Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3331–3336
Marseille, 20-25 June 2022

Multilingual Pragmaticon: Database of Discourse Formulae

Anton Buzanov, Polina Bychkova, Arina Molchanova, Anna Postnikova, Daria Ryzhova
HSE University, Moscow, Russia
aamolchanova_3@edu.hse.ru
{abuzanov, pbychkova, apostnikova, dryzhova}@hse.ru

Abstract

The paper presents a multilingual database aimed to be used as a tool for typological analysis of response constructions called discourse formulae (DF), cf. English *No way!* or French *Ça va!* (≈ 'all right'). The two primary qualities that make DF of theoretical interest for linguists are their idiomaticity and the special nature of their meanings (cf. consent, refusal, negation), determined by their dialogical function. The formal and semantic structures of these items are language-specific. Compiling a database with DF from various languages would help estimate the diversity of DF in both of these aspects, and, at the same time, establish some frequently occurring patterns.

The DF in the database are accompanied with glosses and assigned with multiple tags, such as pragmatic function, additional semantics, the illocutionary type of the context, etc. As a starting point, Russian, Serbian and Slovene DF are included into the database. This data already shows substantial grammatical and lexical variability.

Keywords: linguistic database, pragmaticalization, discourse formulae, pragmatic typology, Construction Grammar

1. Introduction

In the last three decades, there has been a surge in developement of various standardized multilingual ontologies and databases, such as BabelNet (Navigli and Ponzetto, 2012), WordNet (Miller, 1995), FrameNet (Baker et al., 2003), MetaNet (Dodge et al., 2015), DatSemShift (Zalizniak et al., 2012), and many others. These resources are mostly lexically oriented. Besides traditional definitions of the words, they contain data about semantic relations between meanings, syntactic rules of their use, examples, and, often, translational equivalents. Most importantly, every resource follows its own fixed scheme of data description and storage, hence the data coming from different languages and contributors remain fully comparable. Resources of this kind are treated as a new generation of dictionaries and are widely used in both practical application (see, for example, (Vial et al., 2019; Chakravarthi et al., 2019; Marzinotto et al., 2019)) and theoretical research ((Boas, 2001; Kocoń and Maziarz, 2021; Zalizniak, 2021), and many others).

However, theoretical findings, especially in the framework of Construction Grammar (Fillmore, 1988; Goldberg, 1995; Hoffmann and Trousdale, 2013), show that there are many linguistic units other than words, with specific contextual distribution and non-compositional meanings. They equally require representation in the form of semantically annotated databases. The so-called Constructicons that appeared recently for different languages (Lyngfelt et al., 2018) illustrate the resources of this kind. They catalog constructions, providing information on their semantics and the restrictions on their use in a sentence.

In this paper, we present the database of discourse formulae (DF). Discourse formulae are a special class of constructions which serve as idiomatic reactions to other utterances in dialog. They express the speaker's attitude to the speech act of the interlocutor. Depending on whether it is a question, a statement, or an offer, the reactions can vary from negation or refusal to confirmation or consent. For instance, in the dialog (1), the English DF *I'm good* would express refusal, and *Don't mind if I do!* would express consent.

(1) — I'm making waffles. Want some? — I'm good. / Don't mind if I do!

These DF are non-compositional: the literal meaning of the phrases *I'm good* and *Don't mind if I do* does not directly indicate either refusal or consent. Moreover, some other languages might lack this kind of strategies to express the same reactions (for instance, there is no DF that would literally translate as 'I am good/okay/fine' in Russian). Yet, the pragmatic function of the DF seems to be indirectly motivated by their source. A multilingual database could be a powerful tool for investigating correlations between the source meaning and the target pragmatic functions of DF.

2. Data and Annotation

2.1. Data Sources

Just like Wordnet or Framenet databases started with English and then expanded onto other languages, Multilingual Pragmaticon also started with one language, in this case, Russian. Before building a typological resource, our team developed the general principles for representation of DF based on Russian data and designed a monolingual resource for language learners — Russian Pragmaticon (Yaskevich et al., 2021). The list of Russian DF was compiled semi-automatically, based on manual annotation of dramatic texts (the process is described in (Gerasimenko et al., 2019)). It was subsequently used as a starting point for collecting DF in

other languages with the help of parallel corpora, and questionnaires. As a first step of constructing Multilingual Pragmaticon, we included two other Slavic languages apart from Russian, in order to get an insight into how much variation can be found in closely related languages. So far, 773 Russian, 162 Serbian, and 229 Slovene DF have been entered into the database. The language sample will be further increased, as representatives of other language groups will be added.

2.2. Annotation

Every formula in the database has a default form and up to 14 realizations, due to variation of emphatic particles, word order, etc. The annotation of DF includes the following parameters:

- language,
- DF inner structure,
- glosses,
- lemmas,
- pragmatic function,
- additional semantics,
- contextual speech acts,
- dialog structure,
- intonation,
- syntax,
- source construction,
- SC syntax, and
- SC intonation

We do not manually establish direct intra- or cross-linguistic connections between similar DF because the database interface can provide clusters of formulae based on different parameters (not restricted to synonyms or translation equivalents). We will discuss these parameters in further detail in the rest of this section, illustrated with the Russian DF *ne možet byt'*, ≈ 'no way' (see 2 for other realizations of the formula).

(2) a. *ne* *može-t* *by-t'*
NEG.PTCL can-PRS.3SG be-INF

b. *eto-go* *by-t'* *ne* *može-t*
DEM.PROX-GEN.SG be-INF NEG.PTCL can-PRS.3SG

c. *da* *by-t'* *ne* *može-t*
JUXT.PTCL be-INF NEG.PTCL can-PRS.3SG

The **inner structure** tag provides a loose description of the literal meaning of a formula. This parameter is two-leveled: the main field corresponds to a more general classification, while the second field highlights additional distinctions. For instance, for *ne možet byt'*, the inner structure type is EPISTEMIC MODALITY since it contains a modal verb *možet*. The inner structure subtype is IMPOSSIBILITY since there is a negation particle *ne*. The inner structure tag allows to group the DF in the database based on their form, and explore the pragmatic functions that correspond to specific forms across languages. At least two types of tasks can be accomplished using this tag. On the one hand, it simplifies establishing translational equivalents for DF. For instance, there are several DF in Serbian with the same inner structure as *ne možet byt'*: *ne može biti*, *nema šanse*, *nećete valjda*, *nije mogućno*, *nema veze*, and others. They are the most accurate analogues of the Russian DF *ne možet byt'* in Serbian. On the other hand, it helps to compare DF of similar usage cross-linguistically. We can see that the variability of DF with tags EPISTEMIC MODALITY / IMPOSSIBILITY in Serbian is much higher than in Russian. It makes possible to find both common paths and sources of pragmaticalization and constructions which are unique to a particular language.

Another formal parameter is glossing and lemmatization. The interlinear **glosses** (e.g. *ne može-t by-t'* 'NEG.PTCL can-PRS.3SG be-INF') enable search by a particular word or grammatical category (for instance, one can find all DF with the verbs of speech, or all DF that contain imperative forms. **Lemmas** are language-specific and allow to search for a particular lexeme. The glossing was done manually, the process facilitated with the use of FieldWorks.[1]

The next set of features is dedicated to semantic and pragmatic properties of discourse formulae. **Pragmatic function**, or **primary semantics**, reflects the main discourse function. The set of pragmatic functions for DF is quite compact and includes negation, prohibition, refusal, surprise, agreement, assessment, confirmation and indifference. Some DF may have multiple functions: e.g. *ne možet byt'* can be used either as NEGATION or as REFUSAL.

The field **additional semantics** is reserved for more nuanced semantic characteristics, such as negative or positive assessment, disbelief, doubt or confidence, etc. Unlike pragmatic function tags, the tags of additional semantics can be used in combination (cf. doubt + negative assessment). Since these tags refer to more subtle semantic distinctions, they can differ in several realizations of the same DF. For instance, *ne možet byt'* in its default form can express genuine surprise of "credence", however, the rest of its realizations can only mean disbelief.

The field **speech act** specifies the type of the speech act directly preceding the DF, triggering its use and thus defining its meaning. *Ne možet byt'* can react to two types of speech acts: HYPOTHESIS, and NEWS. After the former, it functions as NEGATION, and after the latter, as an expression of SURPRISE.

The field **dialog structure** reflects the number of speech acts that are relevant for the use of the DF in a dialog. The structure can be either BIPARTITE or TRIPARTITE (i.e. one or two utterances before the DF itself) depending on whether the speech act before a trigger is necessary.

Since the DF usually function as full utterances, their **intonation** is described in terms of general intonation patterns in the given language. The field has four

[1]The home page of the project is https://software.sil.org/fieldworks/, and the source code is available at https://github.com/sillsdev/FieldWorks

realization	inner structure	glosses	lang	pragmatics	speech act
ešče čego	irony	yet WHAT.GEN	ru	refusal	suggestion | offer | advice | request
ešče čego	irony	yet WHAT.GEN	ru	negation	hypothesis | polar question
glupost-i	speech act devaluation	nonsense-ACC.PL	ru	negation	hypothesis | opinion
kak by ne tak	volitive modality	how SUBJ.PTCL NEG.PTCL so	ru	refusal	demand
kak by ne tak	volitive modality	how SUBJ.PTCL NEG.PTCL so	ru	negation	hypothesis
ko to zna	epistemic modality	who.NOM it.ACC.SG know.PRS.3SG	sr	refusal	

Table 1: Example of annotation

possible values: STATEMENT, EXCLAMATION, POLAR QUESTION or WH-QUESTION. The **syntax** will be annotated with PoS-tagging and chunking (i.e. marking the borders of syntactic phrases).
The **source construction** is the non-compositional structure, formally similar to the DF, that is used outside of the dialog and is likely to be the source from which the DF originated. Information for the source construction is acquired in the same way as for the main formula.
From a technical point, it was necessary to provide annotators with a user-friendly interface which would be easily converted into the database. Editing instances must also be available for annotators even after loading the data in the database. We created a table on Google Sheets and this allowed us to comfortably annotate the DF collaboratively and, what is more important, to edit the annotation and load it into the database in batches.
A fragment of the annotation table is given in Table 1. There, the formulae are grouped by pragmatic function, and different speech acts are divided with vertical bars, which are removed during preprocessing.

3. Database

3.1. General properties

The relational PostgreSQL database is structured to cover pragmatic, semantic, morphological, and discourse features of the data.
The database is maximally decomposed compared to the markup table where all features were annotated within the same list. The central table is dedicated to particular realizations. The values for every property in the list below are stored as separate tables.

- lemmas,
- languages,
- formulae,
- intonations,
- source constructions,
- glosses,
- inner structures types,
- inner structure subtypes,
- primary semantics (aka pragmatics),
- additional semantics, and
- speech acts

Additionally, there are some cross-reference tables to implement many-to-many relations. These are

- realization2lemma,
- realization2gloss,
- realization2inner_structure,[2]
- realization2speech_acts, and
- semantics

A many-to-many relationship is a relationship between two entities (columns), where values from both of them can correspond to more than one value within another entity (column). The perfect example of such a relationship is that of between realizations and glosses: a realization may contain more than one morpheme, and a certain morpheme can occur across different formulae.
Other relations are one-to-many, thus the central table is directly tied up with auxiliary tables (structures, formulae, source construction, etc; the full database structure is presented in Figure 1). A one-to-many relationship is a relationship between two entities (columns), where values from the first entity can correspond to many values from the second entity and not vice versa. A formula can have many realizations, but a realization can belong to only one formula.

3.2. Preprocessing

Comparison between Table 1 and Figure 1 shows that the data need to be preprocessed before going to the database. If two formally identical realizations have different pragmatic meanings, we consider them to be distinct DF. This is the only criterion to separate formally identical expressions. If a realization can be used

[2]Technically, it is not a many-to-many relation, however due to manual defects in the markup it is easier to create such an auxiliary table.

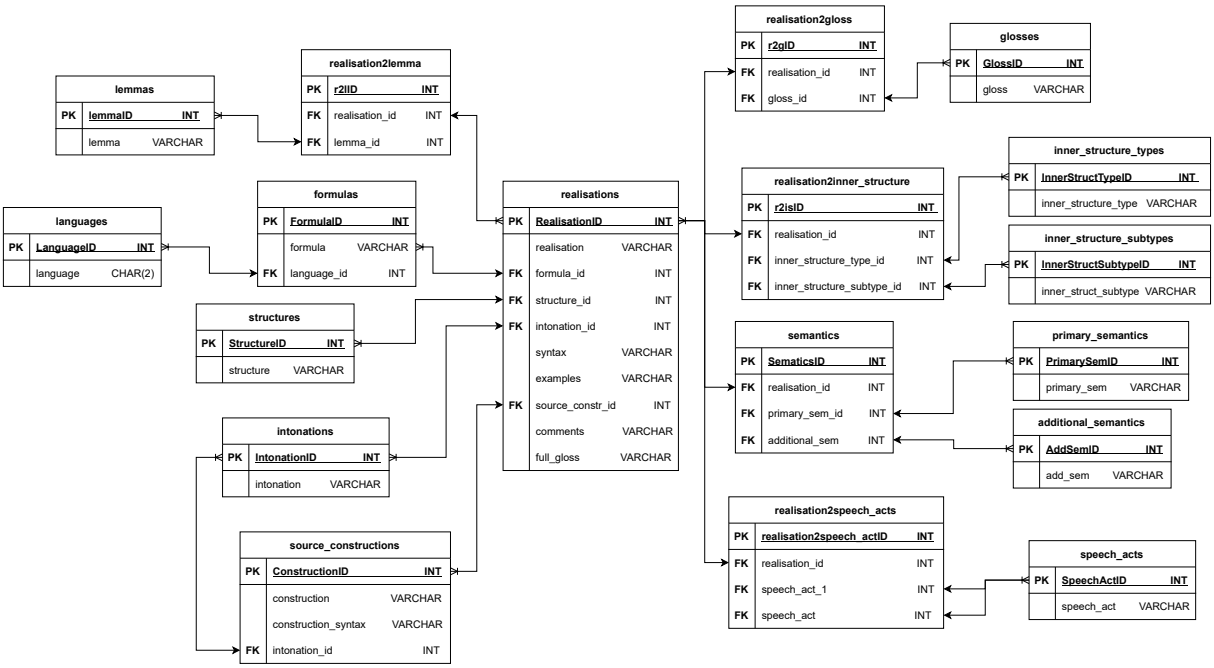

Figure 1: Database structure

as an answer to several different speech acts, we consider it to be a single formula and do not create separate lines for it.

Firstly, we eliminate the DF that are not yet ready to be added to the database, or have been already uploaded. The next step is to form a list of glosses and a list of lemmas. Lemmas and glosses are converted to lowercase and stripped in case of unexpected space characters.

Several speech acts or additional semantics can be assigned to one realization, but they must be separated into multiple lines to be stored in the database. Before adding values to the database, they must be indexed since this is the most convenient way to catalog data.

When all lists from the table are gathered and indices are ascribed to every value, all the data come to the database and spread over the tables forming a network with foreign-key relationships inside.

3.3. Web-interface

To expand the potential audience, we designed the user interface as a web application written in Python on the basis of the Flask framework.[3]

The interface makes two major search modes available. The first mode allows searching for a particular formula and getting all information concerning it. In that mode, other search fields are blocked. The second mode allows for choosing different values from different fields. For most of them, we provide hints with possible values (i.e. multiselect is implemented). Syntactic structure is the only field that requires typing and not choosing from the list at present.

[3]Documentation is available on https://flask.palletsprojects.com/en/2.0.x/

The database allows for various queries. The interlinear glosses are provided for every formula, enabling search by a particular word or grammatical category.

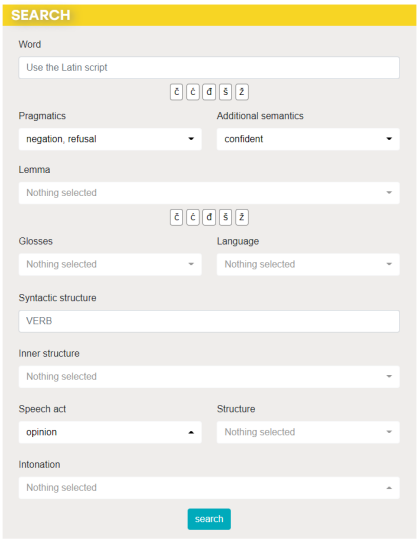

Figure 2: Search form

The main feature of the interface is that it can combine different values within one formula. Selecting multiple values (see, for instance, pragmatics on Figure 2) implies that **all** values must be presented within a formula. The query combining NEGATION and REFUSAL will end up with formulae which can express both meanings (such as *kak by ne tak* 'I don't think so', *pobojsja boga* 'for God's sake').

Although the entries with different pragmatic function are treated as distinct DF, it is possible to perform a search with conjunction of pragmatic tags. In that case,

the DF are grouped by their form, returning a list of polysemous DF. Other search fields enable grouping by the DF itself. Thus, the query in Figure 2 searches for all formulae that can express both NEGATION and REFUSAL, and then preserves only those that are uttered after OPINION and can express CONFIDENT.

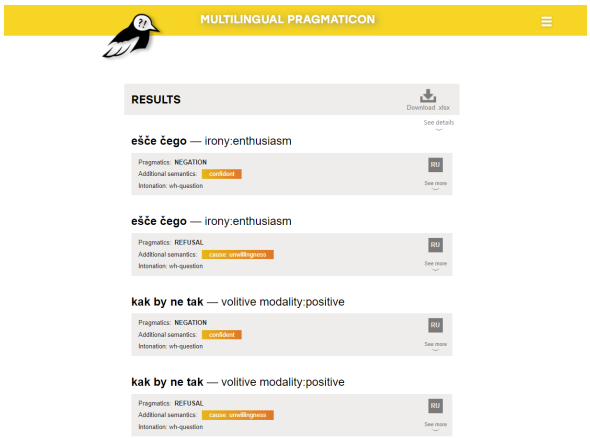

Figure 3: Search results

Search by other fields combines features within one formula (again, one formula corresponds to only one pragmatic meaning, i.e. two *ešče čego* in Figure 3 are **distinct** formulae). Thus, lemmas, glosses, speech acts, and other features must be all present within any of pragmatics (not in all of them).

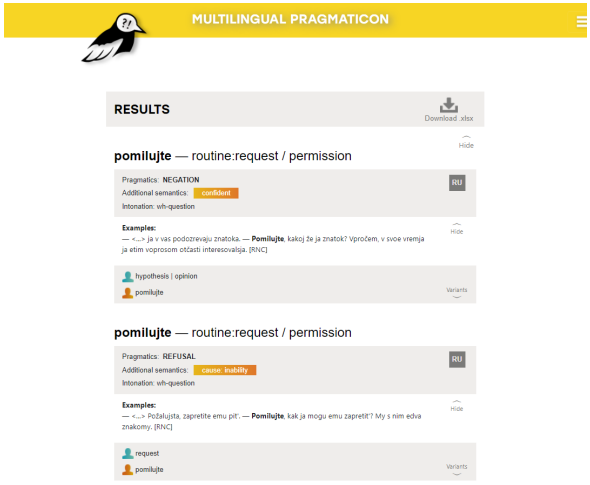

Figure 4: Detailed search results

Search results are both provided within the interface (as is shown in Figures 3 and 4) and can be downloaded as an `xlsx` table. Within the interface, results are shown in two ways, compact and detailed. Compact results as in Figure 3 show only the main information about formulae, while detailed results as in Figure 4 show all the information available, including realizations, glosses, examples, speech acts, etc.

The task of searching all elements from the list, when a many-to-many relationship is established, is not so easy. To be able to search for realizations containing all glosses (lemmas, inner structures, etc.) from a certain list we had to implement the algorithm of Relational Division with Remainder. This algorithm allows searching for many values when they are conjugated. Our solution was inspired by that in (Celko, 2009).

Thus, we provide a web interface that covers many possible queries in which potential user can be interested. However, it is often not possible to obtain results using disjunction of values within a single search query. In that case, multiple queries must be used.

The interface is going to be published on https://linghub.ru/ in May, 2022. Both the source code and the lists of the parameter values (additional semantics, inner structures and speech acts) are available at https://github.com/vantral/Multilingual-Pragmaticon.

4. Conclusion

We introduced a multilingual database of discourse formulae. DF are specific pragmatic items, which are often idiomatic and express pragmatic meanings from a closed set of speaker's attitudes towards the content of the interlocutor's utterance.

From the theoretical point of view, the resource can serve for the analysis of pragmaticalization – the process during which lexical units become pragmatic items (Diewald, 2011). The database contains the models of the source meanings of the DF (see inner structure and glosses), as well as the classification of the resulting pragmatic units (see primary and secondary semantics and speech acts which serve as triggers).

The classification of pragmatic meanings built on the basis of typological data is valuable itself; to the best of our knowledge, nothing similar has been suggested yet. Additionally, the analysis of polysemous DF from many languages will reveal the sets of meanings which often cluster together, and thus it will be possible to establish cognitive distances between the points of this pragmatic meaning space (cf. $CLICS^3$ (Rzymski et al., 2020) for lexical meanings).

As for practical applications, the database can be used to establish translational equivalents for DF in different contexts. This task is far from trivial, and often causes problems to interpreters and language learners.

Besides their high frequency in everyday colloquial speech, DF are not fully implemented in dialog systems, because of the insufficient theoretical knowledge on their nature and their crucial role in the dialog, and hence the lack of training data. We believe that our resource would help to overcome this problem as well.

Finally, other classes of linguistic units, such as routines (*Hello! God bless you!*) and interjections (*Oh my God!*), can be studied and represented following the same principles.

5. Acknowledgements

This work was supported by Russian Foundation for Basic Research, research project no. 20-012-00240. We are also grateful to Ekaterina Taktasheva and Ekaterina Voloshina for technical support.

6. Bibliographical References

Boas, H. C. (2001). Frame semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography. In Paul Rayson, et al., editors, *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster. University Centre for computer corpus research on language, University Centre for computer corpus research on language.

Celko, J. (2009). Divided we stand: The sql of relational division. *Simple Talk*.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019). Wordnet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7.

Diewald, G. (2011). Pragmaticalization (defined) as grammaticalization of discourse functions. *Walter de Gruyter GmbH & Co. KG*.

Fillmore, C. J. (1988). The mechanisms of" construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pages 35–55.

Gerasimenko, E., Puzhaeva, S., Zakharova, E., and Rakhilina, E. (2019). Defining discourse formulae: computational approach. In *Proceedings of Third Workshop" Compu*, volume 4, pages 61–69.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Hoffmann, T. and Trousdale, G. (2013). *The Oxford handbook of construction grammar*. Oxford University Press.

Kocoń, J. and Maziarz, M. (2021). Mapping wordnet onto human brain connectome in emotion processing and semantic similarity recognition. *Information Processing & Management*, 58(3):102530.

Marzinotto, G., Damnati, G., and Béchet, F. (2019). Adapting a framenet semantic parser for spoken language understanding using adversarial learning. *arXiv preprint arXiv:1910.02734*.

Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. *arXiv preprint arXiv:1905.05677*.

Zalizniak, A. A. (2021). Cognitive mechanisms of semantic derivation in the domain of visual perception. In *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics: Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, October 10-16, 2020, Moscow, Russia*, volume 1358, page 267. Springer Nature.

7. Language Resource References

Baker, C. F., Fillmore, C. J., and Cronin, B. (2003). *The structure of the FrameNet database*. Oxford University Press. Available at https://framenet.icsi.berkeley.edu/fndrupal/.

Dodge, E. K., Hong, J., and Stickles, E. (2015). *MetaNet: Deep semantic automatic metaphor analysis*. Available at https://metanet.icsi.berkeley.edu/metanet/.

Lyngfelt, B., Borin, L., Ohara, K., and Torrent, T. T. (2018). *Constructicography: Constructicon development across languages*. John Benjamins Publishing Company.

Miller, G. A. (1995). *WordNet: a lexical database for English*. ACM New York, NY, USA. Available at http://wordnetweb.princeton.edu/perl/webwn.

Navigli, R. and Ponzetto, S. P. (2012). *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*. Elsevier. Available at https://babelnet.org/.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., Chang, S., Lai, Y., Morozova, N., Arjava, H., Hübler, N., Koile, E., Pepper, S., Proos, M., Van Epps, B., Blanco, I., Hundt, C., Monakhov, S., Pianykh, K., Ramesh, S., Gray, R. D., Forkel, R., and List, J.-M. (2020). *The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies*. Available at https://clics.clld.org/.

Yaskevich, A., Bychkova, P., Koziuk, E., Rakhilina, E., Slepak, E., Utkina, A., Zhukova, S., and Tatiana, Z. (2021). *The Russian Pragmaticon. An electronic database of the Russian pragmatic constructions*. Available at https://pragmaticon.ruscorpora.ru/.

Zalizniak, A. A., Bulakh, M., Ganenkov, D., Gruntov, I., Maisak, T., and Russo, M. (2012). *The catalogue of semantic shifts as a database for lexical semantic typology*. De Gruyter Mouton. Available at https://datsemshift.ru/.

Appendix: Glossary

3 — third person
ACC — accusative
DEM — demonstrative
GEN — genitive
INF — infinitive
JUXT — juxtaposition
NEG — negative
PL — plural
PROX — proximate
PRS — present
PTCL — particle
SG — singular
SUBJ — subjunctive