Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Representing the Database of Semantic Shifts by Zalizniak et al. from 2024 in Cross-Linguistic Data Formats

Katja Bocklage[1], Anna Di Natale[1], Annika Tjuka[2], Johann-Mattis List[1]
[1] Chair of Multilingual Computational Linguistics, University of Passau, Passau
[2] Department of Linguistic and Cultural Evolution, MPI for Evolutionary Anthropology, Leipzig

In this brief study, we show how the Database of Semantic Shifts, a large resource on semantic change and semantic motivation, can be represented in Cross-Linguistic Data Formats. The representation allows for a convenient quantitative analysis of the numerous annotations on semantic change and semantic motivation and for the integration of the database with additional resources on semantic change and semantic motivation that have been compiled independently in the last years.

1 Introduction

Although scholars have emphasized for a long time that patterns of semantic change and semantic motivation can be surprisingly similar across unrelated languages, there are not many attempts to prove this by compiling detailed collections that list attested cases of semantic change or semantic motivation in the languages of the world. The Database of Semantic Shifts by (Zalizniak et al. 2024, http://datsemshift.ru) is a notable exception that has considerably grown in recent times. In order to allow us to integrate this resource with additional cross-linguistic semantic resources that have been created over the past years, we have now created a workflow that allows us to convert the Database of Semantic Shifts into Cross-Linguistic Data Formats (CLDF, Forkel 2018, https://cldf.clld.org).

2 Background

The development of the Catalogue of Semantic Shifts started in 2002 and continues to this day. Since 2002, the group of Maria Bulakh, Dmitry Ganenkov, Ilya Gruntov, Timur Maisak and Maxim Russo have joined Anna Zalizniak in her work at the Institute of

Computer-Assisted Language Comparison in Practice
Volume 7, Number 1
URL: https://calc.hypotheses.org/7060 DOI: 10.15475/calcip.2024.1.4

Published on 24/04/2024

Linguistics of the Russian Academy of Sciences. It was first made publicly available in 2013 (http://semshifts.iling-ran.ru/) and received a new web address (http://datsemshift.ru/) in the context of a revision in 2018 (Zalizniak 2018: 771).

The catalogue is neither the first nor the only attempt to collect instances of semantic change. Zalizniak et al. (2012: 637–638) point to the comparative Indo-European dictionary of Buck (1949) and to two proposals, one for a comparative dictionary of semantic change (Schropfer 1956) and one for a semasiological dictionary of Indo-European languages (Trubachev 1964 [2004]), as well as to various published efforts to establish regularities of semantic change — such as Blank (1997), Sweetser (1990), Traugott & Dasher (2002) and others — as inspiration sources for their own work. By now, with more than 8000 different instances of semantic shift in the most recent version (3.0, obtained on February 5, 2024) the Catalogue of Semantic Shifts largely exceeds other collections in scope.

The catalogue defines the concept of semantic shift as a specific relation between two meanings that is realized in concrete processes of semantic change or word formation.

> The Catalogue of Semantic Shifts has clearly defined basic principles, its primary notions being those of a semantic shift, which is understood as a relation of cognitive proximity between two linguistic meanings, and a realization of a semantic shift, i.e. one polysemic word or a pair of cognate words that act as "exponents" of this relation. (Zalizniak 2018: 781)

Five different types of semantic shifts are presented: Synchronic Polysemy (1) occurs when two meanings can be expressed with the same word, such as French *femme* for WIFE and WOMAN (Zalizniak et al. 2012: 634). Word pairs are classified as Morphological Derivation (2) if either a morphological derivative of a word with meaning A can express meaning B such as Italian contare "to count" and raccontare "to narrate" (Zalizniak et al. 2012: 635) or vice versa or if grammatical variations of the same word carry different meanings such as Spanish *celo* "zeal, fervency" and *celos* (plural form) "jealousy" (Zalizniak 2018: 774). Diachronic Semantic Evolution (3) describes the semantic shift between an ancestor variety to a descendant variety, such as Latin demoror "to delay" → French demeurer "to live" (Zalizniak et al. 2012: 634). Cognates (4) are words of a language or of two sister languages that share the same root, such as German *Zahl* "number" and English tale (Zalizniak 2018: 774). A Borrowing (5) takes place if a word with meaning A in one language is adapted into another language with a different meaning B, such as Old French *plain* "flat, smooth" → English *plain* "simple" (Zalizniak et al. 2012: 635).

The goal of the Catalogue of Semantic Shifts is to collect semantic shifts that reoccur across the world's languages in order to provide the data necessary to study universal tendencies of semantic change and semantic motivation. The underlying assumption is that most semantic shifts reflect rather general cognitive principles and are therefore

likely to recur across languages and times. By collecting as many shifts as possible, the shifts showing the strongest universal tendencies should reveal themselves with growing amounts of data. In this form, the Catalogue of Semantic Shifts can be used to uncover cross-linguistic patterns of conceptualization and in this way provide concrete help in semantic reconstruction and etymological studies (Zalizniak et al. 2012: 634).

3 Representing DatSemShift in CLDF

In order to represent the data underlying the Database of Semantic Shifts in Cross-Linguistic Data Formats, we followed the basic workflow for CLDF conversion outlined and developed in earlier studies (List 2021, List et al. 2022), making use of the CLDFBench package (https://pypi.org/project/cldfbench, Forkel and List 2020). Given that the scope and the structure of the database differs from lexical datasets that we have previously modeled in CLDF, the steps to create the CLDF version of the database are not identical with the steps used in previous examples for data conversion.

The steps to obtain the CLDF data and to run the CLDFBench code, however, are the same as with other CLDF dataset. Downloading can be best done via GIT, since the code is curated on GitHub. Thus, via the command line, you can retrieve the DatSemShift CLDF package as follows.

```
$ git clone https://github.com/lexibank/datsemshift.git
$ cd datsemshift/
```

To install the code that you need to run the CLDF conversion but also to convert the data to other formats (as we will show in the analysis examples section), you should create a fresh virtual Python environment and can then install all necessary packages with the help of the pip command.

```
$ pip install -e .
```

3.1 Downloading the Data

In order to do access the data provided by the Database of Semantic Shifts at a given point in time, we created a download command within the CLDFBench application. To trigger this command, one has to change the value of the variable DOWNLOAD in the beginning of the main Python script (lexibank_datsemshift.py) from False to True. Once this has been done, typing the basic download command provided by CLDFBench will download the data directly from the website.

```
$ cldfbench download lexibank_datsemshift.py
```

The download accesses the HTML representation of the database through the website and then uses regular expressions in order to identify the relevant data points, which are stored in individual tables in TSV format that are available from the raw folder of the CLDF repository (available at https://github.com/lexibank/datsemshift). The raw HTML files themselves are not shared in the repository.

3.2 Concept Mapping

The data underlying the CLDF version reported here, was downloaded on February 5 from the website and stored in the automatically prepared tabular format. This dataset contains a total of 4583 semantic glosses, representing all concepts between links of semantic shifts are identified in the database. While there are a few cases where glosses seem to point to identical concepts, we managed to identify them quickly when mapping the data systematically to the Concepticon (https://concepticon.clld.org, List et al. 2024, Version 3.2).

As a starting point we used already available Concepticon mappings that were made for a previous version of the Catalogue of Semantic Shifts (Zalizniak et al. 2020, Version 2.0). These mappings had been checked back in 2020 in the typical peer review process that we have established for the Concepticon project, resulting in the Concept list Zalizniak-2020-2590, which contained also a rudimentary representation of the underlying semantic network. Equipped with these initial mappings that had survived a rigorous peer review process in which several linguists had been involved, we went through an additional round of proposing and discussing Concepticon mappings by first inferring new mappings automatically for the new concepts that had not been available in the version from 2020, using the PySem library (List 2021, https://pypi.org/project/pysem, Version 0.8). We then loaded the data in an online spreadsheet and worked in collaboration to check all automated mappings and to double check the mappings that had been carried out earlier with a team of three reviewers working in collaboration, discussing all ambiguous cases. The new concept list was added to Concepticon's Version 3.2 which was published in March 2024.

3.3 Language Mapping

Glottocodes that link to the Glottolog reference catalogue (Hammarstrom et al. 2024, https://glottolog.org, Version 5.0) are provided in the original database for almost all shifts. When creating a list of languages with their corresponding Glottocode, we only had to correct a couple of codes that were corrupted. The full list is available in the folder etc/languages.csv in the repository.

3.4 Network Representation of Semantic Shifts in Tables

We introduced a new way to represent network data with Concepticon 3.2. This format makes use of the fact that JSON can be used inside of CSV tables, which allows us to

represent a complex directed or undirected network with edge and node attributes in a single CSV table (see Bocklage et al. 2024) for details. The major concept network underlying the Database of Semantic Shifts is represented in the Parameter Table of the CLDF data (file cldf/parameters.csv). This table contains 4228 rows for all 4228 out of 4583 concepts for which valid shifts reflected in Polysemy (1) and Derivation (2) could be identified, ignoring the shifts belonging to the other categories for the time being. Shifts are divided into two categories: those, where the original data indicates directions, and those where no directions are given.

As an example for this format, consider the following JSON representation that can be found as part of a larger list of links in line 5 (concept number 4) in the Parameter Table for the concept “abdomen” (linked to 1251 BELLY in Concepticon) in the column Linked_Concepts (which represents relations where no direction has been provided in the original data).


```
{
  "ID": "2278_pregnant",
  "NAME": "pregnant",
  "Polysemy": 1,
  "Derivation": 0,
  "PolysemyByFamily": 1,
  "DerivationByFamily": 0,
  "Polysemy_Lexemes": ["11312"],
  "Derivation_Lexemes": [],
  "Polysemy_Shifts": ["shift1635"],
  "Derivation_Shifts": [],
  "Polysemy_Families": ["Uralic"],
  "Derivation_Families": []
}
```


This example shows that at the time of synchronizing the CLDF data with the current state of the database, we found example for a shift between BELLY and PREGNANT in which the direction was not indicated in the database. The shift itself, which can be compared with the actual data online (https://datsemshift.ru/shift1635), belongs to the Uralic family and has the internal Lexeme Identifier “11312”. This identifier itself can be found in the Form Table (file cldf/forms.csv) of the CLDF dataset in the column IDS_IN_SOURCE. When searching for this identifier in the Form Table, we find that it belongs to the Language with the ID 244, which — as we can see from the Language Table (file cldf/languages.csv) refers to the variety Komi with Glottocode komi1268. The form itself is listed as *рушку* and also occurs in the shift shift5895 in the original database (https://datsemshift.ru/shift5895).

We can see that — even if it requires a certain amount of attention to be paid to the internal relations among the data in the CLDF dataset — the individual pieces of information displayed in the original database on the website are likewise represented in CLDF. Additionally, our approach provides a unified view on the lexemes referenced in the database. In this regard, the CLDF version of the Database of Semantic Shifts adds information to the original web interface.

3.5 Representation of Word Families in Individual Languages

We use the Form Table (CLDF FormTable) that is traditionally used to provide data on lexemes in wordlists in Cross-Linguistic Data Formats to represent individual word forms which are additionally linked to their respective source forms. The annotation here follows an earlier proposal (Schweikhard and List forthcoming, see also Pulini and List forthcoming), that aims at displaying derivational relations between individual word forms in one and the same wordlist by adding a column that indicates the source of a given form. This means, with a word like German *Mitte* “middle” from which another German word *Mittelfinger* “middle finger” is derived, we place both words into the same table, with one row being reserved for each word, and we add a column Source_Lexeme which provides the ID of the word that serves as the source. This simple format, illustrated in Table 1, was used to extend model the source-target relations underlying the word forms that serve as realization examples for the shifts in the Database of Semantic Shifts in our CLDF representation of the data. In order to keep this version simple, we decided to only include those lexemes that exhibit directed relations and restrict the shift types to cases of polysemy and derivation.

ID	Language	Concept	Form	Source_Lexeme
1	German	MIDDLE	Mitte	
2	German	MIDDLE FINGER	Mittelfinger	1

Table 1: Representing cases of word derivation in tabular data.

3.6 Data Conversion with CLDFBench

To convert the data from its raw form (in which it was downloaded and prepared in February 5, 2024) to CLDF with the help of CLDFBench, recent versions of Glottolog and Concepticon are required and must be downloaded beforehand. This can again be easily done with GIT. To avoid large downloads, we only download the latest versions of Glottolog and Concepticon in the following example.

```
$ git clone https://github.com/glottolog/glottolog -b v5.0 --depth 1 --single-branch glottolog
```

```
$ git clone https://github.com/concepticon/concepticon-data -b v3.2.0 --depth 1 --single-branch concepticon
```

The conversion to CLDF can then be done with the following command.

```
$ cldfbench lexibank.makecldf --glottolog=glottolog --concepticon=concepticon lexibank_datsemshift.py
```

4 Examples

With the data available in the form of a CLDF dataset, we can easily query the data automatically in different ways. As a first example, one can easily query the concept list with the help of the pycldf Python package (Forkel et al. 2024, Version 1.37.1, https://pypi.org/project/pycldf/).

```
from pycldf import Dataset
ds = Dataset.from_metadata("cldf/cldf-metadata.json")
concepts = ds.objects("ParameterTable")
for concept in concepts:
    if concept.data["Concepticon_ID"]:
        print(concept.data["Concepticon_Gloss"])
        for target in concept.data["Target_Concepts"]:
            target = concepts[target["ID"]]
            if target.data["Concepticon_Gloss"]:
                print(" →", target.data["Concepticon_Gloss"])
```

This code will output all directional relations in the data which are also linked to Concepticon in the form shown below for the concept set YOUNGER BROTHER .

```
YOUNGER BROTHER
→ HUSBAND
→ OWNER
→ BEAR
→ LITTLE FINGER
→ PRINCE (RULER)
```

We can also query the word forms and look for word formation processes in individual languages. Here, it is useful to leverage the possibility to convert CLDF datasets to SQLite, since SQLite queries are very fast and can be adjusted easily to account for the question at hand. The conversion of the CLDF dataset to SQLite is straightforward with the following pycldf command.

```
$ cldf createdb cldf/cldf-metadata.json example/dss.sqlite
```

This creates the database dss.sqlite in the folder example/. Querying the DatSemShift data on individual forms requires to carry out a join of the form table with itself in order to compare all forms with each other. When carrying out the join, it is useful to filter the data beforehand, restricting the data to only one language, which one can identify by its Glottocode (we filter on German with the Glottocode stan1295 in the following). In an earlier study, we have illustrated how SQLite queries can be carried out in a shell script in more detail (Shcherbakova and List 2023), so we won't discuss this aspect of the query here, but just briefly explain how we can extract pairs of source and target lexemes from the CLDF data with the help of an SQLite query.

The core idea here is to use a classical JOIN expression in SQL from the table with the lexical forms (FormTable) and to filter the resulting table by searching for all those cases in which the local ID (Local_ID) recurs in the newly introduced column of the lexeme sources (Source_Lexemes). This selection can be achieved by testing four conditions, namely searching for cases where the local identifier is preceded by a space and occurs at the end of the string representing the source lexemes, where it is followed by a space occurring at the beginning, where it is identical with the whole entry for the source lexemes, or where it occurs inside the string for the source lexemes with a space to both sides. This restriction must be added to the ON part of the SQL statement, as illustrated in the following lines of code.

```
-- previous parts of the query
ON
  (
    table_b.Sources like '% ' || table_a.Lexeme_ID || ' %'
    OR table_b.Sources like table_a.Lexeme_ID || ' %'
    OR table_b.Sources like table_a.Lexeme_ID
    OR table_b.Sources like '% ' || table_a.Lexeme_ID
  )
-- final parts of the query
```

Having stored the entire query in a normal shell script, we can easily call it in order to retrieve all German words in the database along with their source forms for those cases in which both the source and the target word are linked to Concepticon (the script is shared in the folder example/ of the CLDF dataset). This query results in 225 individual shifts, as we can easily check when using the word count functionality in the Shell.

```
$ sh query.sh | wc -l
225
```

The resulting CSV file created by the query can be easily imported into the network visualization tool Cytoscape (Smoot et al. 2011, https://cytoscape.org) and inspected in detail from there. While the concepts form smaller groups (since we are dealing with an individual language here), it is very helpful to inspect smaller patterns that emerge from the individual word relations for an individual language, as can be seen from the small example in Figure 1, showing a part of the German word family network where German *Glas* "glass" extends its meaning to "cup (made from glass)", while the plural form is also used to denote "glasses".

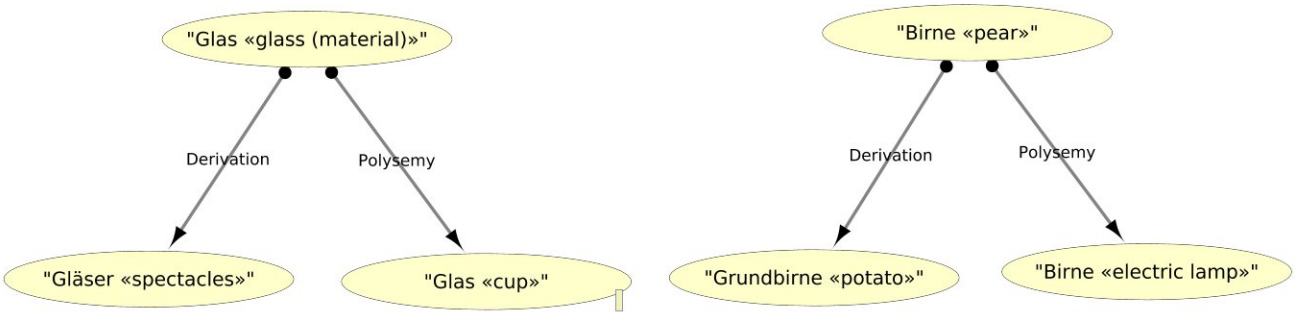

Figure 1: Words derived from the word for "glass" and "pear" in German, as represented in Cytoscape, using the network created from the CLDF dataset.

While these are small examples that one might consider as minor use cases of a database of the size of the Database of Semantic Shifts, it is important to emphasize that the inspection of word family relations in individual networks would not be possible from the database in its current form alone. Only the extended CLDF representation that models not only conceptual relations but also form relations makes it possible to inspect the annotations of the Database of Semantic Shifts in this detail.

5 Conclusion

In this little study we have illustrated how the Database of Semantic Shifts — one of the largest collections of cross-linguistic semantic motivation patterns that are derived from the linguistic literature — can be represented in Cross-Linguistic Data Formats. Apart from allowing direct computational access to the data underlying the original database project, we have also tried to show how the integrated representation in CLDF allows to query the underlying data in ways that go beyond the presentation of the data in the original database. As a result, the CLDF conversion opens new perspectives on this huge resource on semantic change.

Supplementary Material

The database is curated on GitHub (https://github.com/lexibank/datsemshift, Version 1.0) and archived with Zenodo (https://doi.org/10.5281/zenodo.11004150). All examples discussed in this study can be found in the folder example/ of the repository.

References

Blank, A. (1997). Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen [Principles of lexical semantic change, taking the Romance languages as an example]. Tubungen: Niemeyer.

Bocklage, K., A. Di Natale, A. Tjuka, and J.-M. List (2024): Directional Tendencies in Semantic Change. Humanities Commons 2024.2. 1-28. [Preprint, under review, not peer-reviewed] https://doi.org/10.17613/0y0r-f341

Buck, C. D. (1949). A Dictionary of Selected Synonyms in the Principle Indo-European Languages. A Contribution to the History of Ideas. Chicago: University of Chicago Press.

Forkel, R., Rzymski, C., and Bank, S. 2024. PyCLDF [Software Library, Version 1.37.1]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://pypi.org/project/pycldf

Forkel, R. and List, J.-M. (2020): CLDFBench. Give your Cross-Linguistic data a lift. In: Proceedings of the Twelfth International Conference on Language Resources and Evaluation. 6997-7004. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.864.pdf

Forkel, R., J.-M. List, S. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarstrom, M. Haspelmath, G. Kaiping, and R. Gray (2018): Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific Data 5.180205. 1-10. https://doi.org/10.1038/sdata.2018.205

Hammarstrom, H., Haspelmath, M., Forkel, R., and S. Bank (2024): Glottolog [Dataset, Version 5.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://glottolog.org

List, J.-M. (2024): PySem [Software Library, Version 0.8]. With contributions by J. English. Passau: MCL Chair at the University of Passau. https://pypi.org/project/pysem

List, J.-M., Tjuka, A., van Zantwijk, M., Blum, F., Barrientos Ugarte, C., Rzymski, C., Greenhill, S. J., & Forkel, R. (2024). CLLD Concepticon [Dataset, Version 3.2.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org

List, J.-M., R. Forkel, S. Greenhill, C. Rzymski, J. Englisch, and R. Gray (2022): Lexibank, A public repository of standardized wordlists with computed phonological and lexical features. Scientific Data 9.316. 1-31. https://doi.org/10.1038/s41597-022-01432-0

List, J-M. (2021). Converting the Vietic Dataset by Sidwell and Alwes from 2021 to CLDF. Computer-Assisted Language Comparison in Practice. 4.2. https://doi.org/10.58079/m6la

Pulini, M. and J.-M. List (forthcoming): Finding language-internal cognates in Old Chinese. Bulletin of Chinese Linguistics. https://doi.org/10.17613/ftm2-3b58

Schropfer, J. (1956). Wozu ein vergleichendes Worterbuch des Sinnwandels? (Ein Worterbuch semantischer Parallelen) [Why do we need a dictionary of semantic change? (A dictionary of semantic parallels)]. In F. Norman (ed.), Proceedings of the 7th International Congress of Linguists. 366–371. London: Clarendon Press.

Schweikhard, N. and J.-M. List (forthcoming): Modeling word trees in historical linguistics. Preliminary ideas for the reconciliation of word trees and language trees. In: Sprach(en)forschung: Disziplinen und Interdisziplinaritat. Akten der 27. Fachtagung der Gesellschaft fur Sprache und Sprachen. https://doi.org/10.17613/t4wq-h604

Shcherbakova, O. and List, J.-M. (2023). Retrieving and Analyzing Taste Colexifications from Lexibank. Computer-Assisted Language Comparison in Practice, 6(2), 73–86. https://doi.org/10.15475/calcip.2023.2.4

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011): Cytoscape 2.8: New Features for Data Integration and Network Visualization. Bioinformatics 27. 3: 431–32. https://doi.org/10.1093/bioinformatics/btq675.

Sweetser, E. (1990). From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure. Cambridge: Cambridge University Press.

Traugott, E. C., & Dasher, R. B. (2002). Regularity in semantic change. Cambridge: Cambridge University Press.

Trubachev, O. N. (1964[2004]). 'Molchat" i 'tayat". O neobkhodimosti semasiologicheskogo slovarya novogo tipa. ['To be silent' and 'to thaw'. About the Necessity of a Semasiological Dictionary of a New Type]. In Trubachev, O. N. Trudy po etimologii. Slovo, istorija, kultura [Collected works on etymology. Word, etymology, culture]. Volume 1. Moscow: Jazyki Slavjanskoj kultury. 311-318.

Zalizniak, A. A. (2018). The Catalogue of Semantic Shifts: 20 Years Later. Russian Journal of Linguistics 22. 4. 770-787. https://doi.org/10.22363/2312-9182-2018-22-4-770-787.

Zalizniak, A. A., Bulakh, M., Ganenkov, D., Gruntov, I., Maisak, T., & Russo, M. (2012). The catalogue of semantic shifts as a database for lexical semantic typology. Linguistics, 50. 3: 633-669. https://doi.org/10.1515/ling-2012-0020.

Zalizniak, A. A., Smirnitskaya, A., Rousseau, M., Gruntov, I., Maisak, T., Ganenkov, D., Bulakh, M., Orlova, M., Bobrik-Fremke, M., Dereza, O., Mikhailova, T., Bibaeva, M., & Voronov, M. (2024). Database of semantic shifts [Dataset, Version 3.0]. Moscow: Institute of Linguistics at the Russian Academy of Sciences. https://datsemshift.ru/ [accessed on 05/02/2024].

Zalizniak, A. A., Smirnitskaya, A., Russo, M., Mikhailova, T., Bobrik, M., Gruntov, I., Orlova, M., & Voronov, M. (2020). Database of Semantic Shifts [Dataset, Version 2.0]. Moscow: Institute of Linguistics at the Russian Academy of Sciences. http://datsemshift.ru/ [accessed on 07/10/2020].

Supplementary Material

Data and code are curated on GitHub (https://github.com/lexibank/datsemshift, Version 1.0) and archived with Zenodo (https://doi.org/10.5281/zenodo.11004150). All examples discussed in this study can be found in the folder *examples* of the repository.

Acknowledgements

We thank the original authors of the Database of Semantic Shift for their excellent resource that has inspired our own work in many regards.

Funding Information

This project has received funding from the Max Planck Society as part of the CALC^3 Research Project (https://calc.digling.org) and from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101044282). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.