AUTOMATIC DATA COLLECTION IN LEXICAL TYPOLOGY

Ryzhova D.A. (daria.ryzhova@mail.ru), National Research University Higher School of Economics, Moscow, Russia
Melnik A.A. (melnik-a-a@mail.ru), National Research University Higher School of Economics, Moscow, Russia
Yershov I.A. (cornerstone26@yandex.ru), National Research University Higher School of Economics, Moscow, Russia
Panteleeva I. M. (impanteleyeva@gmail.com), National Research University Higher School of Economics, Moscow, Russia
Paperno D.A. (denis.paperno@loria.fr), National Center for Scientific Research (CNRS), Lorraine Research Laboratory of Computer Science and its Applications, France
Singh Ya. (yaju.singh007@gmail.com), Indian Institute of Technology, Kharagpur, WB, India
Sobolev M. (strategy155@gmail.com), National Research University Higher School of Economics, Moscow, Russia

АВТОМАТИЧЕСКИЙ СБОР ДАННЫХ В ЛЕКСИЧЕСКОЙ ТИПОЛОГИИ

Рыжова Д.А.(daria.ryzhova@mail.ru), Научно-исследовательский университет «Высшая школа экономики», Москва, Россия
Ершов И.А. (cornerstone26@yandex.ru), Научно-исследовательский университет «Высшая школа экономики», Москва, Россия
Мельник А. А. (melnik-a-a@mail.ru), Научно-исследовательский университет «Высшая школа экономики», Москва, Россия
Пантелеева И. М. (impanteleyeva@gmail.com), Научно-исследовательский университет «Высшая школа экономики», Москва, Россия
Паперно Д.А. (denis.paperno@loria.fr), Национальный центр научных исследований (CNRS), Лаборатория исследований в области компьютерных наук и их приложений, Франция
Синг Я. (yaju.singh007@gmail.com), Индийский институт технологии, Харагпур, Индия
Соболев М (strategy155@gmail.com), Научно-исследовательский университет «Высшая школа экономики», Москва, Россия

The paper addresses an issue of an automatic data collection for lexical typological studies in the Frame approach paradigm. A research in this framework is based on the analysis of distributional properties of the lexemes in question. Hence, questionnaires for such studies consist of typical contexts where lexical items from a given semantic domain can potentially occur. We aim at filling these questionnaires automatically, and this task can be splitted into two different problems: questionnaire translation and its filling with the relevant data. We suggest three methods for the first task completion (translation via bilingual dictionaries vs. online cloud translators vs. parallel corpora), and two algorithms are focused on the second task (filling of a questionnaire based on monolingual corpora vs. on online translators). We test our algorithm on the data from four semantic domains of qualitative features (‘sharp’, ‘smooth’, ‘thick’, ‘thin’).

Key words: lexical typology, synonyms, corpus study, automatic translation, parallel corpus.

1. Introduction

This paper presents a pipeline for an automatic collection of lexical typological data. Lexical typology is a branch of linguistics interested in differences and similarities between word meanings cross-linguistically. Emerged relatively recently, this domain has already gained an extensive attention from many researchers all over the world. Several approaches have been proposed, such as the *Natural Semantic Metalanguage* (Wierzbicka 1984, Goddard and Wierzbicka 2007), or the denotation-based approaches (Berlin and Kay 1969, Majid et al. 2006); see Koptjevskaja-Tamm et al. (2016) for a recent overview. Many semantic fields have already received typological descriptions (for example, 'put / take' (Narasimhan and Kopecka 2012), 'cut / break' (Majid and Bowerman, 2007), temperature terms (Koptjevskaja-Tamm 2015)).

In the present paper, we adopt the Frame Approach to lexical typology (Rakhilina, Reznikova 2016) that relies on the linguistic behavior of lexical items and consists primarily in comparison of words' distributional properties across languages. As has been assumed in many theoretical and computational studies, different meanings of words occur in different contexts, so the analysis of co-occurrences can serve as a key to the differences in lexical meanings. For example, the English word *thick* can combine with the noun *grass*, while the corresponding Russian adjective *tolstyj* 'thick' never modifies the noun *trava* 'grass'. This distributional phenomenon reflects a fundamental difference in the meanings of the so-called translational equivalents *thick* and *tolstyj*. Besides their shared function of marking a long distance between the opposite sides, edges or surfaces of an object (compare *thick layer* - *tolstyj sloj*), the former can refer to a property of having small items located very closely together, while the latter does not express this meaning (in Russian, such situations are described by the adjective *gustoj*).

The Frame Approach has already proven its viability: various semantic domains in dozens of languages have been examined and described within this paradigm, including verbs of motion in liquid (Lander et al. 2012), pain predicates (Reznikova et al. 2012), verbs of rotation (Kruglyakova 2010), and many others. The results of this research, namely the data on differences in distributional properties between translational equivalents, should be useful in human and machine translation (see Kyuseva et al. 2013); moreover, these data can serve as a new evaluation metric for distributional models (see Ryzhova et al. 2016 for more details). The

main shortcoming of the approach that prevents these data from being used in linguistic applications is that their collection takes too much time and effort. In this paper, we suggest several algorithms that could help to overcome this problem.

The paper is structured as follows. First, in Section 2, we explicate the task. Section 3 reports the algorithms of automatic data collection that we suggest: methods of questionnaire translation (3.1) and possible ways of its automatic filling (3.2). In Section 4 we provide the empirical results, and we conclude in Section 5.

2. Theoretical background

In lexical typology, several effective methodologies of automatic data collection already exist. They rely mostly on two types of resources: machine-readable bilingual dictionaries and multilingual corpora. Dictionaries allow comparing sets of meanings of lexical items, revealing some prominent patterns of colexification (see François 2008, Youn et al. 2016, among others). Multilingual corpora give an opportunity to find contextual synonyms and to detect differences in their usage (Wälchli and Cysouw 2012, Eger 2012).

However, as E. Rakhilina and T. Reznikova (2016) show, one needs to examine all possible language resources (dictionaries, corpora, elicitation results) in order to reveal the structure of any given semantic field and to describe the patterns of its lexicalization in all languages of the sample. Existing methodologies are indeed powerful, but only applied to a very restricted part of the lexicon. Dictionary entries are often poorly structured and hardly comparable, and they can miss some typologically relevant oppositions, simply because these distinctions are not lexicalized in languages in question. For example, a single English adjective *sharp* can be translated into French as *tranchant* or *pointu*, depending on the context (generally speaking, *tranchant* describes instruments with a cutting edge, and *pointu* modifies nouns denoting objects with a piercing point, or objects of a pointed shape). However, this semantic distinction is almost never reflected in English dictionaries[1]. As a result, it cannot be captured automatically based on dictionary data: the CLICS database that accumulates data from intercontinental dictionary series (Key and Comrie 2007) and several other wordlists (see List

[1] For example, in MacMillan dictionary, these meanings are listed together under the first definition of the word *sharp: a sharp object has an edge that can cut or an end that is pointed*. (URL: https://www.macmillandictionary.com/dictionary/british/sharp_1)

et al. 2014 for more details) presents a single lexical concept ‘sharp’ for both cutting and piercing instruments[2].

Multilingual corpora provide much more reliable data (see Eger 2012), but by the moment they are not big enough to serve as a basis for typological analysis of any words other than very frequent lexical items (such as verbs of motion or locative predicates).

In our research, we use manually prepared typological questionnaires and attempt to fulfill them automatically with the data from available lexical resources. Thus, we aim at elaborating a set of tools that would facilitate data collecting, but we rely on a pre-existing set of potentially important types of contexts to be sure that we get relevant information and to have a solid basis for further comparison of the collected data.

We focus on four semantic fields of qualitative features: ‘sharp’, ‘smooth’, ‘thick’ and ‘thin’. We rely on the lexicon of this type for several reasons. First, fine distinctions between words with such meanings are usually underrepresented in dictionaries (see the example from the domain ‘sharp’ above). Second, these domains have already been analyzed in typological perspective (see Kashkin and Vinogradova to appear, Kozlov and Privizentseva to appear, Kyuseva et al. to appear), so that we have manually prepared and verified questionnaires in English and Russian for every domain filled with the data from several languages. Finally, words from these semantic fields usually belong to the class of adjectives, which are easy to handle: in most cases, it’s enough to take into account only nouns that an adjective can modify to capture the most important principles of its distribution (Bolinger 1967, Rakhilina 2000). Thus, a questionnaire for such semantic domain consists of a list of nouns and a list of adjectives, and to fill it one has to indicate if a given adjective can combine with a given noun (see an illustration in Table 1).

	thick	*broad*	*wide*	*large*	*fat*
book	+				
wall	+	+		+	
rope	+				
person					+

[2] It should be mentioned that there is also a concept ‘pointed’ in the database, but the difference between ‘sharp’ and ‘pointed’ concepts is not evident. The English word *pointed* usually describes an object’s shape (*a pointed nose/chin*), but not a sharpened piercing point of an instrument, but the relationship between English lexical items and cross-linguistic concepts in CLICS remains unclear.

Table 1. Fragment of the questionnaire for the domain ‘thick’ filled with the English data

Since a questionnaire in the Frame paradigm consists of context examples, the task of its filling boils down into two subtasks: 1) translation of a questionnaire (i.e. the relevant adjectives and nouns) into the target language; 2) filling it with the data from this language. For each semantic domain, we conduct three experiment series based on two source and target language pairings: English -> Russian and Russian/English -> Italian.

3. Suggested algorithms for collection of lexical typology data

3.1. Questionnaire translation

The task of questionnaire translation is not as trivial as it may seem at the first sight. The most challenging part is finding translation equivalents for the adjectives in question, since in this case one does not need to find the most appropriate (or the most frequent) translation of a word from the source language, but to reveal *all* lexical items belonging to a given semantic domain in the target language. As we have already seen in the previous section, the correspondence between adjectives of one semantic domain in different languages is not one-to-one: at least two adjectives correspond to the single English *sharp* in French, and in Mandarin Chinese this field contains up to eight lexical items.

We have applied three techniques to complete the task of questionnaire translation. The first method consists in translating adjectives and nouns with the help of general-purpose machine translation systems; we used online translation systems offered by Yandex and Google since they presumably represent the state of the art. The second method consists in translation via machine-readable bilingual dictionaries. The third way extracts translation equivalents from bilingual corpora.

3.1.1 Online translators

In this section we will describe the algorithm of questionnaire translation that is based on existing machine translation system treated as a black box. The simplest strategy for noun translation is straightforward -- using noun as input to cloud translator and getting its output. For this purpose two different cloud translators APIs were used, Yandex.Translate and Google Translate. Our method for adjective translation is much more complicated because, unlike the case above, the number of adjectives in a questionnaire fluctuates from one language to another, and a primitive translation scheme won't achieve good results. We decided to modify the

algorithm so that it checks all synonyms of the adjective translated. After that we perform backwards translation for every synonym that we have got on the first stage. If the backwards translation of a synonym coincides with the input adjective we include it the final list of the semantic domain members in the target language.

3.1.2 Machine-readable dictionaries

The second way of questionnaire translation is based on bilingual machine-readable dictionaries. For our experiments, we relied on dictionaries provided by the web-services Freedict.org and Yandex.

Freedict dictionaries are stored in TEI format[3], an XML format standard for text data. The structure of the TEI format is constructed by paired tags as required by XML conventions. Each dictionary entry is distinguished by the <entry> tag, with the <form> and the <sense> tags inside it. Several entries of the <sense> tag are allowed, and their number depends on the amount of translation equivalents for the lexeme in question. Possible translations are enumerated: equivalents for the basic meaning of a lexical item come first, followed by possible translations for the word in question used in figurative contexts. The attribute “n” of the tag <entry> indicates the equivalent’s number.

We translate adjectives under the following scheme. For every adjective from a questionnaire we find its entry, extract its first translational equivalent(s) (indexed by *n="1"*) and translate it backwards using the corresponding reverse dictionary to exclude irrelevant candidates (compare a similar strategy reported in the previous section). We include in the final list of adjectives only those items whose backward translation matched with some of the adjectives from the initial set in the source language.

We translate nouns almost the same way, extracting from the source-target dictionary the candidate for the sense number 1 and omitting the step of backward translation. If there is no source noun in the Freedict dictionary we also consult the Verdict dictionary. This latter dictionary has more simple structure. Each entity takes one row and has four features splitted by tabulation: target word, its part of speech, its translation to the source language and the link to the source dictionary.

3.1.2 Parallel corpora

[3] URL: [http://www.tei-c.org/index.xml]

For this method, we used sentence-aligned tagged bilingual corpora from the Russian National Corpus (English-Russian and Russian-Italian). The English-Russian corpus contains 527,782 sentences, the Russian-Italian corpus contains 101,814 sentences.

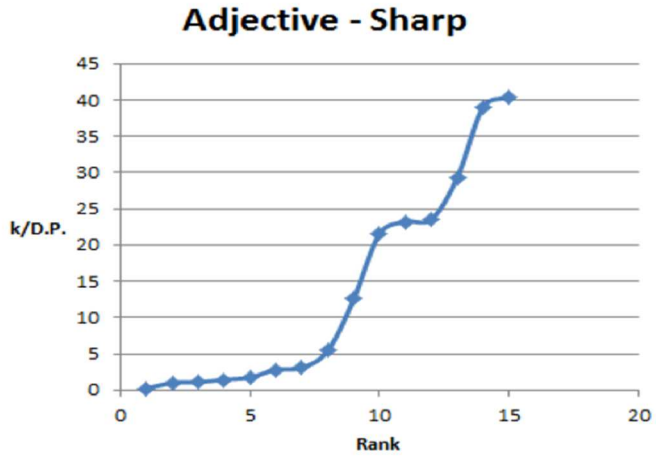
We identify translation candidates with the help of POS tags and sentence alignments based on the *dependency measure* shown in Formula 1.

$$Dependency\ Measure = \frac{P(target\ adjective \mid source\ adjective)^2}{P(target\ adjective) * P(source\ adjective)}$$

where P(A|B) is the probability of the presence of adjective A in the aligned target sentence given adjective B in the source sentence.

Formula 1. Dependency Measure

Dependency Measure for the 2nd Language Adjective was calculated against all the 1st Language Adjectives, which serve as its potential translation equivalents. They can be then ranked according to the dependency measure. If we plot the rank of the L2 adjectives against their 1/DM values for a given L1 adjective, we usually observe a gradual decrease in DM followed by an abrupt rise in slope after a particular rank. It turns out that all the translations before the abrupt rise of 1/DM actually are the possible translations for a particular adjective.



Picture 1. 1/DM vs. DM ranking for Hindi equivalents of the English *sharp*.

To increase the statistical data given the limited size of our corpora, we lemmatize morphological variants of English adjectives including comparative and superlative forms and productive adverbs in *-ly*. Nouns are translated via the same method.

3.2 Filling the questionnaire

After we get the questionnaire translated in some language, we need to check which of the all theoretically possible noun-adjective combinations are indeed legitimate. We test two different methods completing this task. The first one relies on online translators, and the second one is based on monolingual corpora.

3.2.1 Online translators

Similarly to the translation task, we use APIs for two cloud translators, Yandex.Translate and Google Translate. For each potential noun phrase of the type “adjective + noun” in the target language we perform forward translation into some intermediary language (we tried to use English, Russian, French and German for this purpose), and then we translate the result in backward direction. As the final step, we check if the backward result and the source noun phrase are equal, and if they are, they are considered as compatible.

3.2.2 Monolingual corpora

We realize this method on the basis of monolingual WaCky corpora (Baroni et al. 2009), which have the .xml format tag-structure. Each sentence is packed in an <s> tag as a list of words with several features (word form, lexeme, morphological annotation, its number in the sentence, number of its syntactic head, and its syntactic category separated by tabulation), each word on its row. Thus we search in the corpus of the target language all possible lexeme bigrams “adjective + noun” from the questionnaire.

To exclude occasional collocations which are not relevant for the target language we use Positive Pointwise Mutual Information measure (see Formula 2). One can interpret Positive Mutual Information value of two lexemes as a degree of statistic significance of their co-occurrence.

$PMI(x,y) = \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$, where

p(x,y) - probability of co-occurrence of adjective X and noun Y in corpus,
p(x) - corpus-estimated probability of occurrence of the adjective,
p(y) - corpus-estimated probability of occurrence of the noun.

PPMI means only positive values of PMI:

$PPMI(x,y) = PMI(x,y)\ if\ PMI(x,y) > 0\ else\ 0.$

Formula 2. Positive Pointwise Mutual Information Measure.

Results

We evaluate our questionnaire translation and filling results on precision and recall. As soon as we need all possible adjective equivalents and only one-to-one translation of nouns both measures are used for adjectives and precision only for nouns. Say A set of adjectives given by gold standard - and B - set of translated adjectives using some algorithm. Then one could count accuracy as a number of words in B that also belong to A and recall as a number of words in A that belong to B as well.

		Freedict	Google translator	Parallel corpus
en-ru	P:	0 / 0 / 0.5 / 0	0 / 0.3 / 0.67 / 1	0.5 / 0.385 / 0.27 / 0.67
	R:	0 / 0 / 0.33 / 0	0 / 0.19 / 0.36 / 0.2	1 / 1 / 1 / 0.67
ru/en-it	P:	1 / 0 / 0.5 / 0.66	1 / 0.55 / 1 / 0.67	1 / 0.5 / 0.5 / 0.5
	R:	0.33 / 0 / 0.16 / 0.4	0.08 / 0.26 / 0.3 / 0.14	0.17 / 0.09 / 0.375 / 0.25

Table 2. Adjective translation precision (P) and recall (R) for semantic fields ‘sharp’ / ‘smooth’ / ‘thick’ / ‘thin’ respectively.

		Freedict	Yandex. Translate	Google translator	Parallel corpus
en-ru	P:	0.38 / 0.5 / 0.67	0.61 / 0.59	0.80 / 0.68	0.393 / 0.417 / 0.434
ru/en-it	P:	0.31 / 0.59 / 0.5	0.58 / 0.56	0.64 / 0.63	0.5 / 0.396 / 0.404

Table 3. Nouns translation precision (P) for semantic fields ‘sharp’ / ‘smooth’ / ‘size’[4] respectively.

Using Google translator for translation of adjectives gives quite similar and not really perfect results as using the Freedict dictionary while the algorithm based on parallel corpora seems to take into account more adjectives and has a more stable recall.

[4] The same list of nouns constitutes the questionnaire for the domains ‘thick’ and ‘thin’, that is why we merge these fields here into a single test dataset labelled ‘size’.

Comparatively to adjectives nouns are translated best of all with online translators. The main problem with Freedict translation of nouns is the small size of its lexicon. Parallel corpus doesn't show any good results there which means that one better uses different algorithms for different tasks.

		Monolingual corpus	Yandex. Translate	Google translator
en-ru	P:	0.75 / 0.79	0.65 / 0.69	0.43 / 0.65
	R:	0.87 / 0.92	0.68 / 0.75	0.55 / 0.66
ru/en-it	P:	0.4 / 0.54	0.64 / 0.55	0.75 / 0.79
	R:	0.7 / 0.68	0.35 / 0.5	0.32 / 0.37

Table 4. Filling questionnaires precision (P) and recall (R) for semantic fields 'thick' / 'thin' respectively.

As for the quality of the questionnaire population, Yandex and Google translators demonstrate comparable results, while the method based on monolingual corpora analysis outperforms both of them in all cases, except for precision value obtained on the Italian dataset.

5. Discussion

In this research, we tested several algorithms of an automatic data collection for lexical typology, namely three methods of questionnaire translation and two methods of its filling. The translation task appears to be quite challenging: the values of Precision and Recall are rather low for all algorithms suggested. On the contrary, results of the second task, questionnaire completion, seem to be much more promising; it means that at least these algorithms can be already incorporated in the process of lexical typological research.

It is clear that in both cases the result naturally depends on the resource size and quality. We suppose that the algorithm of questionnaire population based on the data from monolingual corpora demonstrates a very good performance because of the reasonable corpora sizes. Similarly, online translators complete this task for some language pairs much better than for

the others. For example, the Yandex.Translate system works more efficiently with the Russian data, exactly as expected.

We suppose that our translational algorithms' performance could also improve in case we enlarge and improve the initial resource database. The results that we get on the data from parallel corpora supports this hypothesis: the usage of the bigger English-Russian corpus results in a better algorithm's performance than the usage of the much smaller Russian-Italian collection. Fortunately, all the resources that we use are constantly updated. Additionally, there is also room for algorithms' improvement: for example, one can choose more frequent or more appropriate nouns with the help of distributional semantic modelling.

References

Berlin, B. and Kay, P. (1969). Basic color terms: their universality and evolution. Berkeley, CA: University of California Press.

Bolinger D. (1967). Adjectives in English: attribution and predication // Lingua 18, 1-34.

Eger, S. (2012). Lexical semantic typologies from bilingual corpora: a framework. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 90-94). Association for Computational Linguistics.

François, A. (2008). Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In: Vanhove, M. (ed.) From Polysemy to Semantic change: Towards a Typology of Lexical Semantic Associations. Studies in Language Companion Series, 106, 163-216.

Goddard, C. and Wierzbicka, A. (2007). NSM analyses of the semantics of physical qualities: sweet, hot, hard, heavy, rough, sharp in cross-linguistic perspective. Studies in Language, 31(4), pp. 765–800.

Kashkin E., Vinogradova O. (to appear). The domain of surface texture // E. Rakhilina, T. Reznikova (Eds.). *The Typology of Physical Qualities*. John Benjamins Publishing Company.

Key, M. R., & Comrie, B. (2007). Intercontinental dictionary series. Online Version: http://lingweb.eva.mpg.de/cgibin/ids/ids.pl.

Koptjevskaja-Tamm, Maria, ed. (2015). The linguistics of temperature. Vol. 107. John Benjamins Publishing Company.

Koptjevskaja-Tamm K., Rakhilina E., and Vanhove M. (2016). The semantics of lexical typology. In: Rimer, Nick (ed.). The Routledge Handbook of Semantics, 434-454.

Kozlov, A., Privizentseva, M. (to appear). Typology of dimensions // E. Rakhilina, T. Reznikova (Eds.). *The Typology of Physical Qualities*. John Benjamins Publishing Company.

Kruglyakova, V. (2010). Semantics of rotation verbs in a typological perspective. [Semantika glagolov vraščenija v tipologičeskoj perspektive]. Ph.D. thesis, Russian State University for Humanities.

Kyuseva, M., Parina, E., Ryzhova, D. (to appear). Methodology at work: semantic fields 'sharp' and 'blunt' // E. Rakhilina, T. Reznikova (Eds.). *The Typology of Physical Qualities*. John Benjamins Publishing Company.

Kyuseva, M., Reznikova, T., & Ryzhova, D. (2013). A typologically oriented database of qualitative features [Tipologicheskaya baza dannykh ad''ektivnoy leksiki]. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog (pp. 419-430).

Lander Yu., Maisak T., and Rakhilina E. (2012). Verbs of aquamotion: semantic domains and lexical systems, in: Motion Encoding in Language and Space, 67-83.

List, J.-M., Mayer, Th., Terhalle, A., and Urban, M. (2014). CLICS: Database of Cross-Linguistic Colexifications. Marburg: Forschungszentrum Deutscher Sprachatlas (Version 1.0, online available at http://CLICS.lingpy.org, accessed on 2018-2-19)

Majid, A. and Bowerman, M. eds. (2007). ''Cutting and breaking" events: A crosslinguistic perspective. [special issue] Cognitive Linguistics, 18(2).

Majid, A., Enfield, N.J. and van Staden, M. eds. (2006). Parts of the body: Cross-linguistic categorisation. [Special issue]. Language Sciences, 28(2-3).

Narasimhan, B. and Kopecka, A. eds. (2012). Events of 'putting' and 'taking': A crosslinguistic perspective. Amsterdam, Philadelphia: John Benjamins.

Rakhilina, E., & Reznikova, T. (2016). A Frame-based methodology for lexical typology. The Lexical Typology of Semantic Shifts, 58, 95-130.

Reznikova T., Rakhilina E., and Bonch-Osmolovskaya A. (2012). Towards a typology of pain predicates. In: Linguistics, 50(3), 421–465.

Ryzhova, D., Kyuseva, M., & Paperno, D. (2016). Typology of Adjectives Benchmark for Compositional Distributional Models. In Proceedings of the Language Resources and Evaluation Conference (pp. 1253-1257). European Language Resources Association (ELRA).

Wälchli, B., Cysouw, M. (2012). Lexical Typology through Similarity Semantics: Toward a Semantic Map of Motion Verbs. De Gruyter, 50.3: 671–710.

Wierzbicka, A. (1984). Cups and mugs: Lexicography and conceptual analysis. Australian Journal of Linguistics, 4(2), pp. 205-255.

Youn H., Suttond L., Smith E., Moore C., Wilkins J. F., Maddieson I., Croft W., and Bhattacharya T. (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.