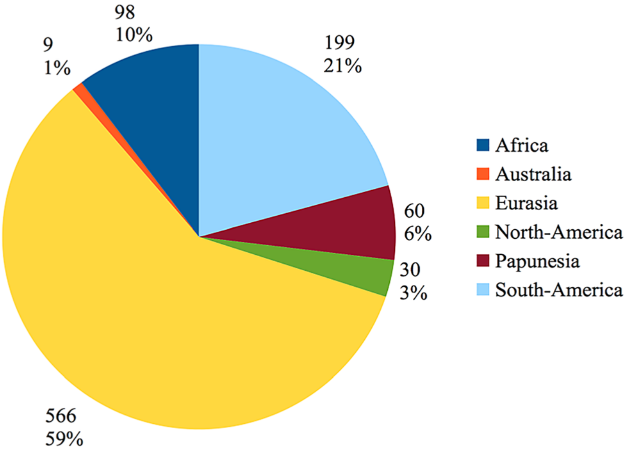

**Figure 10:** Distribution of the data per macro-area in the domains of perception and cognition.

This appears to be the result of two different factors. Australia, as a macro-area, has a very low rate of colexification overall, as shown in Figure 11, which displays the ratio between the number of words that co-express two or more meanings and the number of words expressing at least one meaning in this semantic domain. North America, on the other hand, has a high rate of colexification in this domain, but the language sample is not large enough for a meaningful macro-areal analysis. As such, both Australia and North America have been excluded from the investigations of the areal patterns below.

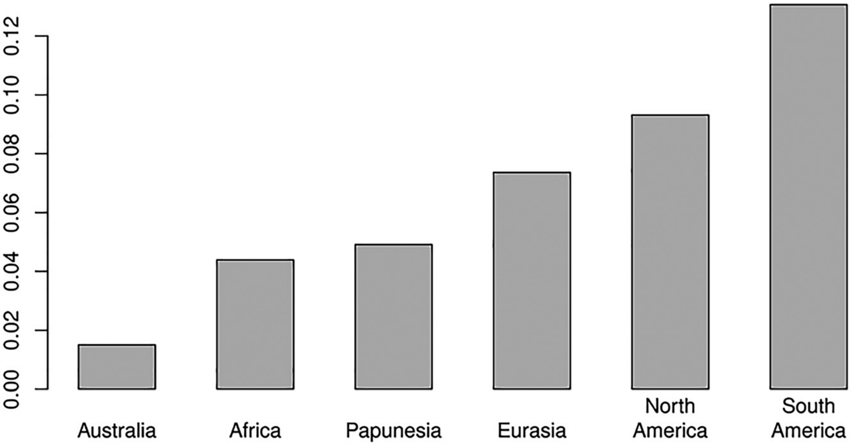

**Figure 11:** Rate of colexification per macro-area in the domains of perception and cognition.

Among the four remaining areas, for which the AMC is slightly better (with a score of 0.328), Africa and Papunesia have a rather low rate of colexification (*c*. one out of 20 verbs), Eurasia stands in the middle, while South America is a heavy colexifier (with *c*. one word-form out of seven colexifying at least two of the 22 meanings. As a matter of fact, South America is also the area where most of the complex colexification patterns for the domain of perception and cognition are found: 69% of the lexical items expressing four meanings or more are from this area.

Table 2 provides details about the number of families, language varieties and colexification patterns in each of these four areas. Individual languages contribute only to a limited extent to the number of colexification patterns available in a given macro-area: between one and three colexification patterns by language on average. Table 2 further lists meanings for the colexification of which information is lacking in CLICS$^2$.

**Table 2:** Distribution of the data per macro-area in the domains of perception and cognition.

| Macro-area | Number of families | Number of language varieties | Number of colexification patterns | Number of meanings colexified | No information about colexification patterns for the meanings |
|---|---|---|---|---|---|
| Africa | 6 | 71 | 98 | 19 | KNOW (SOMEBODY), KNOW OR BE ABLE, SEEM |
| Eurasia | 23 | 273 | 566 | 22 | – |
| Papunesia | 3 | 40 | 60 | 17 | KNOW (SOMEBODY), KNOW OR BE ABLE, LEARN, READ, SEEM |
| South America | 36 | 75 | 199 | 20 | KNOW (SOMEBODY), KNOW OR BE ABLE |

In order to analyze area-specific patterns of colexification, we provide for the four macro-areas:

1. *A full colexification network*, with edges between every single pair of meanings that is colexified by at least one lexical item in the area. The goal of this network is to display the full complexity of the dataset for each macro-area; the size of the labels is based on betweenness centrality.
2. *A correlation plot*, created by representing Pearson correlation coefficients for all pairs of columns in the lexical matrix. Bluer points indicate positive correlations, and conversely, redder points indicate negative correlations. The size

of the points, together with color intensity, indicate the strength of the association. Insignificant correlations (*p*-value > 0.05) were crossed out. Some meanings are not represented in the lexical data from some of the regions, and the respective rows/columns are filled with '?'s. It is important to stress that correlation plots are an *exploratory* method and do not represent an attempt at rigorous hypothesis testing. They show, in a slightly more robust manner, the relative strength of different colexification patterns. Therefore we did not attempt to correct for multiple testing: given the sparsity of the data, most correlations underlying the plots would be rendered insignificant, which would reduce the exploratory value of the plots to nearly zero. Conversely, this correction would not have made our analysis more robust given the uncertain nature of the original sample.

3. *A semantic map*, inferred based on the principles described in Section 3 and for which we fixed a goodness of fit of at least 90% (cf. fn. 22). The semantic maps provide a qualitative look into the data and complement the quantitative approach of the other methods.
4. *A 2-D visualization of the verbs* based on the meanings that these verbs colexify (cf. Figure 8, above).

## 4.1 Eurasia

For the Eurasian macro-area, we have information for all 22 meanings considered in this section. They are connected by 87 edges in the full colexification network (Figure 12a) and by 68 edges in the semantic map (Figure 12c), for which we kept edges with weight four or more, resulting in a goodness of fit of 90% for this model (accounting for 582 out of 648 colexification patterns). As expected, given the higher proportion of colexification patterns from this area, which covers almost the 60% of the observed patterns, the Eurasian map is close to the global map in Figure 7.

As such, four main clusters, similar to those found in the global semantic maps, are identified here (Figure 12c): SEE, KNOW, THINK/BELIEVE, and HEAR. However, these clusters, while showing a significant degree of internal connectedness (as evidenced by the numerous edges among the meanings of individual clusters), are more weakly interconnected and, hence, more independent than in the global map, except for the domains of cognition and hearing, which remain strongly interlinked.

In the domain of cognition, there is a central chain of meanings that are positively correlated with one another: <READ, LEARN>, <LEARN, KNOW>, <KNOW, UNDERSTAND>. KNOW is further significantly connected to FEEL, which, together with UNDERSTAND, links the domain of COGNITION and the domain of HEARING. KNOW is connected to SEE (cf. the general map) and UNDERSTAND is heavily colexified with HEAR, as was also

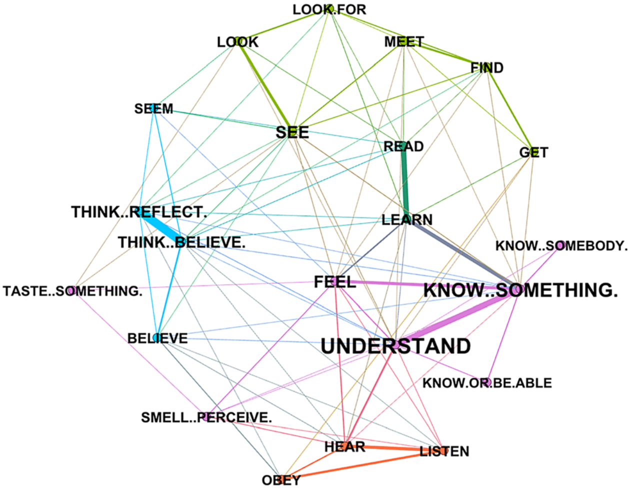

**Figure 12a:** Colexification network (Eurasia).

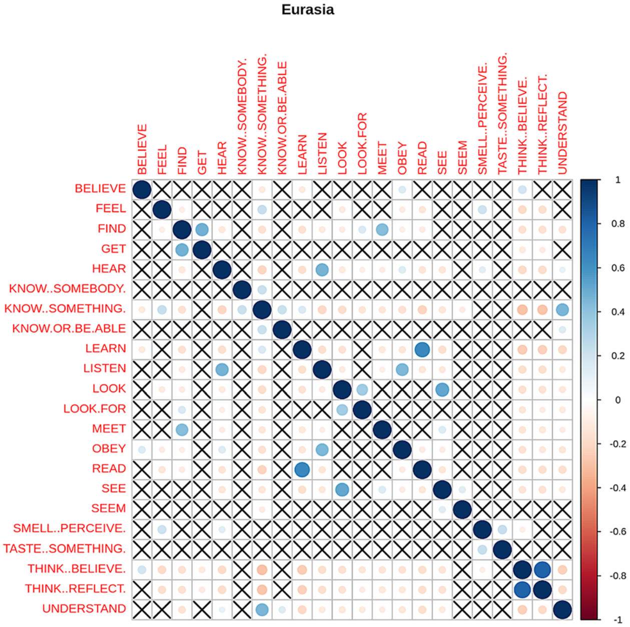

**Figure 12b:** Correlation plot (Eurasia).

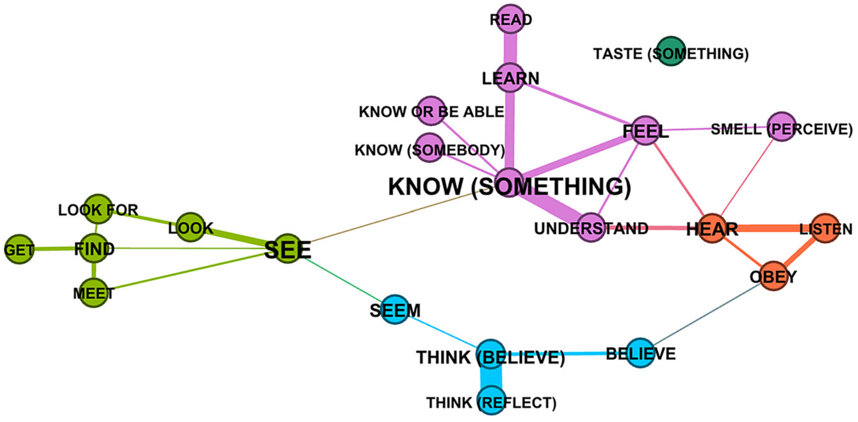

**Figure 12c:** Semantic map (Eurasia).

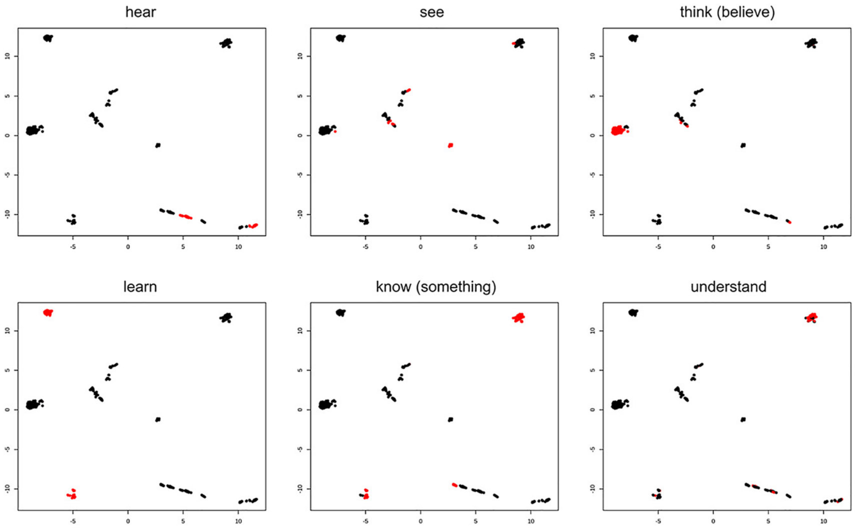

**Figure 12d:** UMAP plots reflecting semantic clustering of perception/cognition lexical items in Eurasian languages. Verbs co-expressing a particular meaning (reflected in the subplot's title) are shown in red.

the case in the general map, albeit the pattern is more robust in terms of frequency in Eurasia.

The semantic domains of VISION and BELIEF are remarkably independent from the other parts of the network in this macro-area. In the domain of VISION, non-controlled visual experience (see) is the bridge to cognition, but the link <see, know> is weak (cf. the negative correlation between the two meanings in Figure 12b), even if represented in four language families (Austroasiatic, Dravidian, Indo-European, Sino-Tibetan).

Among the sensory modalities, taste and smell are rarely colexified with other meanings in Eurasia as appears from the semantic map in Figure 11c. Noteworthy, however, is the significant positive correlation <taste, smell> in this area (Figure 12b), which is attested twice in languages from Eurasia (Northern Yukaghir [Yukaghir] *morej* and Middle High German [Indo-European] *smecken*) and twice in South America (with Lengua [Lengua-Mascoy] *lingaiyi* and Kaingáng *meng*).

The UMAP plot for Eurasia (Figure 12d), the region with the highest number of colexifying verbs by far (566), partly replicates the global structure of several clusters (Figure 8). This is related to the fact that the general UMAP depends strongly on the Eurasian UMAP. Colexified <know, understand> form a separate cluster along with colexified <learn, know> and learn; these three clusters reflect the chain of meaning learn-know-understand that we identified above in the domain of cognition. The three other main clusters correspond to verbs whose meanings are associated with think, hear, and see respectively. As mentioned with reference to the semantic map in Figure 12c, when <know, understand> do not go together, understand tends to be colexified with hear-verbs, and know with learn-verbs.

## 4.2 Africa

For the African macro-area, we have information for 19 meanings. They are connected by 41 edges in the full colexification network (Figure 13a) and by 27 edges in the semantic map (Figure 13c), for which we kept edges with weight two or more, resulting in a goodness of fit of 91% for this model (accounting for 97 out of 107 colexification patterns).

In this macro-area, the heavily colexified meanings see/look and hear/listen belong to two independent clusters (see the negative correlation between these respective meanings in Figure 13b). Interestingly, however, they are connected by the meaning taste, which is a feature that is not visible in the general semantic map of Figure 7. In fact, the <hear, taste> colexification, on one hand, is frequent ($N = 11$) in the global dataset, but is especially well represented in languages from Africa (Atlantic-Congo: five cases; Afro-Asiatic: one case), which explains the positive correlation between the two meanings in the corrplot (Figure 13b).[28] The

[28] Even though some of the correlations are shown as significant (e.g., understand vs. obey and understand vs. feel), these are mostly supported by negative data (cases where neither of the respective meanings are present). Actual positive evidence for them is extremely slim.

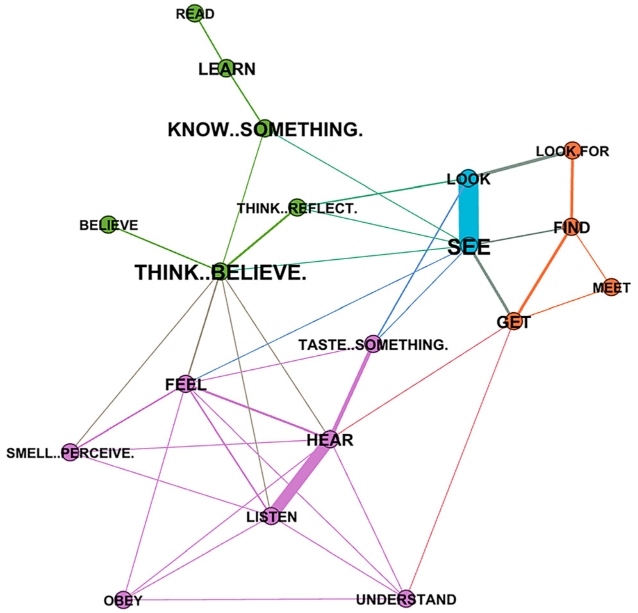

**Figure 13a:** Colexification network (Africa).

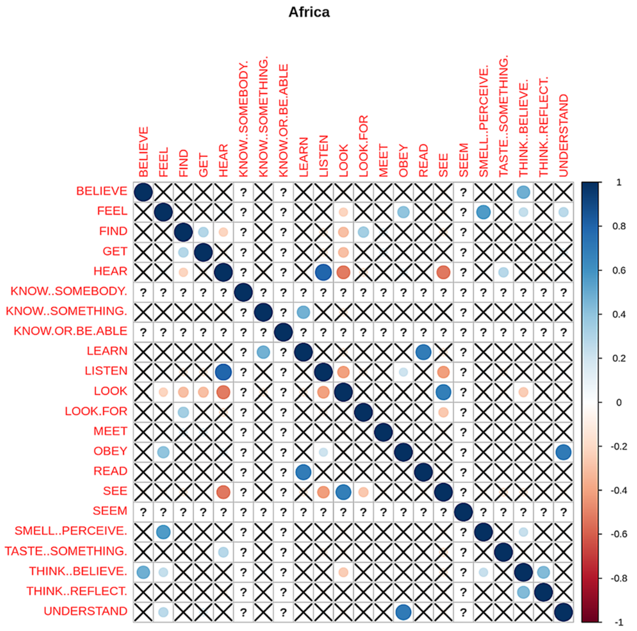

**Figure 13b:** Correlation plot (Africa).

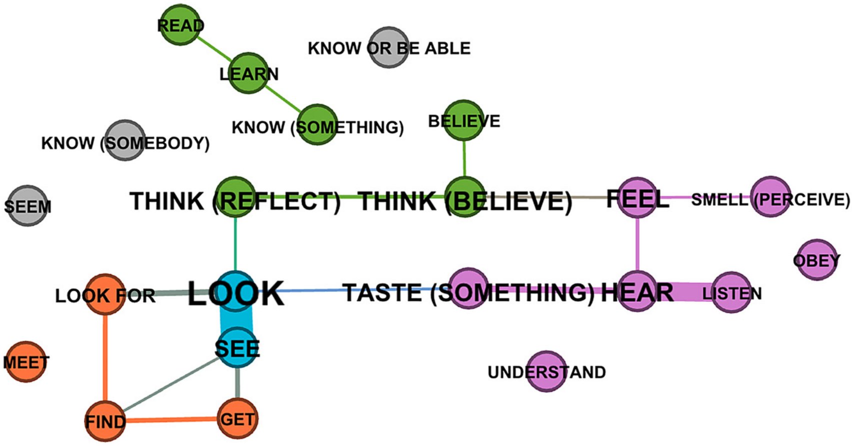

**Figure 13c:** Semantic map (Africa).

colexifications <LOOK, TASTE> (three cases) and <SEE, TASTE> (two cases), on the other hand, are rare and two African languages from the Atlantic-Congo family account for three cases: *iambuya* in Vunjo (<SEE, LOOK, TASTE>), and *ilola* in Machame (<LOOK, TASTE>). As a result, all the sensory modalities are linearly connected in this macro-area (SEE/LOOK-TASTE-HEAR/LISTEN-FEEL-SMELL): contrary to what happens in other areas, the meanings KNOW and UNDERSTAND do not mediate between VISION, on the one hand, and other sensory modalities, on the other.

The group of meanings KNOW, LEARN, and READ, which are significantly correlated in this macro-area, are quite independent from the domain of VISION, with only one case of co-expression <SEE, KNOW> (*we* in Lame (Peve), Afro-Asiatic, Africa). This is a second feature characteristic of this macro-area: cognition verbs like KNOW and UNDERSTAND do not mediate between VISION on the one hand and other sense modalities on the other hand.

Finally, SMELL and FEEL show a strong positive correlation only in Africa, a weaker one in Eurasia and South America, and no correlation in Papunesia. This correlation is however based on a single case for Eurasia, and two in Papunesia and Africa, while the colexification occurs four times in South American languages belonging to four different language families (the correlation is weaker there because there are three cases with verbs having the meaning SMELL but not FEEL vs. zero such cases in Africa).

The structure of the plot for the lexical items in the African subsample (98 observations) is tripartite (Figure 13d). Two major groups are dominated by HEAR and SEE verbs, which totally divide UNDERSTAND between them, and the very small

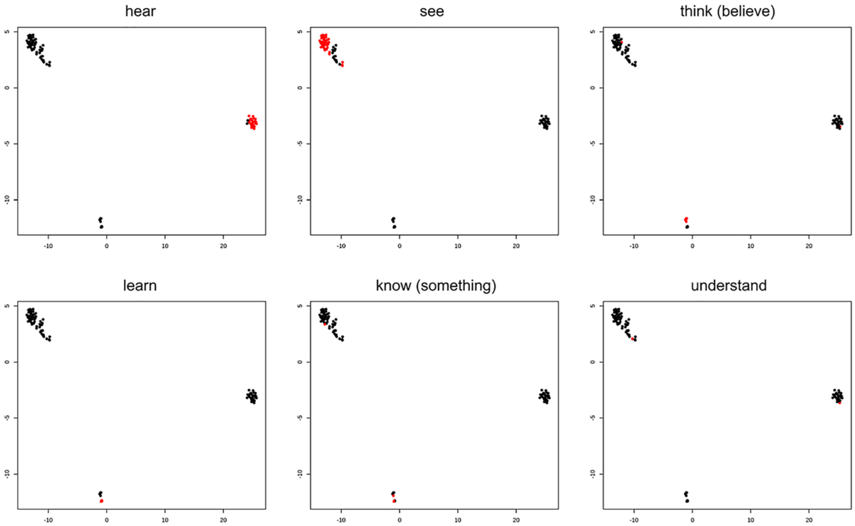

**Figure 13d:** UMAP plots reflecting semantic clustering of perception/cognition lexical items in African languages. Verbs co-expressing a particular meaning (reflected in the subplot's title) are shown in red.

third cluster consists of THINK, LEARN, and KNOW verbs. The very small size of the cognition subsample for this area (four LEARN verbs, eight THINK verbs, and four KNOW verbs) makes the interpretation of these results difficult.

## 4.3 Papunesia

For the Papunesian macro-area, we have information for 17 meanings. They are connected by 32 edges in the full colexification network (Figure 14a) and 22 edges in the semantic map (Figure 14c), for which we kept edges with weight two or more, resulting in a goodness of fit of 90% (accounting for 67 out of 75 colexifications).

This macro-area deviates from the state of affairs found in both Eurasia and Africa in that it strongly associates the TASTE modality with the general FEEL meaning and, to a lesser extent, with HEAR and OBEY, as is reflected in the corrplot of Figure 14b, which displays a positive correlation between these meanings. The semantic map in Figure 14c shows that these meanings are not only correlated, but can be also seen as a chain TASTE-FEEL-HEAR-OBEY. It is important to stress however—and this is a limitation of our approach with a limited number of observations for a macro-area—that this association is not the result of a frequently attested polysemy pattern, since only two lexemes, namely *roŋo* and *roŋo-hia*, which are clearly

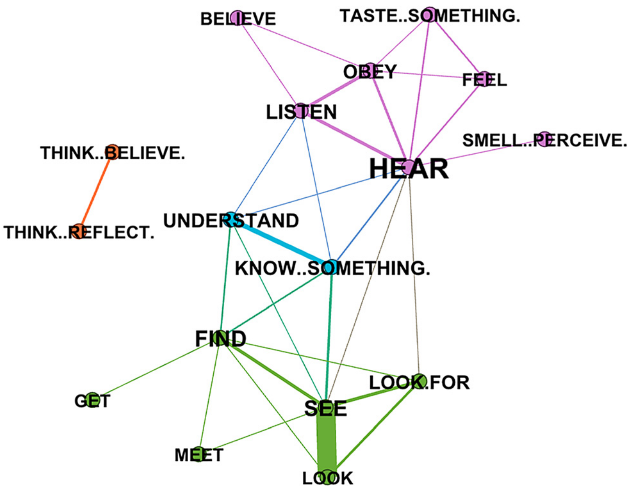

**Figure 14a:** Colexification network (Papunesia).

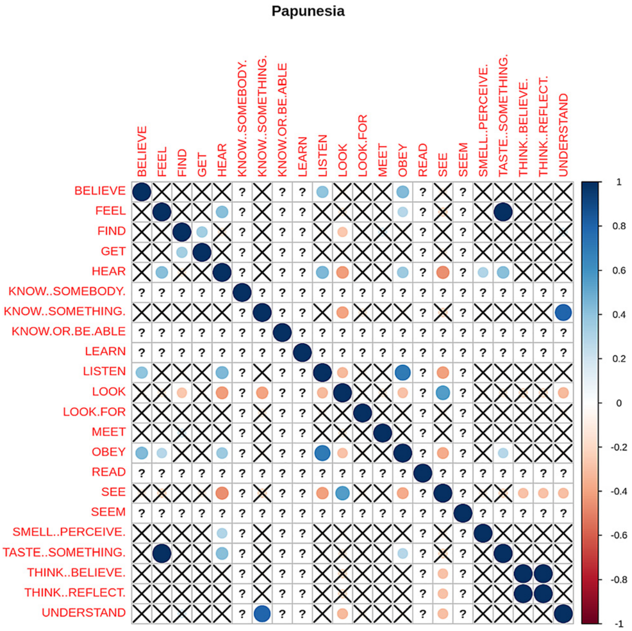

**Figure 14b:** Correlation plot (Papunesia).

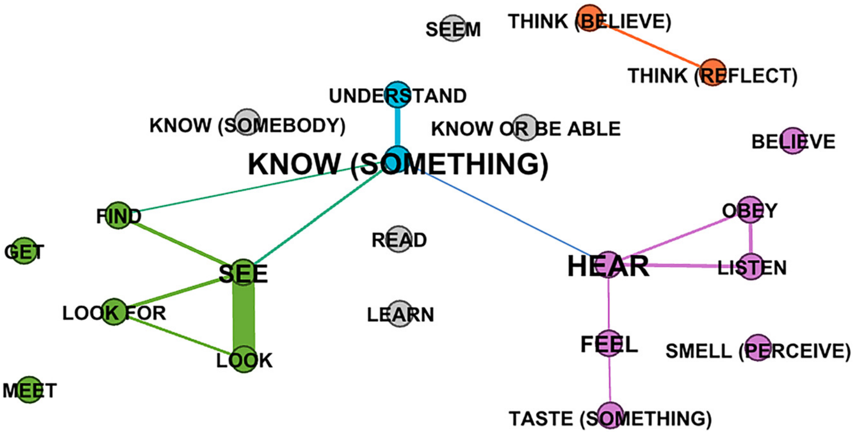

**Figure 14c:** Semantic map (Papunesia).

related and are both found in one language (Maori [Austronesian, Papunesia]), instantiate this pattern: *roŋo* colexifies the meanings <FEEL, TASTE, HEAR> and *roŋo-hia* can additionally mean OBEY.

Another point in which Eurasia and Africa differ from Papunesia is the relation of HEAR to KNOW, rather than to UNDERSTAND, as an intermediary meaning (Figure 14c), in two languages of the sample (Takia [Austronesian, Papunesia] and Kui [Timor-Alor-Pantar, Papunesia]). However, this deviation from the patterns attested in the other two macro-areas could possibly be attributed to a semantic vagueness between the two meanings. As a matter of fact, KNOW and UNDERSTAND are very strongly correlated in this area (cf. Figure 14b): the six verbal lexemes with the meaning UNDERSTAND (all from Austronesian languages) always express at least the meaning KNOW as well. The <HEAR, KNOW> colexification will be discussed further below in the section about South America, where this pattern occurs as well.

Finally, the semantic domain of THINKING is entirely disconnected from the other meanings of the perception/cognition domain in Papunesia. While OBEY is (weakly) connected to BELIEVE, as in other areas worldwide, the connection <BELIEVE, THINK> that one finds elsewhere is not documented in the CLICS$^2$ dataset for Papunesia.

The UMAP plot for the Papunesian lexical subsample (60 observations) shows a simple bipartite structure (Figure 14d). It is based on a somewhat messy separation between HEAR and SEE verbs (the <HEAR, SEE> colexification itself was found twice in Australia and once in Papunesia and South America). There are no LEARN verbs in this subsample; UNDERSTAND and KNOW (SOMETHING) are tilted towards the HEAR cluster. The number of data points makes it hard to draw any definite conclusions from this plot.

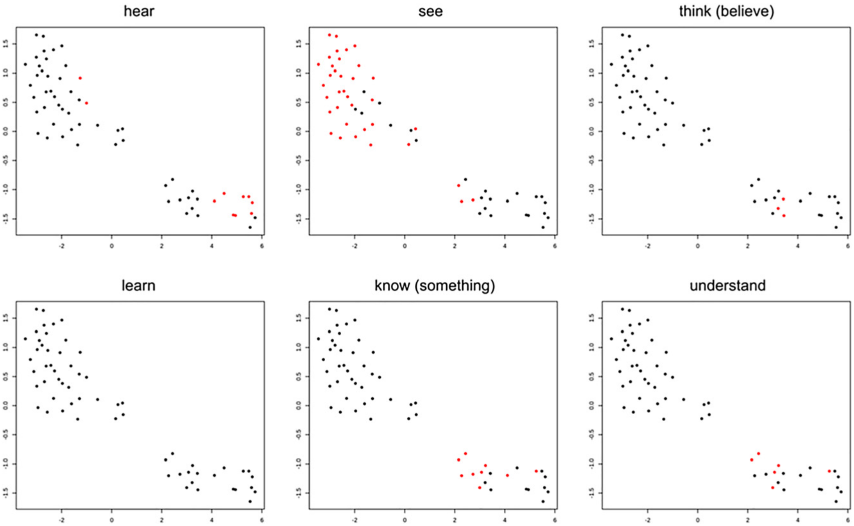

**Figure 14d:** UMAP plots reflecting semantic clustering of perception/cognition lexical items in Papunesian languages. Verbs co-expressing a particular meaning (reflected in the subplot's title) are shown in red.

## 4.4 South America

For the South American macro-area, we have information for 20 meanings. They are connected by 78 edges in the full colexification network (Figure 15a) and 44 edges in the semantic map (Figure 15c), for which we kept edges with weight two or more, resulting in a goodness of fit of 93% for this model (accounting for 279 out of 300 colexification patterns). Like in other macro-areas, four main clusters show up, but unlike in the other subsamples, the cluster around THINK is more tightly connected to cognition (through KNOW) and to the sense modalities around HEAR (through the meanings FEEL and OBEY).

Similarly to the situation found in Papunesia, HEAR and KNOW are connected. Generally speaking, the colexification pattern <HEAR, KNOW> is diverse from a genealogical and areal point of view: 11 language varieties from seven language families. It is found three times in South America, twice in Papunesia and once in Eurasia, while five Australian languages (belonging to the same language family) actualize this rarer pattern (Figure 16).

In South America, this colexification pattern is limited to a northern geographical region containing a group of three languages, Orejón, Yuwana, and Waorani, which belong to different language families (Tucanoan, Jodi-Saliban,

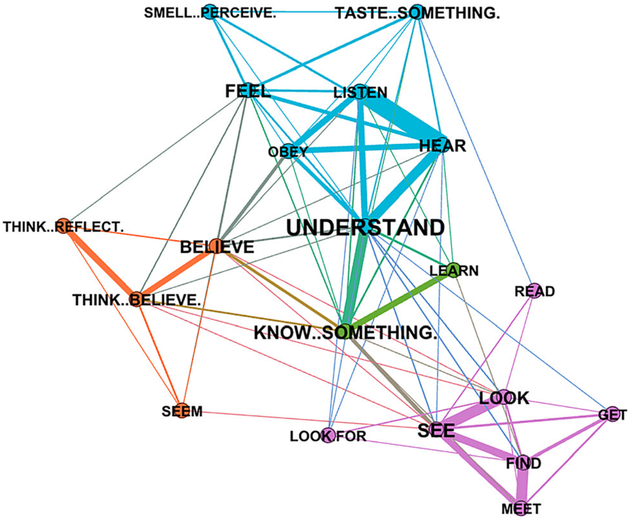

**Figure 15a:** Colexification network (South America).

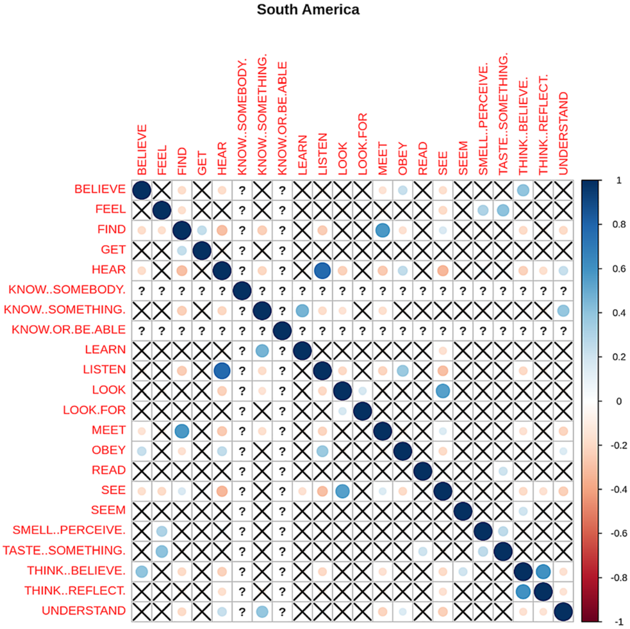

**Figure 15b:** Correlation plot (South America).

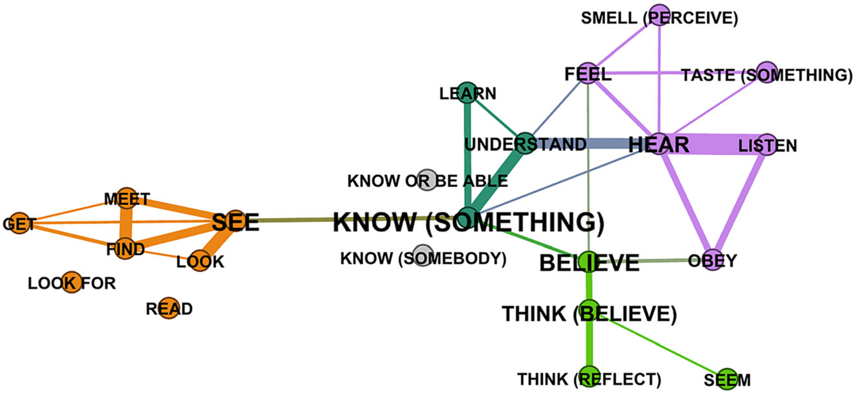

**Figure 15c:** Semantic map (South America).

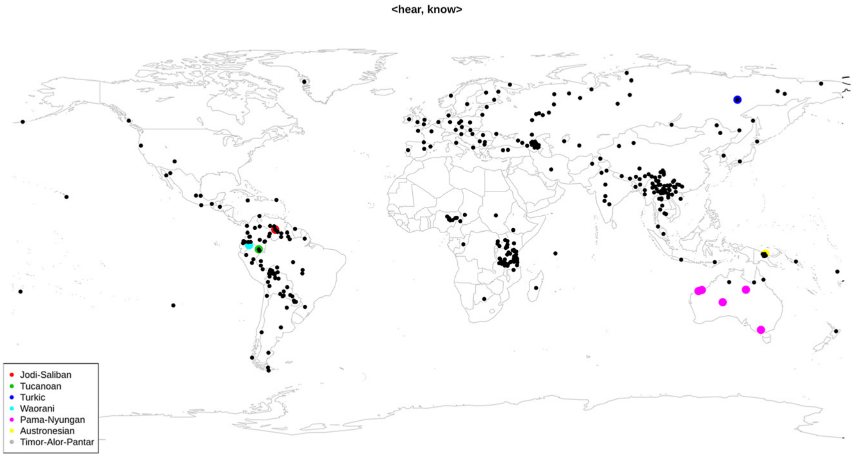

**Figure 16:** The <HEAR, KNOW> colexification worldwide (black-colored dots represent no attestation of the colexification pattern and other-than-black-colored dots signal that the language varieties show this pattern).

and Waorani, respectively). This makes it a possible candidate of an areally-biased colexification pattern (cf. Thomason 2001: 99), even though actual contact between these three languages is far from ascertained. Note that, overall, HEAR and KNOW are negatively correlated in this macro-area. Conversely, there is a positive correlation between HEAR and UNDERSTAND, which is also significant, albeit less strong, in Eurasia.

Another candidate for an areally biased colexification pattern comes from the domain of perception. Specifically, the <TASTE, FEEL> colexification is found in five languages of South America (see Figure 17), which amounts to more than half of the total *N* of this colexification in the CLICS$^2$ language sample ($N_{total} = 8$). The close connection of the two meanings is also reflected in their positive correlation in Figure 14b (cf. the negative or weak correlation between the two meanings in other macro-areas). More importantly, these five languages belong to four different language families and three of the language varieties are found in Brazil (Catuquina [Pano-Tacanan], Yaminahua [Pano-Tacanan], and Waurá [Arawakan]). The fact that a genealogically heterogeneous cluster of languages shares the <TASTE, FEEL> colexification makes it a good candidate for a pattern that results from diffusion events rather than inheritance (see Koptjevskaja-Tamm and Liljegren 2017, who consider shared colexification patterns as an indicator of areality).

The pattern of the UMAP plot in South America, the region with the second largest sample size (199 observations), is tripartite (Figure 18). Two distinct groups are dominated by HEAR and SEE respectively, and the middle one is dominated by KNOW (SOMETHING). Both UNDERSTAND and LEARN are distributed between these three groups; however, the number of LEARN verbs in the sample is rather small ($N = 14$).

As observed in Section 4, South America is a heavy colexifier, with no less than 17 verbs expressing four meanings or more. These polysemic items can be categorized into two main groups: one with verbs expressing the meanings UNDERSTAND, HEAR, LISTEN and OBEY, and one associating KNOW and UNDERSTAND with other sensory or cognitive modalities.

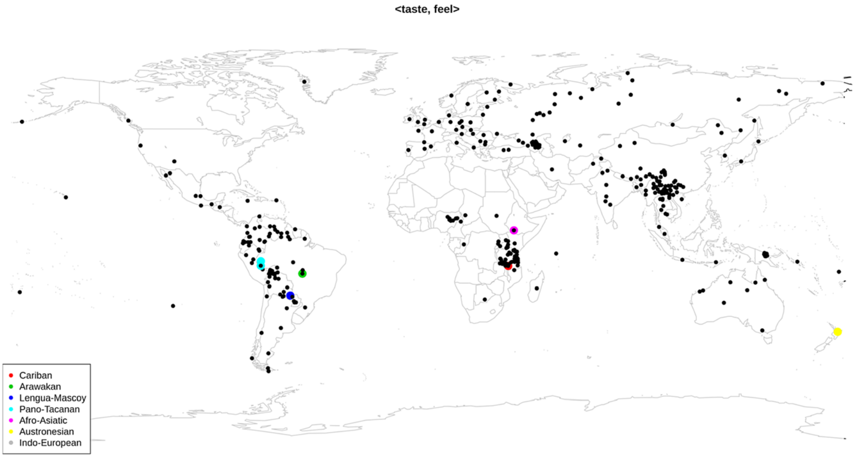

**Figure 17:** The <TASTE, FEEL> colexification worldwide.

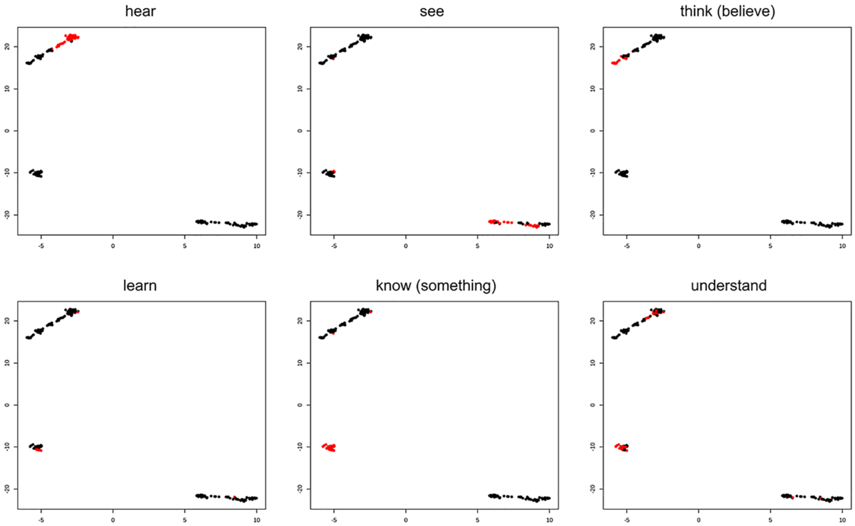

**Figure 18:** UMAP plots reflecting semantic clustering of perception/cognition lexical items in South American languages. Verbs co-expressing a particular meaning (reflected in the subplot's title) are shown in red.

# 5 General discussion

We now summarize the main points made above. First, it should be noted that generalizations about the cross-linguistic organization of the lexicon are not easily or straightforwardly identified. For one to be able to formulate such generalizations, access to large datasets is needed. In the past, the availability of such datasets was rather limited. Given this limitation, it is not surprising that the number of languages of a typical lexico-typological study ranged from 10 to 50 (see Koptjevskaja-Tamm et al. 2015: 436). A major type of exception are studies that rely on massively parallel texts (see Östling 2016; Wälchli 2010; Wälchli and Cysouw 2012; specifically Wälchli 2016 on perception verbs). However, due to the specific genre that these studies are based on (mainly religious texts, in particular the New Testament), the variation in the concepts that can be analyzed is limited. Luckily, the increasing availability of resources that contain a large amount of lexical information makes large-scale typological studies on the lexicon possible nowadays. In particular, large lexical databases such as CLICS and ASJP have recently been used in order to investigate areal factors in lexical typology (Gast and Koptjevskaja-Tamm 2018).

Our study extends this approach, by exploring the possibility of identifying significant generalizations about areal patterns of co-expression in the lexicon. We focused on the domains of perception and cognition, two semantic fields that are central to human experience, as represented in a number of different datasets, i.e., Vanhove's (2008) dataset, Multilingual WordNet, and an improved version of the Database of Cross-Linguistic Colexifications, CLICS$^2$. We applied different methods (Table 3), which proved suitable for answering different types of questions and shed complementary light on the same dataset.

**Table 3:** The main techniques used in the current study.

| | | **Type of approach** | |
|---|---|---|---|
| | | ***Quantitative*** | ***Qualitative*** |
| Scope | *Narrow* | Correlation analysis | Colexification networks |
| | *Wide* | Dimensionality reduction | Semantic maps |

To sum up, one can approach colexification matrices from a qualitative or quantitative point of view. The qualitative viewpoint involved comparing semantic maps and colexification networks of different macro-areas, whereas more quantitative exploratory ones included correlation plots and dimensionality-reduction techniques. Additionally, both types of approach can have a narrow or wide scope. Regarding the qualitative approaches, the distinction between wide- and narrow-scope analysis refers to the possibility of taking into account complex broader colexification patterns, namely patterns that extend beyond pairwise meanings associations. As Gast and Koptjevskaja-Tamm (2018: 52) put it "we should always look at broader patterns of multifunctionality before jumping to conclusions about pairwise meaning associations". Semantic maps allow us to adopt this broader view, whereas colexification networks, as fascinating as they are, are restricted to pairwise associations. As for the quantitative approaches, there is a continuum from narrow-scope methods with a possibility of significance testing (correlations) to wide-scope methods without significance testing (dimensionality reduction).

Table 3 summarizes the differences among the four techniques used in the current study.

All four techniques are suitable for large datasets. In Section 3.1, we presented an additional tool, namely the formal concept lattice, which has the advantage of explicitly displaying the association between form and content as well as the hierarchical structure of the concepts, but on the other hand does not allow for an easy exploration of large-scale datasets due to its representational complexity.

Our results go beyond the clear representation and visualization of data, as the studies revealed new insights about the general and macro-areal structure of the semantic domains explored here. We list some of these in the following. First, our semantic maps show that intrafield connections between verbs of vision and hearing are mediated by interfield connections, i.e., via the cognition domains of knowledge and understanding. This result is robust, turning up in several datasets and across macro-areas. Second, the Multilingual WordNet dataset reveals one particularly interesting finding: controlled activities, as those instantiated by such verbs as *look* or *listen*, are not directly linked to cognition: the verbs expressing un-controlled experiences (SEE and HEAR) are. The same result was obtained from CLICS$^2$, which confirms these observations based on a typologically more diverse dataset. Third, further results from CLICS$^2$ include the finding that knowledge is more closely linked to vision, and mental manipulation (i.e., understanding) to hearing. Finally, meanings belonging to other sensory modalities (taste, smell (perceive), feel) cluster with HEAR rather than with SEE: this is an important result which points in the direction of a conceptual split between visual and other types of perception. All of these are good candidates for a universal principle of the organization of lexicons, which might act as a functional trigger, biasing the probabilities of particular developmental pathways. They also provide interesting hypotheses for other disciplines, in particular experimental approaches to cognition.

In order to uncover patterns that are specific to particular macro-areas (and, potentially) to micro-areas, we used a set of complementary methods. These methods were only applied to the CLICS$^2$ dataset, because it is the richest and most areally balanced dataset at our disposal.

In Africa, the colexification <TASTE, HEAR> is well-attested and the two meanings are positively correlated; most importantly the domains of vision and hearing are not mediated by meanings pertaining to the domain of cognition, such as KNOW and UNDERSTAND (cf. Vanhove 2008), but by the meaning TASTE itself,[29] which yields a continuum of meanings associated sensory modalities in the African macro-area: SEE/LOOK-TASTE-HEAR-FEEL-SMELL. Finally, the cluster that contains the meanings KNOW, LEARN, and READ is independent from the domain of vision. The state of affairs found in Eurasia does not differ much from the global picture. This is not unexpected, given the high proportion of colexification patterns of this area. However, some correlations between meanings appear to be stronger in Eurasia, e.g., UNDERSTAND is strongly correlated with HEAR and KNOW with SEE, and the main clusters (SEE, KNOW,

[29] Interestingly, Nakagawa (2012) stressed the central role of TASTE among the perception verbs of three little documented Khoe languages.

THINK, HEAR) are overall less interconnected. A striking feature of Papunesia is the semantic vagueness between KNOW and UNDERSTAND, which causes HEAR to be associated with KNOW, or more precisely, to the couple <KNOW, UNDERSTAND> through KNOW, rather to UNDERSTAND. In this area, the cluster THINK/BELIEVE appears disconnected from the other meanings belonging to the perception and cognition domains. In South America, a heavily colexifying macro-area, two possible candidates of areally biased colexification patterns have been identified: <TASTE, FEEL> in the north and <HEAR, KNOW> in the central area. Detailed studies are needed in order to assess these hypotheses.

We now turn to some limitations of the present study and the datasets upon which it is based. First, the sample is not ideal. Despite the significant improvement in the new version of CLICS$^2$ and although the sample is putatively global, it is in effect skewed towards Eurasia when macro-areas are considered. For some language varieties and for certain concepts in several language varieties, a lack of data distorts the picture: the rate of the Average Mutual Coverage is as low as 0.3 for the 22 meanings belonging to a central semantic domain such as perception and cognition, which has of course a significant impact when one sets out to study areal phenomena.

The second limitation relates to methodology. If one focuses on one semantic domain, even if shared across all languages, and does not try to identify areally biased patterns first (as do Gast and Koptjevskaja-Tamm 2018), limited results may obtain. In the present study, this is reflected in the low number of possible candidates for areally-specific colexification patterns that were identified. Furthermore, while semantic maps prove definitely useful for capturing cross-linguistic regularities, an important caveat must be kept in mind: they over-generate possible constellations of meanings. While this does not detract from the positive findings, it means that there is a gap between predicted and documented colexification patterns. It might be that these gaps are due to chance, but it also may be that these gaps are hiding places for hitherto unnoticed semantic connections, on the one hand, or historical contact events, on the other.

Regarding the quantitative methods we used for studying areality, the main catch is that even those that allow for significance testing (e.g., correlations) can be fooled by unbalanced samples, as in the correlation based on one positive + positive, one positive-negative, and 96 negative-negative values (cf. fn. 28). The bottom line is that whatever the tools used, the data are often not rich enough to properly assess the causal factors underlying shared colexification patterns in a given area. These may include inherent semantic factors, language contact, inheritance due to genealogy, and more.

These limitations may be seen as challenges for further exploration of quantitative methods for determining the causes for the distribution of colexification patterns in large samples. An especially important goal is the possibility of estimating the base probability of particular semantic shifts, such as SEE > UNDERSTAND. Insight into this matter might be provided by models of semantic change that look for correlations between frequency and semantic change (Dubossarsky et al. 2015; Hamilton et al. 2016; Kutuzov et al. 2018; Tang 2018). Such distributional models, however, still have many problems that make their immediate application to this question impractical. Another goal is the development of dense samples of different areas in the world, with complete datasets for particular domains. This would allow a better chance of directly testing the relative contributions of inheritance, language contact, and inherent semantics. It will also provide more quality data which will assist attempts of inferring diachronic information from synchronic polysemies (Dellert 2016).

Despite these caveats, the present study supports the usefulness of bottom-up exploratory research as a means to bring to light cross-linguistic generalizations about the structure of lexicons. These generalizations, in principle, can be at any level of areality, from the micro-areal to the global, the main limitation being the quantity and quality of data available. These generalizations, beyond their interest for linguists, may potentially provide novel hypotheses to be tested in other disciplines, such as experimental psychology. Finally, the data-related limitations may point to a fruitful avenue for future research, namely, genealogically and/or areally dense samples, which would allow more detailed studies of the actual historical pathways of change involved in the innovation, diffusion, and loss of meanings associated with lexical items.

| Language | Glottocode | Top level family | Macro-area |
|---|---|---|---|
| Albanian | alba1267 | Indo-European | Eurasia |
| Arabic | stan1318 | Afro-Asiatic | Eurasia |
| Araki | arak1252 | Austronesian | Papunesia |
| Basque | basq1248 | | Eurasia |
| Bulgarian | bulg1262 | Indo-European | Eurasia |
| Catalan | stan1289 | Indo-European | Eurasia |
| Catuquina | pano1254 | Pano-Tacanan | South America |
| Chinese (Mandarin) | mand1415 | Sino-Tibetan | Eurasia |
| Chinese (Taiwan) | taib1240 | Sino-Tibetan | Eurasia |
| Croatian | sout1528 | Indo-European | Eurasia |
| Danish | dani1285 | Indo-European | Eurasia |
| Dutch | dutc1256 | Indo-European | Eurasia |
| Djaru | jaru1254 | Pama-Nyungan | Australia |
| Dwot | dass1243 | Afro-Asiatic | Africa |
| English | stan1293 | Indo-European | Eurasia |
| Finnish | finn1318 | Uralic | Eurasia |
| French | stan1290 | Indo-European | Eurasia |

(continued)

| Language | Glottocode | Top level family | Macro-area |
|---|---|---|---|
| Galician | gali1258 | Indo-European | Eurasia |
| Greek | mode1248 | Indo-European | Eurasia |
| German | stan1295 | Indo-European | Eurasia |
| Hebrew | hebr1245 | Afro-Asiatic | Eurasia |
| Indonesian | indo1316 | Austronesian | Papunesia |
| Italian | ital1282 | Indo-European | Eurasia |
| Jahai | jeha1242 | Austroasiatic | Papunesia |
| Japanese | nucl1643 | Japonic | Eurasia |
| Kaingáng | saop1235 | Nuclear-Macro-Je | South America |
| Kolyma Yukaghir | sout2750 | Yukaghir | Eurasia |
| Kui | kuii1253 | Timor-Alor-Pantar | Papunesia |
| Kuku-Yalanji | kuku1273 | Pama-Nyungan | Australia |
| Lame (Peve) | peve1243 | Afro-Asciatic | Africa |
| Lengua (Northern and Southern) | nort2971<br>sout2989 | Lengua-Mascoy | South America |
| Malay | indo1326 | Austronesian | Eurasia |
| Maori | maor1246 | Austronesian | Papunesia |
| Middle High German | midd1343 | Indo-European | Eurasia |
| Manchu | manc1252 | Tungusic | Eurasia |
| Northern Yukaghir | nort2745 | Yukaghir | Eurasia |
| Norwegian | norw1258 | Indo-European | Eurasia |
| Norwegian Bokmål | norw1259 | Indo-European | Eurasia |
| Nunggubuyu | nung1290 | Gunwinyguan | Australia |
| Orejón | orej1242 | Tucanoan | South America |
| Persian | fars1254 | Indo-European | Eurasia |
| Polish | poli1260 | Indo-European | Eurasia |
| Portuguese | port1283 | Indo-European | Eurasia |
| Russian | russ1263 | Indo-European | Eurasia |
| Sar | sarr1246 | Nilo-Saharan | Africa |
| Siona | sion1247 | Tucanoan | South America |
| Slovenian | slov1268 | Indo-European | Eurasia |
| Spanish | stan1288 | Indo-European | Eurasia |
| Swedish | swed1254 | Indo-European | Eurasia |
| Takia | taki1248 | Austronesian | Papunesia |
| Thai | thai1261 | Tai-Kadai | Eurasia |
| Yulu | yulu1243 | Central Sudanic | Africa |
| Yuwana | yuwa1244 | Jodi-Saliban | South America |
| Vunjo | vunj1238 | Atlantic-Congo | Africa |
| Waorani | waor1240 | Waorani | South America |
| Warluwarra | warl1256 | Pama–Nyungan | Australia |
| Waurá | waur1244 | Arawakan | South America |

**Acknowledgments:** We thank two anonymous reviewers as well as Maria Koptjevskaja-Tamm for their incisive and constructive comments.
**Research funding**: Stéphane Polis acknowledges the support of the National Fund for Scientific Research (Belgium). Thanasis Georgakopoulos would like to acknowledge the support of the Marie Curie BeIPD postdoctoral grant from the European Commission (Grant number: 600405). In addition, this article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

# References

Aikhenvald, Alexandra Y. & Anne Storch (eds.). 2013. *Perception and cognition in language and culture*. Leiden: Brill.

van der Auwera, Johan & Vladimir A. Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2(1). 79–124.

Bastian, Mathieu, Sebastien Heymann & Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. In Proceedings of the Third International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154 (accessed 1 December 2018).

Bickel, Balthasar. 2017. Areas and universals. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 40–55. Cambridge: Cambridge University Press.

Bickel, Balthasar. 2020. Large and ancient linguistic areas. In Mily Crevels & Pieter Muysken (eds.), *Language dispersal, diversification, and contact: A global perspective*, 78–99. Oxford: Oxford University Press.

Bickel, Balthasar, Alena Witzlack-Makarevich & Taras Zakharko. 2014. Typological evidence against universal effects of referential scales on case alignment. In Ina Bornkessel-Schlesewsky, Andrej Malchukov & Marc Richards (eds.), *Scales: A cross-disciplinary perspective on referential hierarchies*, 7–43. Berlin: De Gruyter Mouton.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte & Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10(2008). P10008. arXiv:0803.0476 [physics.soc-ph]. (accessed 31 January 2019).

Bond, Francis & Kyonghee Paik. 2012. A survey of wordnets and their licenses.In *Proceedings of the 6th Global WordNet Conference (GXC 2012)*, 64–71. Japan: Matsue.

Brenzinger, Matthias & Anne-Maria Fehn. 2013. From body to knowledge: Perception and cognition in Khwe-Ani and Ts'ixa. In Alexandra Y. Aikhenvald & Anne Storch (eds.), *Perception and cognition in language and culture*, 161–191. Leiden: Brill. https://doi.org/10.1163/9789004210127_008.

Burenhult, Niclas & Asifa Majid. 2011. Olfaction in Aslian ideology and language. *The Senses & Society* 6(1). 19–29.

Caballero, Rosario & Iraide Ibarretxe-Antuñano. 2014. Ways of perceiving, moving, and thinking: Re-vindicating culture in conceptual metaphor research. *Cognitive Semiotics* 5(1–2). 268–290.

Classen, Constance. 1997. Foundations for an anthropology of the senses. *International Social Science Journal* 49(153). 401–412.

ConExp Project. 2006. The concept explorer. http://conexp.sourceforge.net/users/documentation/index.html (accessed 31 January 2019).

Cristofaro, Sonia. 2010. Semantic maps and mental representation. *Linguistic Discovery* 8(1). 35–52.

Croft, William. 2001. *Radical construction grammar. Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Croft, William. 2003. *Typology and universals*, 2nd edn. Cambridge: Cambridge University Press.

Croft, William. 2010. What do semantic maps tell us? Comment on 'Semantic maps and mental representation' by Sonia Cristofaro. *Linguistic Discovery* 8(1). 53–60.

Cysouw, Michael. 2007. Building semantic maps: The case of person marking. In Bernhard Wälchli & Matti Miestamo (eds.), *New challenges in typology*, 225–248. Berlin and New York: Mouton De Gruyter.

Cysouw, Michael, Martin Haspelmath & Andrej Malchukov. 2010. Introduction to the special issue 'Semantic maps: Methods and applications'. *Linguistic Discovery* 8(1). 1–3.

Dellert, Johannes. 2016. Using causal inference to detect directional tendencies in semantic evolution. In Seán G. Roberts, Christine Cuskley, Luke McCrohon, Lluís Barceló-Coblijn, Olga Fehér & Tessa Verhoef (eds.), *The evolution of language: Proceedings of the 11th international conference (EVOLANGX11)*. Available at: http://evolang.org/neworleans/papers/139.html (accessed 1 March 2019).

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13. 257–292.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer & Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In Vito Pirrelli, Claudia Marzi & Marcello Ferro (eds.), *Word structure and word usage. Proceedings of the NetWordS Final Conference*, 66–70. Pisa, Italy.

Evans, Nicholas & David Wilkins. 2000. In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language* 76(3). 546–592.

François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove (ed.), *From polysemy to semantic change*, 163–215. Amsterdam: Benjamins.

Gast, Volker & Maria Koptjevskaja-Tamm. 2018. The areal factor in lexical typology: Some evidence from lexical databases. In Daniel Van Olmen, Tanja Mortelmans & Brisard Frank (eds.), *Aspects of linguistic variation*, 43–82. Berlin: De Gruyter Mouton.

Georgakopoulos, Thanasis & Stéphane Polis. 2018. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass* 12(2). e12270.

Georgakopoulos, Thanasis & Stéphane Polis. 2021. Lexical diachronic Semantic maps. Mapping the evolution of time-related lexemes. *Journal of Historical Linguistics*. Online first article. https://doi.org/10.1075/jhl.19018.geo.

Georgakopoulos, Thanasis, Daniel A. Werning, Jörg Hartlieb, Tomoki Kitazumi, Lidewij E. van de Peut, Annette Sundermayer & Gaëlle Chantrain. 2016. The meaning of ancient words for 'earth': An exercise in visualizing colexification on a semantic map. *eTopoi. Journal for Ancient Studies* 6. 1–36.

Greenberg, Joseph H. 1978. Diachrony, synchrony, and language universals. In Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of human language: Word structure*, Vol. 3, 47–82. Stanford: Stanford University Press.

Guerrero, Lilián. 2010. El amor no surge de los ojos sino de los oídos: Asociaciones semánticas en lenguas yuto-aztecas. *Onomazéin* 21(1). 47–69.

Hamilton, L. William, Jure Leskovec & Dan Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2116–2121.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language*, 211–242. Mahwah, NJ: Lawrence Erlbaum.

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *World loanword database*. Munich: Max Planck Digital Library.

Ibarretxe-Antuñano, Iraide. 2008. Vision metaphors for the intellect: Are they really crosslinguistic? *Atlantis. Journal of the Spanish Association of Anglo-American Studies* 30(1). 15–33.

Ibarretxe-Antuñano, Iraide. 2013. The relationship between conceptual metaphor and culture. *Intercultural Pragmatics* 10(2). 315–339.

Koptjevskaja-Tamm, Maria, Ekaterina Rakhilina & Martine Vanhove. 2015. The semantics of lexical typology. In Nick Riemer (ed.), *The Routledge handbook of semantics*, 434–454. London & New York: Routledge.

Koptjevskaja-Tamm, Maria & Henrik Liljegren. 2017. Semantic patterns from an areal perspective. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 204–236. Cambridge: Cambridge University Press.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*. Available at: arXiv:1806.03537 (accessed 1 June 2019).

Levinson, C. Stephen & Asifa Majid. 2014. Differential ineffability and the senses. *Mind & Language* 29. 407–427.

List, Johann-Mattis, Simon Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi & Robert Forkel (eds.). 2018a. *Database of cross-linguistic colexifications*. Jena: Max Planck Institute for the Science of Human History. http://clics.clld.org (accessed 1 June 2019).

List, Johann-Mattis, Simon Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi & Robert Forkel. 2018b. CLICS$^2$: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology* 22(2). 277–306.

List, Johann-Mattis, Michael Cysouw, Simon Greenhill & Robert Forkel (eds.). 2018c. *Concepticon*. Jena: Max Planck Institute for the Science of Human History. Available at: http://concepticon.clld.org (accessed 15 June 2019).

List, Johann-Mattis, Robert Forkel, Christoph Rzymski & Simon J Greenhill. 2018d. clics/clics2: Creating colexification networks from lexical data (v1.0.1). Zenodo (accessed 6 May 2020).

Majid, Asifa & Niclas Burenhult. 2014. Odors are expressible in language, as long as you speak the right language. *Cognition* 130. 266–270.

Maslova, Elena. 2004. A universal constraint on the sensory lexicon, or when hear can mean 'see'? In Aleksandr P. Volodin (ed.), *Tipologičeskie obosnovanija v grammatike: k 70-letiju professora Xrakovskogo V.S.*, 300–312. Available at: http://anothersumma.net/Publications/Perception.pdf (accessed 1 April 2019).

Matisoff, James. 1978. *Variational semantics in Tibeto-Burman*. Philadelphia: ISHI.

McInnes, Leland & John Healy. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML].

Melville, James. 2018. uwot: The Uniform Manifold Approximation and Projection (UMAP) method for dimensionality reduction. R package version 0.0.0.9002. https://github.com/jlmelville/uwot (accessed 1 June 2019).

Muysken, Pieter. 2008. Introduction. In Pieter Muysken (ed.), *From linguistic areas to areal linguistics*, 1–23. Amsterdam: John Benjamins.

Nakagawa, Hirosi. 2012. The importance of TASTE verbs in some Khoe languages. *Linguistics* 50(3). 395–420.

Newman, Mark E. J. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 74. 036104.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. London/Chicago: University of Chicago Press.

Nichols, Johanna. 2003. Diversity and stability in languages. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 283–310. Malden/Oxford/Melbourne/Berlin: Blackwell Publishing.

Östling, Robert. 2016. Studying colexification through massively parallel corpora. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shifts*, 157–176. Berlin and Boston: De Gruyter Mouton.

Perrin, Loïc-Michel. 2010. Polysemous qualities and universal networks, invariance and diversity. *Linguistic Discovery* 8(1). 259–280.

Rakhilina, Ekaterina & Tatiana Reznikova. 2016. A frame-based methodology for lexical typology. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shifts*, 95–129. Berlin: De Gruyter.

Regier, Terry, Naveen Khetarpal & Asifa Majid. 2013. Inferring semantic maps. *Linguistic Typology* 17. 89–105.

Ryzhova, Daria & Sergei Obiedkov. 2017. Formal concept lattices as semantic maps. In Ekaterina L. Chernyak (ed.), *Computational linguistics and language science*, 78–87. Aachen CEUR Workshop Proceedings.

San Roque, Lila, Kobin H. Kendrick, Elisabeth Norcliffe, Penelope Brown, Rebecca Defina, Mark Dingemanse, Tyko Dirksmeyer, Nick Enfield, Simeon Floyd, Jeremy Hammond, Giovanni Rossi, Sylvia Tufvesson, Saskia van Putten & Asifa Majid. 2015. Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics* 26. 31–60.

San Roque, Lila, Kobin H. Kendrick, Elisabeth Norcliffe & Asifa Majid. 2018. Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics* 29(3). 371–406.

Sinnemäki, Kaius. 2014. A typological perspective on differential object marking. *Linguistics* 52(2). 281–313.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. In *Proceedings of the American Philosophical Society*, Vol. 96, 452–463.

Sweetser, Eve. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Tang, Xuri. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering* 24(5). 649–676.

Thomason, Sarah. 2001. *Language contact: An introduction*. Edinburgh: Edinburgh University Press.

Traugott, C. Elizabeth & B. Richard Dasher. 2002. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Urban, Matthias. 2012. *Analyzibility and semantic associations in referring expressions: A study in comparative lexicology*. Leiden University PhD dissertation.

Vanhove, Martine. 2008. Semantic associations between sensory modalities, prehension and mental perceptions: A crosslinguistic perspective. In Martine Vanhove (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 341–370. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/slcs.106.17van.

Viberg, Åke. 1983. The verbs of perception: A typological study. *Linguistics* 2. 123–162.

Viberg, Åke. 2001. The verbs of perception. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook*, 1294–1309. Berlin & New York: Mouton De Gruyter. https://doi.org/10.1515/9783110171549.2.11.1294.

Wälchli, Bernhard. 2010. Similarity semantics and building probabilistic semantic maps from parallel texts. *Linguistic Discovery* 8(1). 331–371.

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. In Maria Koptjevskaja-Tamm & Martine Vanhove (eds.), *New directions in lexical typology. Special issue of linguistics*, Vol. 50, 671–710.

Wälchli, Bernhard. 2016. Non-specific, specific and obscured perception verbs in Baltic languages. *Baltic Linguistics* 7. 53–135.

Wichmann, Søren & Eric Holman. 2009. *Assessing temporal stability for linguistic typological features*. München: LINCOM Europa.

Winter, Bodo, Marcus Perlman & Asifa Majid. 2018. Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition* 179. 213–220.

Wnuk, Ewelina & Asifa Majid. 2014. Revisiting the limits of language: The odor lexicon of Maniq. *Cognition* 131. 125–138.

Youn, Hyejin, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft & Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences of the United States of America* 113. 1766–1771.