Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

The Universal Database for Lexical Typology

Ekaterina Voloshina
University of Gothenburg
Gothenburg, Sweden
vokat@mail.ru

Polina Leonova
HSE University
Moscow, Russia
00leonovapolina@gmail.com

Abstract

The paper presents the principles of creating a database for research in lexical typology and describes the possibilities of its use as a linguistic resource. The database is built around semantic fields and frames, i. e. units of analysis in the frame-based theory of lexical typology.

The database provides a universal format for storing the data; therefore, any project in lexical typology can be easily added. The database does not only store the data from previous research projects but allows anyone who wants to contribute to submit data via its web interface. The database includes examples provided by native speakers and manually annotated with translations, semantic fields, and frames, following the annotation principles adopted within the frame approach to lexical typology.

Keywords: lexical typology, corpus linguistics, computational lexicography
DOI: 10.28995/2075-7182-2023-22-1133-1140

База данных для лексико-типологических исследований [1]

Екатерина Волошина
Гётеборгский университет
Гётеборг, Швеция
vokat@mail.ru

Полина Леонова
НИУ ВШЭ
Москва, Россия
00leonovapolina@gmail.com

Аннотация

В статье представлены принципы создания базы данных для исследований в области лексической типологии и описаны возможности ее использования в качестве лингвистического ресурса и инструмента для сбора и анализа материала. База данных построена на основе семантических полей и фреймов, т.е. единиц, на которых основан лексико-типологический анализ в рамках фреймового подхода.

База предполагает универсальный формат хранения данных, поэтому любой проект по лексической типологии может быть легко в нее добавлен. База данных не только содержит материал предыдущих исследовательских проектов, но и позволяет любому желающему внести новые данные, используя специально разработанный веб-интерфейс. В базе хранятся примеры, полученные от носителей языка и аннотированные вручную: для каждого примера приводится его перевод на русский язык, семантическое поле, к которому относится иллюстрируемая лексическая единица, и соответствующий примеру фрейм.

Ключевые слова: лексическая типология, корпусная лингвистика, компьютерная лексикография

[1]Статья подготовлена в ходе проведения исследования № 23-00-012 «Смежность семантических полей в типологической перспективе» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)

1 Introduction

The cross-linguistic analysis of lexicon usually requires both a sufficient amount of data and adequate annotations to it. For semantic research, as for linguistic research in general, it is convenient to structure and store this information in databases (in our case - lexical databases). However, as opposed to phonetic or morphological databases, the lexical ones inevitably reflect theoretical assumptions of their developers. In particular, FrameNet (Baker et al., 1998) represents the frame approach to semantics, constructicons (Lyngfelt et al., 2018) provide data annotated in the tradition of Construction Grammar (Fillmore and Kay, 1995), etc. Even CLICS (Rzymski et al., 2020), which seems to be theory neutral, relies on a certain categorisation of semantic domains and concepts within them, and thus selects just one research solution out of many available. Indeed, a detailed analysis of linguistic data within the framework of Moscow Lexical Typology Group (MLexT) [2] results in other sets of categories for the same semantic domains. The fact is that CLICS automatically aggregates data from available wordlists, whereas the MLexT approach involves a manual and data-driven detection of relevant concepts, or, in terms of MLexT, frames (for details, see (Rakhilina and Reznikova, 2016)).

In this paper, we present a database which incorporates the principles of MLexT approach and implements several functions related to data collection, aggregation of primary data, editing the possible set of questions, and data analysis. For data collection any researcher can add their data in a unified format filling a questionnaire and, after moderation, the data are added directly to the database and become available for search. Besides that, moderators can edit questionnaires and uploaded data, as well as add general descriptions and observations related to semantic fields. The database includes a search engine with two modes to cover all possible queries for the purposes of research in both lexical typology and semantics in general.

2 Lexical typology: theory and methods

For a long time lexicon has not been studied from a typological perspective. The first attempts in this area cover just a few semantic domains, namely, color, kinship and body-part terms (Berlin and Kay, 1969; Keesing, 1975; Andersen, 1978). Over the last decades, however, the scope of lexical typology has expanded to encompass a whole range of domains, cf. verbs of cutting and breaking (Majid et al., 2007), verbs of motion, e.g. motion in water (Majsak, 2007), rotation (Krugljakova, 2010), falling (Reznikova et al., 2020), pain predicates (Reznikova et al., 2012), verbs of putting and taking (Kopecka and Narasimhan, 2012), physical qualities (Koptjevskaja-Tamm, 2015; **?**), etc. These studies differ in methodology of data collection and analysis. The approach we follow here, as already stated, is the one developed by MLexT, as it turns to be more versatile than other techniques in this area (for details, see (Rakhilina and Reznikova, 2016)).

Within the MLexT approach, semantic domains are compared in terms of frames, i.e. prototypical situations that are expressed through words of a given semantic field.

The process of establishing the frames starts with browsing through corpora and dictionaries. As a result, relevant contexts for each frame and parameters of their opposition are identified. For example, for verbs of motion in water a relevant parameter would be the opposition of active and passive motion, for verbs of rotation – the type of axis (inner or outer) and the type of rotating object, for verbs of change – the degree of change (full versus partial change). It does not mean that these parameters should be relevant for every language: frames that are lexically opposed in one language can be colexified in another. Hypotheses generated on this stage become the basis for a context-based questionnaire consisting of a list of sentences with gaps which are supposed to be filled with a lexeme from the field under study. The number of sentences is determined by the number of possible oppositions which are relevant for the field.

Questionnaires are used for verifying data collected from corpora and dictionaries, and for collecting data for the languages that do not have such resources. This method of simultaneous work with corpora and native speakers for each language allows us to collect data from various languages, including endangered ones.

[2] `https://lextyp.org`

Language	Field	Frame	Context	Verb	Example	Translation	Source
Spanish	change	partial change of an object	It was obvious that something must be ___ in the legislation.	cambiar	No hay planes de cambiar la legislación en ese sentido, al menos a corto plaz	Nikakih izmenenij v zakonodatel'stvo v jetoj oblasti, po krajnej mere v kratkosrochnoj perspektive, ne planiruetsja.	speakers
French	change	full change of an object	He ___ linoleum for parquet.	changer	Il a changé le linoléum contre le parquet.	On pomenjal linoleum na parket.	speakers
Karel	change	partial change of an object	The chef was angry with the cook because he ___ the recipe.	muuttua	"muuttele syömistä, elä syö yhtä ta samua	Raznoobraz' edu, ne esh' odno i to zhe	corpus

Table 1: An annotation example.

At the same time, every study conducted in this framework involves different language experts dealing with various resources. The database storing the up-to-date version of the questionnaire and the skeleton of the field structure would ensure compatibility of data coming from different languages and facilitate the process of data collection and analysis.

3 Database

3.1 Annotation

The database includes the data from questionnaires filled in by native speakers or examples from corpora. Each example is translated to Russian and annotated in the following way:

- **Language**: the original language of the example;
- **Semantic field**: a generalized part of lexicon the example belongs to, e.g., 'falling', 'hiding', 'changing', etc.;
- **Frame**: one of the prototypical meanings included in a semantic field, e. g. the semantic field of falling can be divided into four main frames: 'loss of vertical orientation', 'detachment', 'falling from elevated surface', and 'crashing down' ((Reznikova et al., 2020));
- **Context**: if the example comes from a questionnaire, context is a given stimulus; if the example is taken from a corpus, it is the closest context to the one in the example;
- **Verb**: a verb that belongs to a given semantic field and is used in the example;
- **Source type**: if the example comes from a corpus or was given by a native speaker.

The annotation examples are given in Table 1.

3.2 Database Structure

The Mongo database is built in a way to represent the annotation described above. It is essential that users can search for any word or word part in examples and their translations; therefore, a non-SQL Mongo database was implemented.

The database schema is represented in Figure 1. The database includes several ways of searching and aggregating the data: two search modes, creating semantic maps and semantic field profiles (for details see Section 4). Some search modes use different types of information (e.g. one can only use fields and languages). To make the search process more efficient, the database is decomposed into several tables instead of storing all information in one.

The database includes five collections: *contexts, fields, frames, languages,* and *verbs*. In languages collections, an id and the text name of an item are given. Fields collection, in addition to an id and a text name, includes an overall description of this field. Frames collections are connected with fields, as they belong to one field. Contexts include the list of frames, since they can belong to several frames. Verbs collection includes a verb and the language it is taken from and a list of examples where each example's annotation includes the example itself, the translation, the source type, and the context id, which connects the example to frames and semantic fields.

```
{
   "contexts":{
      "_id": ObjectId,
      "context": string,
      "frame": string
   },
   "fields":{
      "_id": ObjectId,
      "field": string,
      "description": string,
   },
   "frames":{
      "_id": ObjectId,
      "frame": string,
      "field": ObjectId
   },
   "languages":{
      "_id": ObjectId,
      "lang": string
   },
   "verbs":{
      "_id": ObjectId,
      "verb": string,
      "lang_id": ObjectId,
      "examples":{
         "example": string,
         "translation": string,
         "source": string,
         "context": ObjectId
      }
   }
}
```

Listing 1: The database schema

4 Web-interface and Usage Cases

The database has the user interface which allows for working in three different scenarios: as a data collector, as a database editor, and as a researcher. Therefore, our database covers all stages of working with data: collection, processing and analysis.

As for technical details, the web-interface is written as a web application in Python on the basis of the Flask framework [3] and Mongo-DB API. [4]

4.1 Scenario I: Data Collection

One of the most important parts of research in lexical typology is data collection. The data is collected through a questionnaire which consists of a set of contexts related to the given semantic field (for an example, see Appendix 5). To make the answers to the questionnaire comparable, it is important not only to store them in the same format but also to collect them in a similar setup.

The database supports two main forms of submitting the data: via manual form editing or uploading the data in the table format (`csv` or `xlsx`). In the first case, a researcher chooses a project to which they want to contribute data and then the form with questions is generated automatically based on the pre-uploaded questionnaires. Every question requires an example with a translation and allows to add a comment. Additionally, a researcher can put some extra examples in the free form.

The second option is to upload a spreadsheet that must follow the template with the questionnaire to be further processed.

The data is uploaded automatically to the database after it is approved by a moderator to exclude data in wrong formats or a scam data since the form can be filled without logging into the database.

4.2 Scenario II: Database Editing

As mentioned above, the data could be added by any user, therefore, only few people can get access to the database: to add, edit or delete files. It is also important that the moderator would have an expertise in lexical typology, so they could exclude non-relevant examples from the filled questionnaires. Besides that, a moderator might fill in information about frames, as frames are usually determined empirically, on the basis of collected data.

The moderator checks that all the fields are filled correctly and the data can be uploaded to the database. Before uploading to the database, the data is stored as files on the server, which can be edited through the web-interface.

Moreover, the moderator can edit or delete data from the database if the data appears to be outdated. Besides that, the editor's functions include adding descriptions about projects, and in the future they will be expanded so editors could add meta-information about languages in projects (for example, to specify the type of lexicalisation system in a given language).

4.3 Scenario III: Research

The main instrument for the research is the database search engine. There are two main search modes: the full text search supported by MongoDB and search by filters. The full-text mode allows searching by words or word parts in examples and translations. The filters are made to query the database annotation, therefore, there are 6 main search criteria that should be selected from a list: semantic field, language, frame, context, verb, and source type. Both types can be combined.

Besides the search engine, the database allows to build automatic semantic maps (Haspelmath, 2003) that aggregate information uploaded into the database. Semantic maps represent how different frames are connected within the semantic field. In the graph, nodes are frames or contexts (depending on the desired granularity), and they are connected if there is a verb that can be used in both frames (or contexts). The frequency of such connections are edge weights.

However, fully connected graphs, or vacuous maps, are less informative, as the more edges a graph has, the less combinations of senses within one lexeme it excludes.

[3] `https://flask.palletsprojects.com/en/2.0.x/`

[4] `https://pymongo.readthedocs.io/en/stable/index.html`

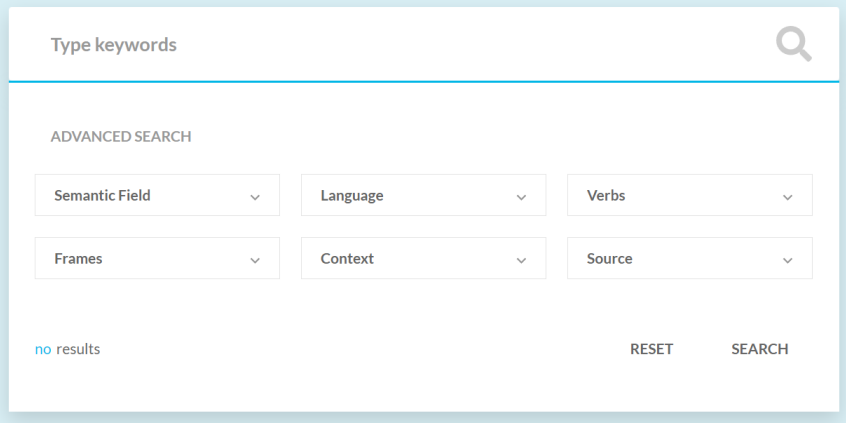

Figure 1: Web-interface for queries

Therefore, semantic maps are maximum spanning trees built with Kruskal's algorithm on the basis of uploaded examples. In other words, instead of graphs built on the basis of the data, semantic maps are subsets with the greatest weight and the minimum number of edges.

While building a semantic map, it is possible to determine a subset of languages that will be taken into account.

Moreover, the database includes subpages for all uploaded projects. Every project corresponds to one semantic field. The description is added manually by a moderator while the other information, such as frames constituting this field, languages and related projects are added automatically on the basis of uploaded examples.

As an example of the possible research, the base can be used for solving problems that are relevant for lexical typology and lexical system theory in general. For instance, it can be used to find out which factor makes the greater contribution to the type of lexical system – areal or genealogical.

5 Conclusion

In this paper, we present a database for research in lexical typology. While previous databases are built specifically for certain semantic fields [5], the database described in this paper is universal in that it unifies the data from different semantic fields. It is especially necessary since the number of projects has increased significantly. Storing all the data in the same database allows to solve new research questions, e.g. which fields are related to each other in the cross-linguistic perspective, or what has a greater impact on the degree of similarity between lexical systems - their genetic or areal closeness.

From the practical point of view, the database is meant to make the process of data collection easier due to the possibility of data contribution and universal format of data storage. The database will be published on `https://linghub.ru/` but it can already be accessed through local interface published here: `https://anonymous.4open.science/r/LexTypDB-4C8C`.

References

Elaine S Andersen. 1978. Lexical universals of body-part terminology. *Universals of human language*, 3:335–368.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. // *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Brent Berlin and Paul Kay. 1969. *Basic color terms: Their universality and evolution*. University of California Press.

[5] `http://www.web-corpora.net/zvukimu/`, `https://linghub.ru/aquamotion/`

Charles J Fillmore and Paul Kay. 1995. Construction grammar. *Language*, 64(501-538):30.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. // *The new psychology of language*, P 217–248. Psychology Press.

Roger M Keesing. 1975. *Kin groups and social structure*. Holt, Rinehart and Winston New York.

Anetta Kopecka and Bhuvana Narasimhan. 2012. *Events of putting and taking: A crosslinguistic perspective*, volume 100. John Benjamins Publishing.

Maria Koptjevskaja-Tamm. 2015. *The linguistics of temperature*, volume 107. John Benjamins Publishing Company.

V. A. Krugljakova. 2010. *Semantika glagolov vraščenija v tipologičeskoj perspektive*. Ph.D. thesis, (RGGU).

Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent. 2018. *Constructicography: Constructicon development across languages*, volume 22. John Benjamins Publishing Company.

Asifa Majid, Melissa Bowerman, Miriam van Staden, and James S Boster. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective.

Raxilina E. V. Majsak, . . 2007. Glagoly dviženija v vode: leksičeskaja tipologija. verbs of motion in water: lexical typology.

Ekaterina Rakhilina and Tatiana Reznikova. 2016. A frame-based methodology for lexical typology. *The lexical typology of semantic shifts*, 58:95–129.

Tatiana Reznikova, Ekaterina Rakhilina, and Anastasia Bonch-Osmolovskaya. 2012. Towards a typology of pain predicates. *Linguistics*, 50(3):421–465.

TI Reznikova, EV Rakhilina, and DA Ryzhova. 2020. Verbs of falling in the languages of the world: Frames, parameters, and types of the systems. *RUSSIAN ACADEMY OF SCIENCES*, P 10.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):13.

Appendix

A. The Sample Questionnaire

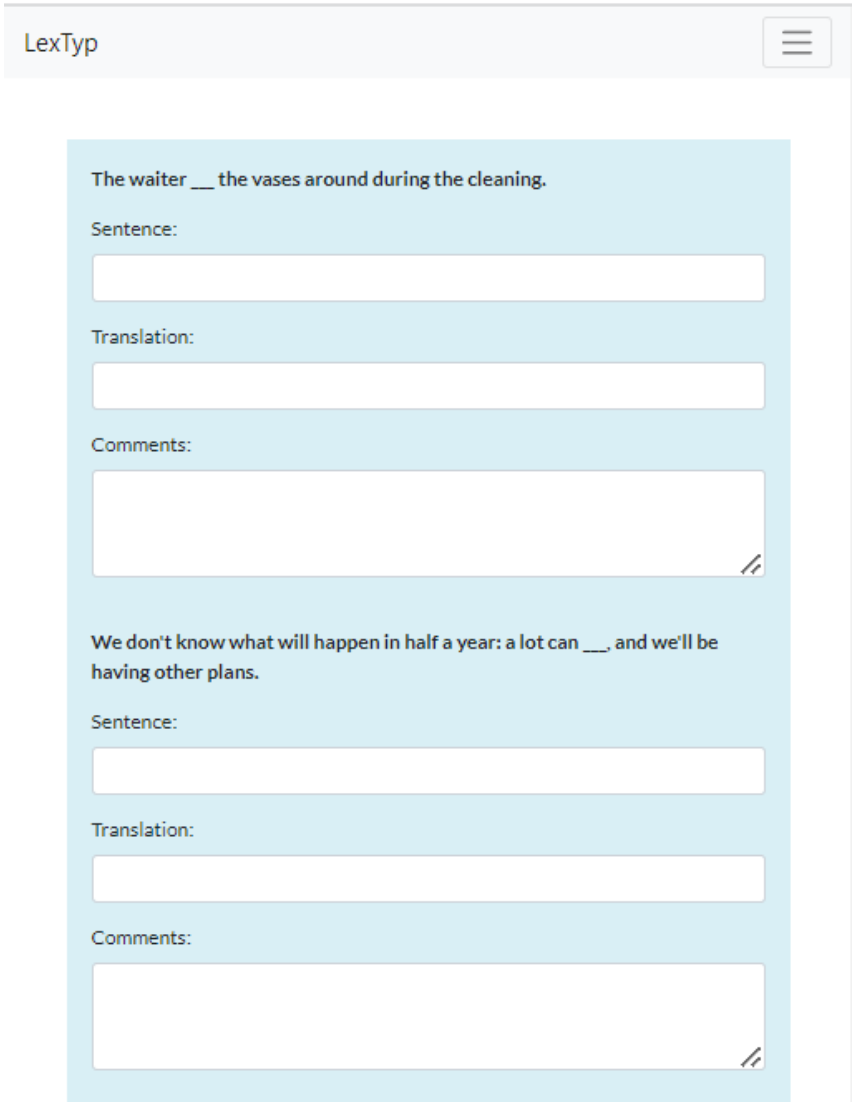

Figure 2: Web-interface for queries