Volker Gast, Maria Koptjevskaja-Tamm

The areal factor in lexical typology

Some evidence from lexical databases

Abstract: Our study aims to explore how much information about areal patterns of colexification we can gain from lexical databases such as CLICS and ASJP. We adopt a bottom-up (rather than hypothesis-driven) approach, identifying areal patterns in three steps: (i) determine spatial autocorrelations in the data, (ii) identify clusters as candidates for convergence areas and (iii) test the clusters resulting from the second step controlling for genealogical relatedness. Moreover, we identify a (genealogical) diversity index for each cluster. This approach yields promising results, which we regard as a proof of concept, but we also point out some drawbacks of the use of major lexical databases.

Keywords: areality, colexification, lexical database, lexical typology

1 Introduction

1.1 Lexical typology and areal linguistics

The lexicon is arguably one of the most difficult domains for cross-linguistic and typological generalizations. It is much more loosely structured than, for instance, sound or tense systems and its structure is hardly reflected in linguistic form. Attempts at identifying principles of lexical organization, for example, with sense relations in the structuralist tradition or through prototypes and family resemblances therefore mostly rely on diagnostic tests and intuitions, even in the analysis of individual languages (e.g., Lipka 1992; Cruse 1986; Kleiber 1990; Geeraerts 2010). Comparing the lexicons of different languages obviously poses an even greater challenge and some would say that such comparison is not even possible,

Volker Gast, Friedrich Schiller University of Jena, Department of English and American Studies, Ernst-Abbe-Platz 8, 07743 Jena, Germany, volker.gast@uni-jena.de

Maria Koptjevskaja-Tamm, Stockholm University, Department of Linguistics, 106 91 Stockholm, Sweden, tamm@ling.su.se

Open Access. © 2018 Volker Gast, Maria Koptjevskaja-Tamm, published by De Gruyter. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. <https://doi.org/10.1515/9783110607963-003>

as word meanings are only defined relative to the systems they form part of (cf. Evans 2011: §1.3 for some discussion). Despite such difficulties, some non-trivial generalizations have been formulated about the organization of lexicons. Most of the insights concern individual domains of meaning, such as color terms (Berlin and Kay 1969 and follow-up work), verbs of motion (Talmy 1985 and follow-up work), kinship terms (e.g., Nerlove and Romney 1967; Dahl and Koptjevskaja-Tamm 2001), body parts (e.g., Majid, Enfield and Van Staden 2006; Brown 2013a, 2013b), perception (e.g., Viberg 1984; Evans and Wilkins 2000; Vanhove 2008a), temperature terms (Koptjevskaja-Tamm 2015) and verbs of aqua-motion (Koptjevskaja-Tamm, Divjak and Rakhilina 2010), to name just a selection of the domains that have figured prominently in this branch of typology. The findings of lexical-typological studies often take the form of “holistic characterizations” (e.g., “language x is a satellite-framed language” in terms of Talmy 1985) or implicational relations (e.g., “if a language has three basic color terms, one of them is ‘red’”; Berlin and Kay’s 1969 stage II).

Broader generalizations about possible systems of lexical domains have often been represented using the semantic map methodology, particularly in the domain of function words such as indefinites (e.g., Haspelmath 1997; van der Auwera and Van Alsenoy 2011) and impersonal pronouns (e.g., van der Auwera, Gast and Vanderbiesen 2012; Gast and van der Auwera 2013) but also in other, “more lexical” domains (cf. Koptjevskaja-Tamm, Rakhilina and Vanhove 2015 for an overview of implicational and probabilistic semantic maps in lexical typology; Rakhilina and Reznikova 2016 on the use of semantic maps in the frame-based approach to lexical typology). Semantic maps can be regarded as networks (technically, graphs; cf. Gast and van der Auwera 2013) representing patterns of multifunctionality manifested by semantically/functionally “comparable” linguistic expressions (e.g., morphemes, words, constructions) of particular languages, where the main guiding principle is the “contiguity/connectivity requirement”. More specifically, functions (uses, meanings, contexts) that are often associated with one and the same linguistic expression, represented as “nodes” (or “vertices”) in the graph, are connected by “edges” or they cover a contiguous region on a semantic map.

While semantic maps may be used to represent the range of meanings associated with any kind of linguistic expression, the term “colexification”, coined by François (2008) and since then widely adopted, specifically targets the expression of two (supposedly different) concepts with one word. Its major advantage is that it is non-committal with respect to why two concepts are expressed with the same linguistic element. Polysemy is certainly the most interesting case in cross-linguistic studies but colexification may also emerge for other reasons,

such as the reanalysis of scope relations (cf. the case of impersonal pronouns discussed in Gast and van der Auwera 2013). Colexification is also less rigorous when it comes to the contiguity/connectivity requirement insofar as linking elements on semantic maps may sometimes be lost in historical developments, leaving us with two words that do not cover a contiguous space on the map but are nevertheless colexified (cf. van der Auwera and Temürçü 2006; van der Auwera and Van Alsenoy 2013).

The concept of colexification can therefore be fruitfully used for the purposes of areal linguistics and areal typology (e.g., Urban 2012; Koptjevskaja-Tamm and Liljegren 2017). As is well-known, languages do not only borrow “matter” (e.g., loan words), but also “patterns” (Matras and Sakel 2004). A particularly common type of pattern transfer has been called “polysemy copying” (Heine and Kuteva 2003, 2005) or “distributional assimilation” (Gast and van der Auwera 2012): As a result of “interlingual identification”, linguistic elements from contact languages may assimilate their distributions. The results of such transfer can further be expected to be reflected in areal patterns of distribution. In fact, some colexification patterns are cross-linguistically frequent, such as the colexification of ‘finger’ and ‘toe’. Others show a genetically and/or areally restricted distribution, such as the colexification of ‘eat’ and ‘drink’ in many Papuan and Australian languages (Aikhenvald 2009), as well as in some other languages of the world (for further examples, see Vanhove 2008b; Urban 2012; Juvonen and Koptjevskaja-Tamm 2016). Still others are very local or even language-specific, such as ‘beef’ expressed as ‘big meat’ in some of the languages of Hindukush – the mountainous region comprising northern Pakistan, northeastern Afghanistan and the northern-most part of Indian Kashmir (Koptjevskaja-Tamm and Liljegren 2017). It has been shown, even before the notion of colexification came into use, that there are clear areal patterns in, for example, the distribution of languages distinguishing specific color terms (cf. Kay and Maffi 2013a, 2013b) and the body parts ‘arm’, ‘finger’ and ‘hand’ (cf. Brown 2013 a, 2013b). There is therefore a challenge to both identify such areal patterns and to provide explanations for them.

A systematic cross-linguistic study of colexification patterns needs a lot of data. Most of the relevant investigations have been based on retrieving dictionary data and/or elicitation, both of which are very time-consuming, which explains why this research has so far been relatively restricted; it would require developing new and more efficient methods of data collection. For instance, the range of data for the comparative study of lexical-semantic patterns can be broadened by relying on parallel texts. This approach, pursued by Östling (2016), among others, represents a highly promising direction for future research, specifically when the range of texts and registers can be broadened. At present, most studies using

“massively parallel corpora” (Cysouw and Wälchli 2007) rely on the Bible and, consequently, on a rather specific (written) register, which introduces a certain bias in terms of both the topics covered and the vocabulary used. For instance, the richness of kinship terminology found in some parts of the world is obviously not retrievable from Bible texts. Still, parallel texts clearly provide a useful resource for the cross-linguistic study of lexical semantics that should be explored further.

The recent development of typological resources with a global coverage of data has made the study of a broader range of colexification patterns possible. For example, the *Database of cross-linguistic colexifications* (CLICS version 1.0; see List et al. 2014) provides an interface for the visualization and graphical inspection of colexification patterns in a sample of 221 languages. Such databases, and their potential (as well as limitations), constitute the main topic of this contribution. In addition to the CLICS database, which has been designed specifically for the study of colexification, we will use data from another lexical database for comparison, the database of the *Automated similarity judgment programme* (ASJP; see Wichmann, Holman and Brown 2016).

1.2 Main questions addressed in this study

The main question addressed in this contribution can be formulated as follows: what can we learn from cross-linguistic lexical databases such as CLICS and ASJP to gain a better understanding of global patterns of lexical organization and their areal distributions? As outlined in Koptjevskaja-Tamm and Liljégren (2017), at least the following groups of lexico-semantic phenomena may serve as indicators of areality:

- lexico-semantic parallels – shared colexification patterns and/or shared lexico-constructional patterns/calques, such as the colexification of ‘fruit’ and ‘child’, or ‘fruit’ being expressed as ‘child of tree’ across many West African languages, both cases involving a semantic association between ‘child’ and ‘fruit’;
- shared formulaic expressions, such as the farewell expressions *au revoir* (French), *auf Wiedersehen* (German), *på återseende* (Swedish), *do svidanija* (Russian) and *näkemiin* (Finnish), which follow the same model across a number of European languages;
- area-specific lexicalizations and a shared or similar-looking internal organization of certain semantic domains, such as a highly specialized vocabulary describing dairy practices and dairy products across the languages of the

Greater Hindukush or the different areally defined patterns in the systems of phasal adverbials ‘still’, ‘no longer’, ‘not yet’ and ‘already’ in European languages (see van der Auwera 1998a).

Since the existing cross-linguistic lexical databases such as CLICS and ASJP are restricted in their data coverage, our main question will specifically target shared colexification patterns. In other words, we want to determine what information we can gain about colexifications from the databases. “Gaining information from” the databases basically comprises two complementary aspects: we can test existing claims and hypotheses about colexification patterns and we can explore new patterns that have not been identified so far. In this study, we pursue the latter approach.

We would like to make it clear from the outset that our approach is entirely unprejudiced. Neither of us was involved in either project (CLICS or ASJP) and we are using the data in a way in which anyone else could use them as well. The study has an exploratory character and a certain methodological focus. The discussion will therefore contain some critical remarks, which are not intended to call into question the merits of open-access resources like CLICS and ASJP, or the value of large-scale databases in general. Moreover, it should be borne in mind that we are using the data from these databases in ways that were not intended or anticipated by the creators. Any kind of limitation pointed out is thus not to be seen as criticism of the databases.

Following some remarks on the data and methodology in Section 2, we present some rather general observations about the scope and limits of the databases in Section 3. In Section 4, we present a bottom-up approach to the identification of areal colexification patterns, as well as some results obtained in this way. Section 5 contains some conclusions.

2 Remarks on the data and methodology

2.1 Lexical databases used in the study

CLICS is an online database of colexifications (called “synchronic lexical associations” on the homepage)¹ with data from 221 language varieties of the world. It draws on four types of (digital) resources (see Mayer et al. 2014):

¹ <http://clics.lingpy.org/main.php> (accessed 1 June 2017).

- the *Intercontinental dictionary series* (IDS; Key and Comrie 2015), which emerged from a long-term project that started in the 1980s and implied the compilation of word lists (by experts) for the expression of 1,310 concepts in 233 languages – for CLICS, a reduced set of (cleaned-up) data from 178 languages was used;
- the *World loan word database* (WOLD; Haspelmath and Tadmor 2009), which contains a large amount of vocabulary (1,000–2,000 items) from 41 languages, compiled by experts – data from 33 of the languages were included in CLICS;
- the online dictionary LOGOS, from which the authors extracted data for four languages not represented in IDS or WOLD;
- the Språkbanken research unit at the University of Gothenburg provides ten word lists of South Asian and Himalayan languages,² six of which were used for CLICS.

The database has been specifically developed for extracting information on colexifications across languages. As its creators explain, “[i]t is designed to serve as a data source for work in lexical typology, diachronic semantics, and research in cognitive science that focuses on natural language semantics from the viewpoint of cross-linguistic diversity. Furthermore, CLICS can be used as a helpful tool to assess the plausibility of semantic connections between possible cognates in the establishment of genetic relations between languages” (CLICS website).³

Information on colexifications can be extracted in various ways. Using the query interface, it is possible to find out whether two specific concepts are linked in the language varieties in CLICS and to determine how many links to different concepts are reported for a specific concept. Users can also browse the concept networks that have been extracted from the data by the database creators and download parts of the data for large-scale quantitative investigations, which is what was done for the purposes of this study.

Given the heterogeneity of sources, there are no consistent, language-independent definitions of the concepts represented in the database.⁴ As far as areal

² <https://spraakbanken.gu.se/eng/research/digital-areal-linguistics/word-lists> (accessed 11 April 2017).

³ <http://clics.lingpy.org/main.php> (accessed 11 April 2017).

⁴ We will leave the discussion of labelling and (the absence of) definitions for the entries in the different lexical databases for a future occasion.

coverage is concerned, the authors make the following disclaimer (see CLICS website):⁵

Coverage of the world's languages in both IDS and WOLD is biased towards certain regions of the world. In the case of IDS, South American languages and languages of the Caucasus are overrepresented. In the case of WOLD, languages of Europe figure particularly prominently. Since it is possible and even expectable that certain polysemies in the lexicon are frequent or even restricted to certain areas of the world, we advise researchers interested in cross-linguistic diversity to take appropriate measures to rule out unwarranted generalizations due to areal effects.

The database of ASJP was created for the purposes of comparative historical linguistics, as a means for evaluating the similarity of words from different languages with the same meaning and, ultimately, for classifying languages computationally on the basis of the observed lexical similarities. It grew out of a collaboration of “25 professional linguists and other interested parties working as volunteer transcribers and/or extending aid to the project in other ways” (Wikipedia).⁶ The database (version 17, April 2017) provides information on the expression of mainly 40 concepts from the (100-item) Swadesh list in 4,664 languages and was collected with guidelines administered to the contributors.⁷ As the ASJP data is based on the Swadesh list, there is, as far as we can tell, no language-independent definition of the concepts. The English words thus function as a *terminum comparationis*. Given the sheer quantity of languages represented in the data, there is, inevitably, a certain amount of heterogeneity in it but it should be borne in mind that the project was coordinated by specialists (especially Cecil H. Brown, Søren Wichmann and Eric W. Holman) and the database continues to be curated.

2.2 Extracting and processing the data

As both CLICS and ASJP associate concepts with words, we can (automatically) identify colexifications through a simple comparison of words and their associated meanings. For the CLICS data, we used the file “links.csv” from the “official” download link.⁸ It contains 32,536 links (colexification types), each of them for a

⁵ <http://clics.lingpy.org/faq.php#data1> (accessed 4 April, 2017).

⁶ https://en.wikipedia.org/wiki/Automated_Similarity_Judgment_Program#The_ASJP_Consortium (accessed 4 April 2017).

⁷ <http://asjp.cild.org/static/Guidelines.pdf> (accessed 1 June 2017).

⁸ <http://clics.lingpy.org/download.php> (accessed 1 June 2017).

certain number of languages. Altogether, it provides information about 91,673 instances of colexification in 221 languages.

The file with the ASJP data⁹ contains 7,221 word lists. Nearly all consist of a 40-item subset of the 100-item Swadesh list found to constitute the diachronically most stable items (Holman et al. 2008). Most of these 40-item lists are incomplete. In addition, there are a little over 300 full 100-item lists (also generally not complete), most of which hark back to the beginning of the project, as described in Brown et al. (2008), before the selection of stable items was made. Often, more than one word is listed for a given concept. We mainly made use of the data from the 40-item lists because of the scarcity of data for the 60 items not on that list (words for ‘feather’ and ‘bark’ constitute exceptions, see Section 4).

As we wanted to keep things comparable and consistent, specifically with respect to geographical data, we mapped all the varieties to Glottolog codes. This led to some loss of information. The 7,221 word lists of the ASJP file were associated with 4,675 ISO 639-3 codes. We lumped the data from different word lists with the same ISO 639-3 code. This may have led to a certain loss of accuracy in the data, as we may have mixed data from distinct language varieties. Some data was also lost because some mappings from ISO 639-3 codes to Glottolog codes were missing. In this way, we obtained information about 690 colexification types in 4,554 languages. The CLICS data was reduced for the same reason, leaving us with 4,064 colexification types from 196 languages.

In order to identify the areal distribution of colexification patterns, we also need negative evidence, i.e., information about the absence of a colexification pattern. As CLICS focuses specifically on colexification, not differentiation, this information is not *per se* included in the database. We can, however, retrieve a certain amount of negative evidence from the data, as the colexification patterns associate pairs of concepts with sets of forms. The ASJP data provides information about specific form-meaning pairings. Whenever we had information about the form encoding a given concept and when the two forms corresponding to the members of a colexification pair were different, we assumed that the pair of concepts in question was differentiated in the given language.

After adding cases of differentiation to our data, our (enriched) ASJP database contained 2,060,856 data points (a subset of the 4,554 languages multiplied by 690 colexification patterns) and the CLICS-database contained 92,805 data points (a subset of the 196 languages x 4,064 colexification patterns).¹⁰ Each data

⁹ The file “dataset.tab” from <http://asjp.clld.org/download> (accessed 1 June 2017).

¹⁰ The data frames can be downloaded at <http://www.uni-jena.de/~mu65qev/data> in csv-files (accessed 4 April 2018).

point is a quadruple of two concepts, a (Glottolog) language code and a value of “t” (true/colexified) or “f” (false/differentiated): <‘arm’, ‘hand’, russ1263, t>, for instance, says that ‘arm’ and ‘hand’ are colexified in Russian. The Glottolog code can, moreover, be mapped to geographical and genealogical information.

As far as the types of concepts represented in the databases are concerned, the ASJP data covers a certain range of nominal concepts from the domains of body parts (e.g., ‘eye’, ‘ear’, ‘nose’), animals and plants (e.g., ‘louse’, ‘dog’, ‘fish’) and nature (e.g., ‘sun’, ‘star’, ‘water’), as well as ‘person’ and ‘name’. The database furthermore contains data on five verbs (i.e., ‘drink’, ‘die’, ‘see’, ‘hear’ and ‘come’), two adjectives (i.e., ‘new’ and ‘full’), the numerals ‘one’ and ‘two’ and the pronouns ‘I’, ‘we’ and ‘you’. As pointed out above, we have also made use of some of the other material in the data, which covers a few additional basic concepts such as ‘feather’, ‘hair’ and ‘bark’. CLICS contains an overall much more varied set of concepts, including a high number of kinship terms, some of them highly specific (see the examples discussed in Section 4.5).

3 What the databases can(not) do for us

Our data indicates, for pairs of concepts, whether they are colexified or differentiated in a given language or variety. This is obviously a simplification. For example, many languages may have different words for ‘day’ and ‘night’ and still use ‘day’ as a cover term. If two languages are shown to colexify a given concept, this does not necessarily mean that they do not have different words for the individual concepts as well. ‘Arm’ and ‘hand’ provide a well-known problem in this respect, as many languages (e.g., Slavic ones) have a cover term for both body parts while also having more specific terms for each part. For example, Russian *ruka* covers both the arm and the hand but there is a more specific word for ‘hand’ as well, i.e., *kist’*, which is used much more rarely and only in specific contexts. There is thus both colexification and differentiation. We assume that the data in the databases, by and large, represent the “preferred” patterns of colexification or differentiation in the languages in question but it should be clear that the distinction between the two cases is an idealization (note that this does not only apply to the data used for this study but also to *World atlas of language structures* maps like those of Brown 2013a, 2013b).

Moreover, we should bear in mind that colexification is a very general term, as pointed out above, and that, unlike in the semantic map methodology, it does not necessarily indicate immediate or exclusive relatedness of two concepts. This

point can be illustrated with patterns that emerged from an inspection of extremely rare cases of colexification. For instance, the colexification pair <‘daughter-in-law (of a woman)’, ‘father-in-law (of a man)’>, found in only two languages, seems quite remarkable. Closer inspection of the data shows, however, that both cases of this type – one from WOLD, one from the IDS – are instances of extremely general kinship terms, rather than unusual cases of polysemy. Swahili (swah1253) *mkwe*, for example, roughly means ‘in-law’ and is thus not only used for the two concepts mentioned above but also for many other in-law relationships. Similarly, Polci (polc1243) *kwam* is used for all types of in-law relations holding between contiguous generations, according to the information given in the IDS. What this shows is that we should always look at broader patterns of multifunctionality before jumping to conclusions about pairwise meaning associations, specifically when we explore the data in a bottom-up fashion as intended in the present study.

Finally, we will briefly address the questions of areal/lexical typology for which neither of the two databases can give us any information (and for which neither of them was made). CLICS and ASJP do not allow us to extract any information on shared lexico-constructional patterns or formulaic expressions. They only associate linguistic forms with concepts, so we can only identify pairs of concepts that are expressed by exactly identical linguistic forms, with the linguistic forms (lexemes) not being further analyzed. This identity of lexemes in synchrony corresponds to François’s (2008: 171) notion of “strict colexification”. Colexification and semantic associations can, however, be understood more broadly – both diachronically, as two concepts being expressed by the same lexeme at different periods in its history, and panchronically, as linked to each other by derivation, composition or other constructions. To give one example: while ‘eye’ and ‘eyelid’ in Khasi (khas1269) and in Kumyk (kumy1244) are expressed by the same form (*ñiuhmat* in Khasi, *köz* in Kumyk), there are plenty of languages, including English, where ‘eyelid’ is expressed by a compound involving ‘eye’. Such cases of “loose colexification” (François 2008: 171), involving various kinds of semantic shifts are, arguably, much more difficult to represent and identify in databases. A laudable attempt to systematize cross-linguistically recurrent patterns of loose colexification is represented in the *Catalogue of semantic shifts in the languages of the world* at the Institute of Linguistics of the Russian Academy of Sciences in Moscow. It is a database that currently contains more than 3,000 semantic shifts found in 319 languages (see Zalizniak 2008; Zalizniak et al. 2012) but the organization of the database does not make it possible at present to draw any (statistical) conclusions on the distribution of these patterns across the languages of the world. In this paper, we will therefore only use data from CLICS and

ASJP, restricting ourselves to cases where the databases associate two different concepts with exactly the same lexeme.

4 Areal clusters of colexification patterns

One of our intentions in using the data from CLICS and ASJP was to detect new patterns and generate new hypotheses about areal clusters of colexification patterns in a bottom-up approach. Our starting point is a map like the one in Figure 1, which shows the distribution of languages colexifying and differentiating ‘feather’ and ‘hair’ (see also Urban 2012 for this pattern) in the ASJP data. A black square stands for colexification, a red/empty circle for differentiation. Note that we are using this example because the relative small number of data points allows us to illustrate the method more easily. The noun ‘feather’ is contained in the Swadesh list but it is not among the 40 core items used for ASJP, so the number of data points is relatively small, and certain regions are heavily underrepresented (e.g., Africa). The full dataset of the (40-item) ASJP project has worldwide coverage (see the remarks made in Section 2.2).

Fig. 1: Colexification (black squares) versus differentiation (red/empty circles) of ‘feather’ and ‘hair’ in the ASJP data

In order to identify areal clusters of colexifications, we proceeded in three steps:

- identification of areally biased colexification patterns;

- identification of areal clusters in the biased patterns as candidates for “cluster areas”, i.e., areas that are characterized by a given colexification pattern;
- testing the cluster areas controlling for genealogical relatedness.

4.1 Identifying areally biased colexification patterns

In order to identify areally biased colexification patterns, we used the Join Count statistic (Cliff and Ord 1981), which is commonly applied to test for spatial autocorrelations in binary data.¹¹ Assume that there is a grid of, say, eight by eight cells and all of the cells are either black or white. If black and white cells are distributed over that grid as on a chessboard, no cell has a (horizontal or vertical) neighbor of the same color – there are no (same color) “joins” at all. In this case, there is a “negative autocorrelation”. If, by contrast, all the black cells are on the left side of the grid and all the white cells on the right side, there are many joins between cells of the same color – to be precise, 52 black-black joins and 52 white-white joins, as against eight black-white joins (in the middle of the board). In this case, there is a positive spatial autocorrelation. The Join Count statistic compares the observed number of same-color joins to the number expected on the basis of a random distribution of colors. It is defined as the observed frequency minus the expected frequency of identical joins, divided by the standard deviation of the expected frequency. The statistic indicates a direction of the correlation (positive or negative) and we can calculate a p-value for it (i.e., a value that indicates the probability of finding the distribution in question under the hypothesis that the colors are distributed randomly).

In order to apply the Join Count test to linguistic data,¹² we have to transform our data points into a grid or network of “neighbors”. This implies that we have to decide what languages count as neighbors. As we are dealing with questions of language contact, “being a neighbor” should mean “potentially being in contact with each other”. Obviously, it is hard to generalize over distances making language contact (im)possible, as this varies with the local habitat (e.g., sea versus mountains). Moreover, the distance has to be adapted to the density of data points, i.e., we need larger distances for sparser data. We experimented with dis-

¹¹ Note that spatial autocorrelation is a common problem in many natural sciences such as biology and geography. As in linguistic typology, data points are often influenced by neighboring data points.

¹² We used the function ‘joincount()’ from the R package *spdep* (see Bivand, Hauke and Kossowski 2013; Bivand and Piras 2015).

tances between 500 km and 2,000 km and – comparing the results with the clustering applied at the second stage – found that a neighbor distance of 2,500 km was a reasonable choice for the CLICS data and a distance of 1,000 km for the ASJP data. On the basis of these distances, the maps were transformed into networks of neighbors, as illustrated in Figure 2 for Mesoamerica.¹³ Note that only those languages are shown for which we have information on colexification or differentiation for the pair <‘feather’, ‘hair’>.

Fig. 2: Neighbor network for languages of Mesoamerica and Central America

4.2 Identifying clusters

For those colexification patterns which showed an areal bias (positive autocorrelation) according to the Join Count test (at a p-value < 0.05), we determined clusters using hierarchical cluster analysis on the basis of a (geographical) distance matrix.¹⁴ We distinguished three types of clusters, defined on the basis of their (maximal) distance to a neighboring cluster, i.e., “micro-clusters” (2,000 km), “meso-clusters” (4,000 km) and “macro-clusters” (6,000 km).

¹³ For the identification of neighbors, we used the function ‘dnearneigh()’ of the spdep package for R (Bivand, Hauke and Kossowski 2013, Bivand and Piras 2015).

¹⁴ As we are dealing with geographical data, we used the function ‘distm()’ from the ‘geosphere’-package for R (Hijmans 2016), rather than the native ‘dist()’-function of R. By default, the function uses the Haversine Great Circle method (fun=distHaversine). For the cluster analysis, we used the native ‘hclust()’-function of R, with the linkage method ‘complete’ (for “round” clusters).

Fig. 3: Dendrogram for the colexification pair <'feather', 'hair'> (meso-clusters)

The results of such cluster analyses are commonly represented in the form of dendrograms, as shown in Figure 3. The dendrogram shows the grouping of the languages exhibiting colexification of ‘feather’ and ‘hair’, represented by their Glottolog codes at the bottom, into meso-clusters. The y-axis of the diagram (“height”) indicates distances between nodes. The splits between groups of languages in Figure 3 are based on the maximum distances between pairs of elements from sisters in the tree. As the figure shows, the clustering of the relevant languages into meso-clusters, the cut-off distance is (by definition) 4,000 km. The (red) boxes at the bottom illustrate the clusters emerging from this cut-off point (the upper edge of any box is located at that “height”, i.e., the distance). Note that the critical distances are upper boundaries, so different cut-off points may deliver the same clusters. The largest distance between pairs of elements from a cluster is thus 2,000 km, 4,000 km or 6,000km, depending on the type of cluster.¹⁵ The geographical locations of the five clusters emerging from Figure 3 are shown in Figure 4. The dotted circles are positioned around the geographical center of a cluster¹⁶ and their radius corresponds to the largest distance of any one cluster language to the center. The areas indicated by the circles enclosing a cluster will be referred to as the “cluster area” in each case. Each cluster area has a numerical identifier for the purpose of data processing and textual reference. The cluster areas determined on this basis will serve as hypotheses for linguistic contact areas characterized by the relevant colexification patterns (potentially among other features, of course).

15 We also applied model-based clustering, where the cut-off points between clusters are determined on the basis of specific diagnostics of the relevant models, such as the Bayesian Information Criterion, as in the function ‘mclust()’ of the R package mclust (Fraley and Raftery 2002; Fraley et al. 2012). This approach yielded highly heterogeneous results in terms of scaling, however, so comparability of clusters was compromised.

16 The centers and radii were calculated by mapping the geographical coordinates to a Cartesian coordinate system, determining the relevant data and mapping them back to geographical coordinates.

Fig. 4: Hypothesized meso-clusters for the colexification pattern <feather, 'hair'>

4.3 Testing and analyzing clusters

Obviously, a (hypothesized) cluster area may have emerged for reasons that are independent of the geographical location of the languages concerned, in particular, as it may just reflect the genealogical relatedness of neighboring languages with inherited colexification patterns. Given that we are most interested in clusters that are areally conditioned (i.e., clusters that emerged through language contact or some other geographical factor), we determined the influence of the independent variable “membership of a language L to a given cluster area” on the dependent variable “presence of a colexification pattern in L ”, controlling for genealogical relationships. We fitted (for each colexification pattern and cluster size, i.e., micro, meso and macro) a mixed effects model, treating the (highest-level) language family as a random effect.¹⁷ The data was pre-filtered and we only ran regression analyses for colexification patterns with at least 20 TRUE cases for

¹⁷ We used Bayesian logistic regression as implemented in the function ‘MCMCglmm()’ of the MCMCglmm-package for R (Hadfield 2010), with weakly informative prior assumptions (Gelman et al. 2008). The main reason for this choice was the structure of the data, with many systematic cases of complete separation. We compared the results from Bayesian regression with frequentist (mixed-effects) methods and inspected the results for “sanity”, using binned residual plots (Gelman and Su 2016). So-called Gelman priors assume a Cauchy distribution with center 0 and scale 2.5. The MCMCglmm package offers a function ‘gelman.priors()’, which we used for this purpose. The number of iterations was set to 130,000, with 30,000 burnin iterations and a thinning interval of 10.

the ASJP data and 15 TRUE cases for the CLICS data. In each model, we included only data points from families such that one member of the family either exhibited the colexification pattern in question or was a member of one of the (hypothesized) cluster areas. The (Bayesian) regression model identifies a posterior mean-value for each cluster area (“pm”, a rough indicator of effect size) and a p-value (“p.pm”) showing the probability that the posterior mean is not higher than zero (and that membership to the cluster area thus has no effect). In the following discussion, we will mainly focus on the p-value. For the <‘feather’, ‘hair’>-pattern, the analysis delivered the values shown in Table 1.

Tab. 1: Results of regression analysis for <‘feather’, ‘hair’> (meso-clusters)

Colexification	Cluster	Long	Lat	Radius	pm	p.pm
<feather, hair>	1	-91.2	16.0	937	1.19	0.16
<feather, hair>	2	-64.5	-2.8	2,043	0.89	0.27
<feather, hair>	3	106.2	16.5	544	2.75	< 0.01
<feather, hair>	4	8.0	10.7	1,380	4.29	< 0.001
<feather, hair>	5	153.2	-3.3	1,470	1.54	0.054

To provide a somewhat better idea of the internal make-up of clusters, Figure 5 shows cluster areas 1¹⁸ and 2 in more detail.

18 The prevalence of the <‘feather’, ‘hair’> colexification in Mesoamerica has been noted earlier, among others by Smith-Stark (1994), who tests but also discards its applicability as a Mesoamerican areal trait (“lexical calque”).

Fig. 5: Cluster areas 1 and 2 for the colexification pair <‘feather’, ‘hair’>

Membership to a cluster area may be a significant predictor even when the cluster is genealogically entirely homogeneous if there is a high number of non-cluster-members from the same family which do not exhibit the colexification pattern in question. The regression analysis only excludes that a cluster primarily contains languages from families whose members tend to exhibit the pattern independently of membership to the cluster. We therefore determined an indicator of the genealogical diversity within each cluster as well. We used Shannon’s diversity index (Simpson 1949) for this purpose, with a correction factor for small sample sizes.¹⁹ We determined this index only for those languages of a cluster area that showed the colexification pattern in question. In the <‘feather’, ‘hair’> clusters in Figure 4, we find different degrees of homogeneity, as Table 2 shows.

Tab. 2: Diversity indices for <‘feather’, ‘hair’> (meso-clusters) (ASJP)

Cluster area	Diversity index
1	1.01
2	1.38
3	0
4	0.77
5	0.44

¹⁹ With the R DiversitySampler package’s ‘Hs()’-function (Lau 2012), with the option “corr=T”.

As Table 2 shows, cluster area 3 in Southeast Asia is totally homogeneous and it actually provides a nice illustration that significant clusters need not be genealogically diverse. All languages showing the colexification pattern in question are Austroasiatic.²⁰ Given that, at the same time, all Austroasiatic languages showing the colexification in question are members of this cluster, membership to this cluster correlates positively with the colexification pattern in question. The ratio of languages to families is one, however, and the diversity index is therefore zero. By contrast, cluster area 2 (in South America) is the most heterogeneous one, as the six languages showing colexification of ‘feather’ and ‘hair’ belong to six different families.²¹ Cluster area 5 (in Melanesia) exhibits a relatively low degree of diversity (six languages from two families). This is illustrated in Figure 6, where families are identified with their Glottolog codes.²² (It should, of course, be borne in mind that the density of data points is very low for this colexification pair, particularly in Melanesia, and that the data is only used for illustrative purposes at this point.)

Fig. 6: Cluster areas 2 and 5 for <‘feather’, ‘hair’>

20 Note that the degree of genealogical diversity of a cluster should ideally be measured not relative to top-level families but to lower-level branches. The Austroasiatic languages in cluster 4 actually belong to four different branches of that family: Bahnar (bahn1262) and Jeh (jehh1245): Bahnaric, Kuy (kuyy1240); Katuric, Ksingmul (puoc1238); Khmuic, Chut (chut1247); Vietic.

21 Viz., Jodi-Saliban (jodi1234), Kakua-Nukak (kaku1242), Nadahup (nada1235), Tupian (tupi1275), Pano-Tacanan (pano1259) and Aikanā (aika1237).

22 Viz., Austronesian (aust1307) and Nuclear Trans New Guinea (nucl1709).

On the basis of the procedure described above, we can now identify clusters in a bottom-up fashion, generating some statistics that will allow us to estimate how interesting they are from the point of view of lexical typology and in an areal perspective. Obviously, the best evidence for contact-induced clusters is provided by examples with a high degree of distinctiveness (which means that many languages in the cluster exhibit the colexification in question and few non-cluster languages have it) and a high degree of genealogical diversity. We will now take a closer look at the clusters delivered by the data from ASJP and CLICS (Sections 4.4 and 4.5).

4.4 Clusters emerging from the ASJP data

The procedure described in Sections 4.1 to 4.3 brought to light 23 colexification patterns showing a significant positive spatial autocorrelation and 120 cluster areas that turned out to be significant predictors for a given colexification in the ASJP data, controlling for genealogical relatedness. The 23 colexification pairs are shown in (1), the 36 meso-clusters in Figure 7.

- (1) <'I', 'fish'>; <'I', 'we'>; <'I', 'you'>; <'bark', 'skin'>; <'blood', 'die'>; <'bone', 'die'>; <'come', 'dog'>; <'die', 'eye'>; <'drink', 'water'>; <'ear', 'hear'>; <'ear', 'leaf'>; <'ear', 'name'>; <'eye', 'name'>; <'feather', 'hair'>; <'fire', 'tree'>; <'horn', 'knee'>; <'horn', 'tooth'>; <'liver', 'two'>; <'louse', 'we'>; <'man', 'person'>; <'mountain', 'stone'>; <'name', 'tooth'>; <'see', 'we'>

Fig. 7: All meso-clusters of the ASJP data

We can now apply further filters in order to identify the most interesting cases of association between a colexification pattern and a cluster area. We can use the diversity index for that purpose. Table 3 shows the top ten of the clusters, ordered by the diversity index (all cluster areas are significant predictors at a 0.05 level, according to the regression model).²³ It also indicates (in this order) the position and radius of each cluster, the numbers of languages and families, the diversity index, the posterior mean and the p-value for the latter value. A list of all significant cluster areas is provided online.²⁴ The ten clusters in Table 3 correspond to four colexification patterns: <‘fire’, ‘tree’>, <‘mountain’, ‘stone’>, <‘ear’, ‘leaf’> and <‘bark’, ‘skin’>. We will now take a closer look at each of these patterns.

Tab. 3: Top ten clusters emerging from the ASJP data

Colexification	Size	No	Long	Lat	Radius	No lng	No fam	Div	pm	p.pm
<fire, tree>	meso	1	142.9	-7.8	2557	53	16	2.3	2.87	< 0.001
<mountain, stone>	macro	1	16.5	6.9	4,320	69	16	2.3	2.90	< 0.001
<ear, leaf>	macro	1	29.8	7.8	3,876	39	14	2.2	2.98	< 0.001
<ear, leaf>	meso	1	30.1	8.6	2,670	38	13	2.1	3.24	< 0.001
<bark, skin>	macro	1	-65.3	-7.2	2,466	14	11	1.9	1.86	0.033
<ear, leaf>	micro	2	35.0	4.9	921	30	10	1.9	3.04	< 0.001
<mountain, stone>	meso	2	133.1	-19.0	1,967	27	10	1.9	1.15	0.027
<mountain, stone>	meso	4	-64.8	-8.4	1,533	9	8	1.6	1.42	0.030
<fire, tree>	micro	6	130.6	-12.1	822	8	7	1.5	1.76	0.019
<mountain, stone>	meso	3	5.9	11.1	1,808	45	7	1.5	1.92	< 0.001

²³ Note that, in some cases, the same cluster areas were identified for different cluster sizes (e.g., meso and macro), as the cut-off points for clusters are maximal distances. In such cases, we obviously regarded the cluster as being of the smallest category.

²⁴ See <http://www.uni-jena.de/~mu65qev/data/colex-tables/ASJP-clusters.htm> (accessed 4 April 2018).

4.4.1 <‘Fire’, ‘tree’>

<‘Fire’, ‘tree’> is a well-known colexification pattern, also noted by Urban (2012) and Östling (2016) and comprehensively discussed by Schapper, San Roque and Hendery (2016). We will therefore restrict ourselves to presenting our data. As has been shown by Schapper, San Roque and Hendery (2016), colexification (either strict or loose) of ‘fire’ and ‘tree’ is well-attested in the Sahul area comprising the languages of Australia, New Guinea and surrounding islands, though not to the extent that this had been suggested in earlier research. In fact, Schapper, San Roque and Hendery (2016) argue that it is much more common for the Sahul languages to colexify ‘fire’ with ‘firewood’, to the exclusion of ‘tree’. However, since ‘firewood’ is not in the Swadesh list, this pattern cannot be extracted from the ASJP data. As we will see in Section 4.5, some relevant information can be obtained from the CLICS data, however. The ASJP data is shown in Figure 8 (in the following, only cluster areas that are significant predictors for the colexification pattern in question are indicated).

Fig. 8: Significant meso-cluster for <‘fire’, ‘tree’> in the ASJP data

4.4.2 <‘Mountain’, ‘stone’>

The colexification of ‘mountain’ and ‘stone’ has been discussed by Östling (2016) and is pervasive in the data used by Urban (2012). As Urban (2012) points out, Buck (1949: §1.22) already noticed a certain affinity between the meanings ‘mountain’ and ‘rock’ in Indo-European (e.g., Goth *hallus* ‘rock’, Old Norse *hallr* ‘large

stone’ or ‘sloping’ as an adjective, probably related to Latin *collis* ‘hill’ and Lithuanian *kalnas* ‘mountain’). Sometimes, such instances of colexification are probably mediated by a word for ‘cliff’ (e.g., Old High German *felis* ‘rock’, Old Norse *fjall* ‘mountain’, Irish *all* ‘rock, cliff’).

The macro-cluster shown in the second row of Table 3 covers the whole continent of Africa. As Figure 9 shows, it is particularly prominent in two parts of Africa, i.e., a central area that can perhaps, roughly, be negatively defined as a “non-Afroasiatic” and “non-Bantu” belt²⁵ and the Kalahari Basin in the south (cf. Güldemann 2010). The whole continent of Australia constitutes a particularly clear case of a cluster. It is, however, genealogically less diverse than the cluster in Africa (the diversity index is 1.9 in Australia versus 2.3 in Africa; note that the Australian cluster is listed as a meso-cluster in Table 3). A relatively weak cluster was identified in South America.

Fig. 9: Significant macro-clusters for <‘mountain’, ‘stone’> in the ASJP data

It seems plausible to assume that the colexification of ‘mountain’ and ‘rock’ may be rooted, to a certain extent, in the physical environment of the speakers, at least in Australia, with rocky environments and/or arid regions. A common alternative colexification partner of ‘mountain’ is ‘forest’, specifically in regions with a rich vegetation (e.g., [Mexican] Spanish *selva*, originally ‘forest’ from Latin *silva*, is often used for ‘rain forest’ as well as ‘montane forest’; see Urban 2012). The African cluster cannot, of course, be explained in this way. Tom Güldemann (p.c.)

²⁵ We owe this observation to Tom Güldemann.

has hypothesized that the colexification of ‘mountain’ and ‘stone’ shown in Figure 9 may represent a pattern that was widespread before the expansion of Afroasiatic and Bantu languages – a ‘remnant areal pattern’ between the zones covered by the two major families, as it were. We would obviously need other types of evidence (especially historical) to test this hypothesis. In any case, the “belt” of black squares in Figure 9 does not seem to correspond to an area that we could define positively – for instance, in terms of geographical characteristics or patterns of language contact.

4.4.3 <‘Ear’, ‘leaf’>

<‘Ear’, ‘leaf’> is a particularly common pattern in Eastern Africa, most of the languages being located in Güldemann’s (forthc.) “Nilotic-Surmic spread zone”. As Table 3 and Figure 10 show, there is a highly distinctive (meso-)cluster in this area, with 39 languages from 14 families. Our method has also identified some weaker and much smaller clusters in other parts of the world, i.e., in the Americas and in Australia, but they are far less diverse than the African cluster and it is questionable whether clusters with two members (as in Australia) should be taken into consideration at all (we included them because even two-member clusters may have resulted from language contact).

Fig. 10: Significant meso-clusters <‘ear’, ‘leaf’> in the ASJP data

Given the density of languages in the Nilotic-Surmic spread zone, it is hard to plot the specific languages or language families of the <‘ear’, ‘leaf’> cluster in this area. Figure 11 shows the Glottolog codes of the families: Afro-Asiatic (afro1255), Atlantic-Congo (atla1278), Central Sudanic (cent2225), Dizoid (dizo1235), Ta-Ne-Omoti (gong1255), Heibanic (heib1242), Khoe-Kwadi (khoe1240), Lafota (lafo1243), Maban (maba1274), Blue Nile Mao (maoo1243), Nilotic (nilo1247), Nubian (nubi1251), Songhay (song1307) and Surmic (surm1244).

Fig. 11: Macro-cluster <‘ear’, ‘leaf’> in Eastern Africa in the ASJP data – language families

As we are not specialists of African languages, we cannot interpret the facts from Eastern Africa any further. Specifically, it would be interesting to know what other meanings are involved in this pattern. Moreover, it would be intriguing to see if there is any connection between the physical environment and this colexification pattern (e.g., insofar as there are plants with ear-like leaves). While this

may seem a bit far-fetched, a cursory glance at materials available to us in fact points in that direction. According to Hellenthal (2010: 493), in Sheko (shek1245) (which is not in our data, but its relative Dizin [dizi1235] is), *haay* means both ‘ear’ and ‘leaf of ensete or yam’ (“ensete” is also known as “Ethiopian banana”) but there are other words for other types of leaves. What ensete and yam leaves have in common is that they are relatively large in comparison to their “hosts” and prominently stick out laterally (while still being curled up, the “top leaf” of an ensete is described with a different word in Sheko, i.e., *mùkmūrī*). This conjecture gets some support from the fact that the pattern is also found in the desert regions of Mexico and the Southern United States, where the agave (*americana*) and similar plants are widespread. We leave it to specialists, however, to explain the particularly strong (areal) association of ‘ear’ and ‘leaf’ in this part of the world.²⁶

4.4.4 <‘Bark’, ‘skin’>

Before we consider the colexification of ‘bark’ and ‘skin’ that emerged from our analysis, it should be pointed out that ‘bark’ is not in the 40-item word list actually used for ASJP (while being in the 100-item Swadesh list). The number of data points is therefore considerably lower than for most other pairs (cf. the case of ‘feather’ and ‘hair’ discussed in Sections 4.1 to 4.3). The two significant macro-cluster areas are shown in Figure 12. The cluster in South America is the “strongest” one.

What is remarkable about cluster area 1 is the genealogical heterogeneity of the languages exhibiting colexification of ‘bark’ and ‘skin’. As Table 3 shows, the cluster comprises fourteen languages from eleven families (note that some of the languages listed in the following are also mentioned by Urban 2012): Abipon (abip1241) is Guaicuruan; Masaká (aika1237) is Aikanã; Apinayé (apin1244) is Nuclear-Macro-Je; Bororo (boro1282) is Bororoan; Cha’palaa (chac1249), Tsafiki (colo1256) and Guambiano (guam1248) are Barbacoan; Hixkaryána (hixk1239) is Cariban; Hupdë (hupd1244) is Nadahup; Minica Huitoto (mini1256) is Huitotoan;

²⁶ Guillaume Segerer (p.c.) has pointed out to us that one interesting aspect of the <‘ear’, ‘leaf’> colexification pattern is that, with a few exceptions, it seems to be absent from Niger-Congo. In his data, it is attested in Sheko (shek1245), the Omotic languages of Wolaytta (wola1242), Gofa (gofa1235), Dorze (dorz1235), Dawro (dawr1236), Haro (kach1284), Basketo (bask1236) [Ta-Ne-Omotic] and Dime (dime1235) [South Omotic], and the Central Sudanic languages of Proto-SBB, Modo (modo1248), Bongo (bong1285), Mangbetu (mang1394), Avokaya (avok1242), Kaliko (kali1312), Logo (logo1259), Lugbara (lugb1240) and Ma’di (madi1260).

Páez (paez1247) is Páez; Paraguayan Guaraní (para1311) and Parakanã (para1312) are Tupian; and Tacana (taca1256) is Pano-Tacanan.

Figure 12 shows that there is also a cluster in Melanesia where ‘bark’ and ‘skin’ are colexified. It comprises seven languages from two families (Austronesian and South Bougainville) and is therefore not as distinctive as the cluster in South America. Remember, however, that the number of data points is very small for this part of the world, as ‘bark’ is not included in the 40-item ASJP list.

Fig. 12: Macro-clusters for <‘bark’, ‘skin’> in the ASJP data

The colexification of ‘bark’ and ‘skin’ has also been discussed, but actually discarded, as a potential lexical trait of Mesoamerican languages (cf. Smith-Stark 1994). As Figure 12 shows, there is no significant cluster in this area, according to the ASJP data. This actually confirms Smith-Stark’s (1994) observation that ‘bark’ and ‘skin’ are not characteristic of that area, as they are also colexified in neighboring, non-Mesoamerican languages.²⁷ However, our data suggests that, even within the cluster, the degree of colexification is limited. Figure 13 shows the hypothesized cluster area, identified by the hierarchical clustering process, and the languages with and without colexification of ‘bark’ and ‘skin’.

²⁷ For ways of quantifying membership to the Mesoamerican linguistic area, see van der Auwera (1998b) and Gast (2007).

Fig. 13: Colexification of ‘bark’ and ‘skin’ in Mesoamerica in the ASJP data

While Figure 13 suggests that colexification of ‘bark’ and ‘skin’ is actually much less widespread in Mesoamerica than we may have thought, it seems to us that what we find here in many cases is “loose colexification”. For example, in the Mixe-Zoquean language Copainalá Zoque (copa1236), *naca* means ‘skin’ and *ku’yu-naca* ‘bark’ or, literally, ‘tree-skin’ (Harrison, Harrison and García H. 1981: 280, 355). This example thus shows, once again, that a separate treatment of loose colexification will be beneficial to detect a broader range of colexification patterns.

4.5 Clusters emerging from the CLICS-data

Before we start to explore the data in a bottom-up way, it seems worthwhile to see if we can get any information from CLICS that would allow for a comparison with the ASJP data. As pointed out in Section 2.1, the types of concepts covered in CLICS are very different from those in ASJP and there is little overlap. One pair of elements, however, can be reasonably compared: CLICS provides information about ‘fire’ and ‘firewood’, which, in a way, complements the information on ‘fire’ and ‘tree’ discussed in Section 4.4.1. Interestingly, the <‘fire’, ‘tree’> pattern is not found in CLICS at all. As can be seen in Figure 14 (showing hypothesized

cluster areas), the data is very sparse, however, and it is not surprising that the clusters – even though the colexification pattern shows a spatial autocorrelation according to the Join Count test – did not pass the significance test in the regression analysis.

Fig. 14: Hypothesized cluster areas for <fire, 'firewood'> in the CLICS data

While the <fire, 'firewood'> clusters in Southeast Asia and Australia/New Guinea are in accordance with the ASJP data on the colexification of 'tree' and 'fire', shown in Figure 8, it is perhaps surprising that the CLICS data shows some relevant data points in South America, as this part of the world does not exhibit a single instance of colexification of 'fire' and 'tree' in the ASJP data. CLICS lists the following languages for this pattern: Qawasqar (qawa1238), Araona (arao1248), Wayuu (wayu1243), Kaingang (kain1272), E'ñapa Woromaipu (also known as Panare, enap1235) and Yavitero (yavi1244). To be sure, 'firewood' and 'tree' are different things but it still seems surprising that, according to the combined ASJP and CLICS data, not a single South American language seems to colexify both 'fire' and 'firewood' and 'firewood' and 'tree' – a pattern that is attested in 10% of the Sahul sample in Schapper, San Roque and Hendery (2016). As we are not familiar with any of the South American languages involved, we have no further comments to make at this point.

We can now turn to a data-driven inspection of the clusters emerging from the CLICS data. Our method identified 36 colexification patterns with a significant spatial autocorrelation and 37 cluster areas that were significant predictors for a given colexification (at a 0.05 level). There are twelve micro-clusters, thirteen

meso-clusters and twelve macro-clusters. The colexification patterns are listed in (2) and the thirteen significant meso-clusters are shown in Figure 15. Given the areal bias of the data, it comes as no surprise that most of the clusters are located in South America and Eurasia.

- (2) <'air', 'weather'>; <'beak', 'mouth'>; <'believe', 'think'>; <'body', 'flesh'>; <'brother', 'sister'>; <'buy', 'take'>; <'catch (ball)', 'take'>; <'count', 'measure'>; <'count', 'think (be of the opinion)'>; <'country', 'earth (ground, soil)'>; <'daughter in law (of a woman)', 'mother in law (of a woman)'>; <'daughter', 'son'>; <'dig', 'drop (verb)'>; <'dye', 'paint (noun)'>; <'earth (ground, soil)', 'world'>; <'father in law (of a man)', 'son in law (of a man)'>; <'female (adjective)', 'woman'>; <'female', 'woman'>; <'furs', 'skin (hide)'>; <'get (obtain)', 'take'>; <'grandson', 'nephew'>; <'grass', 'pasture'>; <'grass', 'plant'>; <'green', 'green (unripe)'>; <'hold', 'keep (retain)'>; <'hold', 'seize (grasp)'>; <'hold', 'take'>; <'language', 'voice'>; <'leg', 'thigh'>; <'male (adjective)', 'man (vs. woman)'>; <'male', 'man (vs. woman)'>; <'nephew', 'niece'>; <'offspring (son or daughter)', 'son'>; <'post (pole)', 'tree'>; <'release (let go)', 'send'>; <'time', 'weather'>

Fig. 15: All meso-clusters emerging from the CLICS data

Some of the clusters emerging from the data are not terribly interesting. For instance, there are many cases of adjectives and nouns being colexified, such as 'woman' and 'female', which is a matter of morphosyntax rather than the lexicon.

The top ten of the remaining significant clusters are shown in Table 4 (the data for all clusters with a p-value lower than 0.1 is provided online).²⁸

Tab. 4: Significant lexical clusters from the CLICS data

Colexification	Size	No	Long	Lat	Radius	No lg	No fam	Div	pm	p.pm
<count, measure>	macro	1	-68.0	-6.1	2,578	18	15	2.2	3.1	0.010
<catch (ball), take>	meso	1	-67.4	-8.6	2,181	16	13	2.1	1.85	0.048
<count, measure>	meso	1	-70.8	1.2	1,237	12	11	1.9	2.74	0.026
<hold, take>	macro	1	-62.0	-30.9	2,754	9	8	1.6	2.17	0.034
<count, measure>	micro	1	-75.4	-1.8	643	7	7	1.5	3.07	0.025
<air, weather>	macro	1	44.5	41.3	3,603	22	5	1.2	2.73	0.042
<get (obtain), take>	macro	1	48.3	46.3	3,088	24	5	1.2	2.38	0.039
<hold (seize), grasp>	micro	2	46.7	41.1	1,484	24	4	1.0	3.60	0.004
<count (think) be of the opinion>	meso	1	46.2	41.9	1,557	18	4	1.0	2.98	0.011
<hold (seize), grasp>	meso	2	46.7	41.1	1,484	24	4	1.0	3.93	0.004

Table 4 can be split into two major groups of clusters, those from South America (the first five) and those from Eurasia (the last five). The five South American clusters comprise three colexification types: <‘count’, ‘measure’> (micro, meso and macro), <‘catch (ball)’, ‘take’> and <‘hold’, ‘take’>. The most prominent pair is clearly <‘count’, ‘measure’>. The macro-cluster listed at the top of Table 4 is shown in Figure 16. As the data in the table shows, it is extremely diverse genealogically speaking, and is found in fifteen languages from eighteen families.

²⁸ See <http://www.uni-jena.de/~mu65qev/data/colex-tables/CLICS-clusters.htm> (accessed 5 April 2018).

Fig. 16: Macro-clusters for '<count>', '<measure>' in the CLICS data

The following clusters are all located in Eurasia: '<air>', '<weather>' (macro), '<get (obtain)>', '<take>' (macro), '<hold>', '<seize (grasp)>' (micro and meso) and '<count>', '<think (be of the opinion)>' (meso and macro). The “strongest” macro-cluster for '<air>', '<weather>' is shown in Figure 17. As the map shows, this colexification is divided rather categorically between South America and Eurasia.

Fig. 17: Macro-clusters for '<air>', '<weather>' in the CLICS data

Further inspection of the data reveals some more interesting patterns, which we do not have the space to discuss at this point. We will therefore restrict ourselves to some comments on one group of clusters discussed above. It is interesting to see that CLICS brings to light a number of verbal patterns. The lexical typology of verbal meanings is particularly difficult, because comparability is even harder to establish in this domain than it is for nominal meanings, and elicitation represents an additional challenge. We have to reckon with elicitation artefacts, for instance, stemming from the language used during the interviews or the language in which the concepts are coded. For example, the fact that colexification of ‘count’ and ‘think (be of the opinion)’ (see Table 4) is widespread in Eurasia may be related to the fact that the concepts are colexified in Russian as well. *Schitat’* is actually listed as an elicitation form for both concepts in the IDS data (though, for ‘think’, an alternative form is given, i.e., *dumat’*). But then, it is possible that Russian is just another language colexifying these concepts, perhaps under areal influence (note that the elicitation languages of the IDS data are not listed in CLICS). A more comprehensive study of such questions obviously requires more data from other parts of the world.

5 Conclusions

The aim of this study has been of an exploratory nature. We intended to determine how much information about areal patterns of colexification we can gain from lexical databases such as CLICS and ASJP. We chose a bottom-up (rather than hypothesis-driven) approach, identifying areal patterns in three steps: (i) determine spatial autocorrelations in the data, (ii) identify clusters as candidates for convergence areas and (iii) test the clusters resulting from the second step controlling for genealogical relatedness. Moreover, we identified a (genealogical) diversity index for each cluster. For the ASJP data, we identified clusters associated with four colexification pairs in this way: <‘fire’, ‘tree’>, <‘mountain’, ‘stone’>, <‘ear’, ‘leaf’> and <‘bark’, ‘skin’>. One of these patterns has figured prominently in recent research carried out by specialists, i.e., <‘fire’, ‘tree’> (see Schapper, San Roque and Hendery 2016). Two of the colexification types have been discussed before, though not in very much detail, i.e., <‘mountain’, ‘stone’> and <‘bark’, ‘skin’> (see Urban 2012). The colexification of ‘ear’ and ‘leaf’, which seems to be prominent in Eastern Africa, has not been noted in the lexical-typological literature, as far as we are aware (we are, of course, not familiar with the whole range of specialized literature). We regard these results as a proof of concept, in the sense that our bottom-up approach has yielded promising results.

Inspection of the other patterns emerging from the ASJP data shows a number of further interesting pairs, some of which may inspire more detailed research, such as <'ear', 'name'>, <'horn', 'tooth'> and <'horn', 'knee'> (see the more comprehensive cluster lists in our online data repository, cf. footnote 25).

The CLICS data shows a heavily biased areal distribution but it can be used to identify some differences between South America and Eurasia. Once interesting pairs of concepts have been identified, other regions of the world could be investigated more thoroughly. What makes the CLICS data particularly interesting is the inclusion of a broad range of concepts, including verbal ones. Given the relative scarcity of data and given the rather strict selection criteria that we applied (e.g., a 0.05 level of significance), we only identified a relatively small number of clusters. Still, some of these clusters are potentially interesting from an areal point of view. Moreover, closer inspection of the data in the online repository (where all clusters with a p-value < 0.1 are listed, cf. footnote 29), again, brings to light some further relevant patterns, such as clusters for various kinship terms (e.g., <'daughter', 'son'>, <'brother', 'sister'>, <'grandson', 'nephew'>) and body parts (e.g., <'leg', 'thigh'>), as well as for the particularly interesting case of <'language', 'voice'>. The conceptualization of 'language' seems to vary greatly across the regions of the world, including metonymies such as 'tongue', 'voice' and 'word' (see Radden 2004). The ways in which the (rather abstract) concept 'language' is encoded deserves a typological study of its own. Here as well, inspiration for follow-up research was found in the data.

We have also pointed out some drawbacks of the use of major lexical databases. First, the data has been collected from various sources, which means that they are not based on consistent definitions, and most of the primary data was probably elicited through English (or some other major language, such as Russian), so that the elicitation stimulus functioned as a *tertium comparationis*. This implies the danger of elicitation artefacts of various types. As we saw in Section 4.5, some languages of Eurasia, including Russian, colexify 'count' and 'think (be of the opinion)'. It is possible that this is a genuine areal pattern but we cannot rule out that at least some data points were influenced by the (probably Russian) elicitation word. Note that the problems concerning the status of verbal meanings in cross-linguistic comparison are, of course, of a more general nature. Elicitation of verbal concepts represents a well-known problem for language documentation and linguistic typology, for at least two reasons. First, it is hard to know *a priori* what verbal concepts a language encodes, specifically when dealing with cultural practices that most of us are unfamiliar with (such as hunting). Second, actions are mostly harder to describe or paraphrase than nominal concepts. Multimodal (or even behavioral) elicitation techniques are therefore required (e.g.,

videos). Elicitation is thus “expensive” and it is likely that, in general, lexicons from little described languages exhibit a “nominal bias” for this reason. It is therefore a great asset that CLICS contains a considerable number of verbal concepts – most of them originating from the IDS data – even though the data, obviously, has to be handled with care.

Having started the experiment of detecting areal colexification patterns bottom-up in an entirely unprejudiced way, our conclusion is predominantly positive. Even in our small-scale exploratory study we have identified various topics that deserve closer investigation. In spite of the inevitable noise in data gathered on a large scale and in collaborative efforts, we hope to have shown that lexical databases represent a valuable tool for typological research, even when used for purposes that they were not originally intended for.

Acknowledgement: We would like to thank Søren Wichmann, Matthias Urban, Johann-Mattis List, Tom Güldemann, Robert Östling, Karsten Schmidtke-Bode, Daniël Van Olmen and an anonymous reviewer for comments on an earlier version of this paper and valuable feedback. Any mistakes or inaccuracies are, of course, our own responsibility. MKT’s research is supported by the grant SAB17-0588:1 from Riksbankens Jubileumsfond.

References

- Aikhenvald, Alexandra Y. 2009. ‘Eating’, ‘drinking’ and ‘smoking’: A generic verb and its semantics in Manambu. In John Newman (ed.), *The linguistics of eating and drinking*, 91–108. Amsterdam: John Benjamins.
- Berlin, Brent & Paul Kay. 1969. *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.
- Bivand, Roger & Gianfranco Piras. 2015. Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software* 63 (18). 1–36. <http://www.jstatsoft.org/v63/i18> (accessed 4 April 2018).
- Bivand, Roger, Jan Hauke & Tomasz Kossowski. 2013. Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis* 45 (2). 150–179.
- Brown, Cecil H. 2013a. Finger and hand. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/130> (accessed 25 April 2017).
- Brown, Cecil H. 2013b. Hand and arm. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/129> (accessed 25 April 2017).
- Brown, Cecil H., Eric W. Holman, Søren Wichmann & Viveka Velupillai. 2008. Automated classification of the world’s languages: A description of the method and preliminary results. *STUF/Language Typology and Universals* 61 (4). 285–308.

- Buck, Carl Darling. 1949. *A dictionary of selected synonyms in the principal Indo-European languages*. Chicago: University of Chicago Press.
- Cliff, Andrew D. & J. Keith Ord. 1981. *Spatial processes: Models and applications*. London: Pion.
- Cruse, D. Alan. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.
- Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF/Language Typology and Universals* 60 (2). 95–99.
- Dahl, Östen & Maria Koptjevskaja-Tamm. 2001. Kinship in grammar. In Irène Baron, Michael Herslund & Finn Sørensen (eds.), *Dimensions of possession*, 201–225. Amsterdam: John Benjamins.
- Evans, Nicholas. 2011. Semantic typology. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 504–548. Oxford: Oxford University Press.
- Evans, Nicholas & David Wilkins. 2000. In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language* 76 (3). 546–592.
- Fraley, Chris & Adrian E. Raftery. 2002. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* 97 (458). 611–631.
- Fraley, Chris, Adrian E. Raftery, T. Brendan Murphy & Lucca Scrucca. 2012. *mclust version 4 for R: Normal mixture modeling for model-based clustering, classification and density estimation*. Seattle: University of Washington.
- 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: John Benjamins.
- Gast, Volker. 2007. From phylogenetic diversity to structural homogeneity: On right-branching constituent order in Mesoamerica. *SKY Journal of Linguistics* 20: 171–202.
- Gast, Volker & Johan van der Auwera. 2012. What is “contact-induced grammaticalization”? Examples from Mesoamerican languages. In Björn Wiemer, Bernhard Wälchli & Björn Hansen (eds.), *Grammatical replication and grammatical borrowing in language contact*, 381–426. Berlin: De Gruyter.
- Gast, Volker & Johan van der Auwera. 2013. Towards a distributional typology of human impersonal pronouns, based on data from European languages. In Dik Bakker & Martin Haspelmath (eds.), *Languages across boundaries: Studies in memory of Anna Siewierska*, 119–158. Berlin: De Gruyter.
- Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.
- Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2 (4). 1360–1383.
- Gelman, Andrew & Yu-Sung Su. 2016. *arm: Data analysis using regression and multilevel/hierarchical models: R package version 1.9-3*. <http://CRAN.R-project.org/package=arm> (accessed 4 April 2018).
- Güldemann, Tom. 2010. Sprachraum and geography: Linguistic macro-areas in Africa. In Alfred Lameli, Roland Kehrein & Stefan Rabanus (eds.), *Language and space: An international handbook of linguistic variation, volume 2: Language mapping*, 561–585. Berlin: De Gruyter.
- Güldemann, Tom. forthc. Areal linguistics beyond contact and linguistic areas in Afrabia. In Tom Güldemann (ed.), *The languages and linguistics of Africa*. Berlin: De Gruyter.
- Hadfield, Jarrod D. 2010. MCMC methods for Multi-response Generalised Linear Mixed Models: The MCMCglmm R package. *Journal of Statistical Software* 33 (2). 1–22.

- Harrison, Roy, Margaret Harrison & Cástula García H. 1981. *Diccionario zoque de Copainalá*. Mexico City: Instituto Lingüístico de Verano.
- Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Clarendon.
- Haspelmath, Martin & Uri Tadmor (eds.). 2009. *Loanwords in the world's languages: A comparative handbook*. Berlin: De Gruyter.
- Heine, Bernd & Tania Kuteva. 2003. On contact-induced grammaticalization. *Studies in Language* 27 (3). 529–572.
- Heine, Bernd & Tania Kuteva. 2005. *Language contact and grammatical change*. Cambridge: Cambridge University Press.
- Hellenthal, Anne-Christie. 2010. *A grammar of Sheko*. Utrecht: LOT.
- Hijmans, Robert J. 2016. *geosphere: Spherical trigonometry, R package version 1.5-5*. <http://CRAN.R-project.org/package=geosphere> (accessed 4 April 2018).
- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller & Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica* 42 (2). 331–354.
- Juvonen, Päivi & Maria Koptjevskaja-Tamm (eds.). 2016. *The lexical typology of semantic shifts*. Berlin: De Gruyter.
- Kay, Paul & Luisa Maffi 2013a. Green and blue. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/134> (accessed 25 April 2017).
- Kay, Paul & Luisa Maffi 2013b. Red and yellow. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/135> (accessed 25 April 2017).
- Key, Mary Ritchie & Bernard Comrie (eds.) 2015. *The intercontinental dictionary series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://ids.cld.org> (accessed 26 April 2017).
- Kleiber, Georges. 1990. *La sémantique du prototype: Catégories et sens lexical*. Paris: Presses Universitaires de France.
- Koptjevskaja-Tamm, Maria (ed.). 2015. *The linguistics of temperature*. Amsterdam: John Benjamins.
- Koptjevskaja-Tamm, Maria, Dagmar Divjak & Ekaterina Rakhilina. 2010. Aquamotion verbs in Slavic and Germanic: A case study in lexical typology. In Victoria Driagina-Hasko & Renee Perelmutter (eds.), *New approaches to Slavic verbs of motion*, 315–341. Amsterdam: John Benjamins.
- 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: Routledge.
- Koptjevskaja-Tamm, Maria & Henrik Liljégren. 2017. Semantic patterns from an areal perspective. In Raymond Hickey, R. (ed.), *The Cambridge handbook of areal linguistics*, 204–236. Cambridge: Cambridge University Press.
- Lau, Matthew K. 2012. *DiversitySampler: Functions for re-sampling a community matrix to compute diversity indices at different sampling levels, R package version 2.1*. <https://CRAN.R-project.org/package=DiversitySampler> (accessed 4 April 2018).
- Lipka, Leonhard. 1992. *An outline of English lexicology: Lexical structure, word semantics and word-formation*. Tübingen: Niemeyer.

- List, Johann-Mattis, Thomas Mayer, Anselm Terhalle & Matthias Urban. 2014. *CLICS: Database of cross-linguistic colexifications*. Marburg: Forschungszentrum Deutscher Sprachatlas. <http://CLICS.lingpy.org> (accessed 4 April 2017).
- Maisak, Timur & Ekaterina Rakhilina (eds.). 2007. *Glagoly dviženija v vode: Leksičeskaja tipologija*. [Verbs of motion and location in water: Lexical typology]. Moscow: Indrik.
- Matras, Yaron & Jeanette Sakel. 2004. Investigating the mechanisms of pattern replication in language convergence. *Studies in Language* 31 (4). 829–865.
- Majid, Asifa, Nick J. Enfield & Miriam van Staden (eds.). 2006. Parts of the body: Cross-linguistic categorisation. [Special issue]. *Language Sciences* 28 (2/3).
- Mayer, Thomas, Johann-Mattis List, Anselm Terhalle & Matthias Urban. 2014. An interactive visualization of cross-linguistic colexification patterns. In *Proceedings of the VisLR workshop at the International Conference on Language Resources and Evaluation 2014*, 1–8. <http://linguist.de/documents/mayer-et-al-2014-clics-visualization.pdf> (accessed 4 April 2018)
- Nerlove, Sara & A. Kimball Romney. 1967. Sibling terminology and cross-sex behavior. *American Anthropologist* 74. 1249–1253.
- 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: De Gruyter.
- Radden, Günter. 2004. The metonymic folk model of language. In Barbara Lewandowska-Tomaszczyk & Alina Kwiatkowska (eds.), *Imagery in language: Festschrift in honour of Professor Ronald W. Langacker*, 543–565. Frankfurt am Main: Lang.
- 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.
- Schapper, Antoinette, Lila San Roque & Rachel Hendery. 2016. Tree, firewood and fire in the languages of Sahul. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shifts*, 355–422. Berlin: De Gruyter.
- Simpson, Edward H. 1949. Measurement of diversity. *Nature* 163. 688.
- Smith-Stark, Thomas C. 1994. Mesoamerican calques. In Carolyn J. MacKay & Verónica Vásques (eds.), *Investigaciones lingüísticas en Mesoamérica*, 15–50. Mexico City: Universidad Nacional Autónoma de México.
- Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical forms. In Timothy Shopen (ed.), *Language typology and syntactic description, volume 3: Grammatical categories and the lexicon*, 57–149. Cambridge: Cambridge University Press.
- Urban, Matthias. 2012. *Analyzability and semantic associations in referring expressions: A study in comparative lexicology*. Leiden: University of Leiden PhD dissertation.
- van der Auwera, Johan. 1998a. Phasal adverbials. In Johan van der Auwera with Dónall P. Ó Baoill (eds.), *Adverbial constructions in the languages of Europe*, 25–145. Berlin: De Gruyter.
- van der Auwera, Johan. 1998b. Revisiting the Balkan and Mesoamerican linguistic areas. *Language Sciences* 20 (3). 259–270.
- van der Auwera, Johan & Ceyhan Temürçü. 2006. Semantic maps in typology. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, 131–134. Oxford: Elsevier.
- van der Auwera, Johan, Volker Gast & Jeroen Vanderbiesen. 2012. Human impersonal pronouns in English, Dutch and German. *Leuvense Bijdragen* 98 (1). 27–64.

- van der Auwera, Johan & Lauren Van Alsenoy. 2011. Mapping indefinites: Towards a neo-Aristotelian approach. In Eliza Kitis, Nikolaos Lavidas, Nina Topintzi & Tasos Tsangalidis (eds.), *Selected papers from the 19th International Symposium on Theoretical and Applied Linguistics*, 1–14. Thessaloniki: Monochromia.
- van der Auwera, Johan & Lauren Van Alsenoy. 2013. Mapping the West Germanic *any*'s. In Eva Csipak, Regine Eckardt, Mingya Liu & Manfred Sailer (eds.), *Beyond 'any' and 'ever': New explorations in negative polarity sensitivity*, 267–298. Berlin: De Gruyter.
- Vanhove, Martine. 2008a. 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: John Benjamins.
- Vanhove, Martine (ed.). 2008b. *From polysemy to semantic change: Towards a typology of lexical semantic associations*. Amsterdam: John Benjamins.
- Viberg, Åke. 1984. The verbs of perception: A typological study. In Brian Butterworth, Bernard Comrie & Östen Dahl (eds.), *Explanations for language universals*, 123–162. Berlin: De Gruyter.
- Wichmann, Søren, Eric W. Holman & Cecil H. Brown (eds.). 2016. *The ASJP database (version 17)*. <http://asjp.cld.org/> (accessed 4 April 2018).
- Zalizniak, Anna. 2008. A catalogue of semantic shifts: Towards a typology of semantic derivation. In Martine Vanhove (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 217–238. Amsterdam: John Benjamins.
- Zalizniak, Anna, Maria Bulakh, Dmitry Ganenkov, Ilya Gruntov, Timur Maisak & Maxim Russo. 2012. The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics* 50 (3). 633–669.