

Introduction

Linguistic typology is traditionally viewed as grammatical or phonological typology. This is easily accounted for by the fact that phonology and grammar are represented by regular markers, observable and comparable across languages. Cross-linguistic analysis of lexical categorization in conceptual domains is a specific task due to seemingly blurred relations between lexical items as system-forming units and is a relatively new area of typological study. However, several prominent approaches to cross-linguistic analysis of lexical systems have already become known. Those include Natural Semantic Metalanguage or NSM (Anna Wierzbicka and Cliff Goddard) and denotation-based approach (MPI Nijmegen: Asifa Majid, Stephen Levinson, et al.), which use two different methodologies. NSM treats lexical meanings as constructed from semantically primitive units [Semantic and Lexical Universals..., 1994; Words and Meanings..., 2014]², while the MPI Group uses experimental techniques that aim at eliciting lexicalized reactions to visual or sensory stimuli [Cutting and Breaking Events..., 2007; Levinson, 2008; The Senses..., 2011]. On advantages and disadvantages of these two approaches see [Rakhilina, Reznikova, 2013].

The case study to be analyzed in this paper was carried out in the framework of the Moscow Lexical Typology Group (MLexT). MLexT roots its studies in investigating linguistic behavior of lexical items and detecting colexification strategies that various languages follow. The method is largely based on the approaches introduced by the Moscow Semantic School: semantic oppositions within a given semantic domain are obtained through comparisons of close synonyms [Apresjan, 2009] and their contextual behavior. The method is based on collecting and analyzing lexical items belonging to the domain and their distribution in the languages from the research sample (starting from the native language of the researcher and

¹ The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2016.

² Similar approach to lexical semantics is partially taken by the Moscow Semantic School and Meaning-Text model, though these do not operate with primitives.

juxtaposing the results with other languages). Data is mostly elicited with the help of corpora, questionnaires and bilingual dictionaries. Building blocks for meaning representation are *frames* — typical situations characteristic for the domain, which also serve as entries for typological questionnaires. Frames and semantics maps that represent colexification strategies are illustrated in Fig. 1 below.

This methodology has already been implemented in the following projects performed by MLexT: verbs of aquamotion [Glagoly dviženija..., 2007], pain metaphors [Reznikova et al., 2012], sound metaphors [Reznikova et al., 2015], verbs of rotation [Kruglyakova, 2010], verbs of oscillation [Rakhilina, Prokofieva, 2005; Shapiro, 2015], and physical qualities (e.g. “sharp”, “wet”, “soft”, “even/flat”, cf. [Kyuseva, 2012; Kashkin, Vinogradova, 2015] and others).

In this paper we would like to have a closer look at the research of the semantic domain of OLD across languages. For MLexT the typological study of OLD was a novice project — it was largely dictionary based and the sample included more than a hundred languages (much more than the average number of 5–8 languages analyzed in a regular research of this kind). The list of frames elicited within the domain includes (see also [Rakhilina, 1999]):

- 1) objects changing in time (like old people or old trees);
- 2) objects having a fixed life-time (old clothes);
- 3) changeable, renewed objects (old director “previous”);
- 4) objects belonging to previous age (old town, old coins).

The data was ultimately checked with the help of questionnaires and experts. Variation in colexification patterns leads to three major types of systems for OLD found in world languages.

In dominant systems one lexeme covers all the four situations despite competing with quasysynonyms in some of them, in binary systems two lexemes share the domain and distributing systems have three or four lexemes accommodating the full set of frames.

Learner Corpora as a tool for typological studies

As we have seen above, implementing one tool is not enough for conducting a full-fledged typological study of lexis. Thus, within the framework proposed by MLexT, there are several instruments (dictionaries, corpora, questionnaires among them), each performing a separate function that contributes to the ultimate result.

The subject of the present section is to introduce another tool that may prove helpful for lexical typology, even though it is not often associated with this branch of linguistic study. Below we will talk about Learner Corpora.

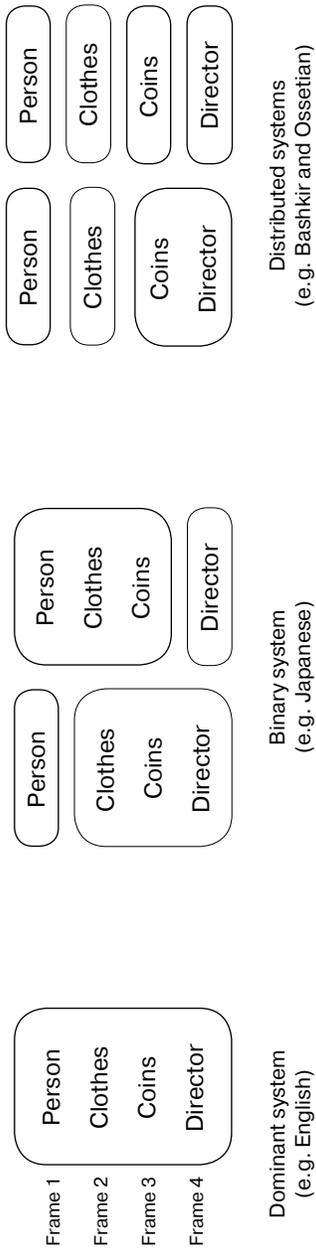


Fig. 1. Semantic maps for the domain of OLD

Corpora have already been mentioned as a means of obtaining lexical data. A regular corpus provides so called “positive” data — phrases grammatically acceptable in a certain language — thus giving basis for generating initial hypotheses and their subsequent verification. Within the framework of MLeXT group, corpora are used to analyze data coming from the native language of a researcher at the start of a study and later for verification of questionnaire results produced by speakers of other languages in the sample.

As collections of oral and written texts produced by people who are in the process of mastering a language, learner corpora are originally designed for research in SLA or in pedagogical purposes and have been by far largely disregarded by typologists. At the same time they contain linguistic data from speakers who have access to two or more language systems — one fully acquired, others as represented in their interlanguage. The latter fact is particularly interesting because errors (or samples of deviational language use) produced by such speakers bring together data from speaker’s interfering languages and can thus serve as a natural instrument for visualization of different patterns of colexification. From this perspective, learner corpora can be seen as a valuable verification tool at the final stage of a typological research. In what follows, we will present an example of how this tool was applied to make the results of OLD research more precise.

As it has been shown in the previous section, lexical systems of OLD in different languages fall into three major groups — dominant (one lexeme applicable to all frames belonging to the domain), binary (all frames distributed between two lexemes) and distributed (three or four lexemes cover the domain). It is naturally expected that a person studying a foreign language will make lexical errors in the domain if their L1 and interfering L2 follow different typological strategies (e.g. dominant and classifying) and, on the contrary, the lexical choice will be accurate if L1 and L2 belong to the same type.

We used the Russian Learner Corpus³ to conduct an experiment and prove the latter fact. We compared corpus data coming from speakers of Russian as a second language and speakers of Heritage Russian whose dominant language is English. English and Russian belong to the same typological type and have a dominant lexeme in the domain of OLD — *old* in English and *staryj* “old” in Russian. Our goal was to find if there were errors in the use of these words.

³ The Russian Learner Corpus comprises texts produced by two categories of non-standard speakers of Russian: L2 learners and HL speakers with various dominant languages. Texts produced by speakers with dominant English were provided by Anna Alsufieva, Evgeny Dengub, Irina Dubinina, and Olessya Kisselev. A preliminary linguistic analysis and tagging were done by the members of the Heritage Russian Research Group (Higher School of Economics), with technical support provided by Elmira Mustakimova, Ekaterina Uetova and Timofey Arkhangelskiy.

The results of the experiment showed that even though the lexical system of OLD is the same in two given languages, errors of different kinds occur in the speech of non-native and heritage speakers. These errors pointed at two kinds of linguistic phenomena: those that initially lay beyond the tasks of the research and those that revealed methodological gaps. We will further take a closer look at both kinds. It should be noted, however, that in the present work our aim is not to give an exhaustive account of how learner corpora can be applied in lexical typology but rather to outline the directions for future work in the field.

Contexts that lie beyond the scope and tasks of MLexT research

The purpose for examining these contexts is twofold: they help checking and formalizing research borders and give direction for further analysis of the domain. There are three types of contexts found in RLC search results:

- contexts with abstract (deverbal) nouns;
- predicative non-stative contexts;
- substantivized use of adjectives.

Deverbal nouns

MLexT research of OLD mainly focused on the names of concrete objects (people, artifacts, plants, etc.), while contexts containing abstract deverbal nouns lied beyond its scope. Search results from RLC, however, show that lexicalization process for patterns like OLD + ABSTRACT CONCEPT (including deverbal nouns) follow particular distributional rules, cf. the following example:

- (1) *mne prosto nadoela moja *byvšaja strižka* L2 Russian
lit. "I was just bored with my former haircut"

In (1) the speaker used a quasisynonymous word *byvšyj* with a deverbal abstract noun *strižka* referring to the third frame for entities that are no longer in use. A more acceptable Russian phrase to express the idea is to use the dominant lexeme:

- (2) *mne prosto nadoela moja staraja strižka* Standard Russian
lit. "I was just bored with my previous haircut"

This deviation from Standard Russian can be explained by interference from English, where the use of quasisynonymous lexeme is preferred to the use of dominant one, cf. *my previous / °old haircut*.

Note, however, that *staryj* is not universally preferred with deverbal nouns, cf. (3) and (4):

- (3) ?? *Mne ne nrvjatsja ego staryje ispolnenija Šopena*
lit. "I don't like his old Chopin performances"

- (4) *Ego *staryj prijezd nadelal mnogo šuma*
lit. "His old visit was scandalous"

These examples show that deverbal noun contexts with adjectives from OLD domain follow specific rules and require additional typological research.

Non-stative contexts

Another phenomenon that was not specifically addressed in the original research on the typology of OLD is the syntactic and semantic opposition between attributive and predicative constructions.

We were mainly dealing with attributive contexts which always display a stative meaning of an adjective, cf. *old tree, old paintings, old regime*. Predicative position, despite having a stative reading, may also trigger an idea of a process or result. Cf. the following sentences in Standard and L2 Russian, where the error in (6) is caused by this reason:

- (5) *Staryje ljudi / oni staryje* [stative meaning] *Standard Russian*
lit. "Old people / they are old"
- (6) *Potom kogda oni uže vmeste I *staryje*
u nix umirajet syn [result: "they became old"] *L2 Russian*
lit. "Then when they are already together and old
their son dies"

(6) contains two clauses with different actionality characteristics. The main clause describes a punctual event ("their son dies"), while the subordinate one is intended to describe a result, which is supported by the use of *uže* "already" ("they are already old"). However, the use of *old* as an adjective blocks this interpretation leaving room only for stative reading. This indicates that two clauses should obey the rule of actionality based concord that prohibits the co-occurrence of states and punctual events in one complex sentence. To express the aspectual relations between the two clauses correctly, the speaker should have chosen the telic verb *sostarit'sja* "to get old", cf.:

- (7) *Kogda ljudi staryje* [state], *oni obyčno ploxo*
vidjat [state] *Standard Russian*
lit. "When people are old [state] they usually see
badly [state]"
- (8) *Kogda oni sostarilisъ* [result],
u nix umiraet syn [punctual event] *Standard Russian*
lit. "When they got old [result],
their son dies [punctual event]"

Nominalization

The research of OLD as a property domain covered mainly adjectives, and the process of nominalization was beyond our interest. Working with corpus data, however, showed that this kind of morphological derivation is worth a detailed typological analysis. The difference in distributional strategies for deadjectivals with this meaning in English and Russian is exemplified by (9):

- (9) *Ogon' gorel vseгда, dnjom i noč'ju,*
*i *staryje i deti spali na peči.* L2 Russian
lit. “The fire was always burning, day and night,
and old and children were sleeping
on the Russian stove”

It is most likely that in producing this sentence the speaker was relying on derivational patterns of their native language, where adjectives undergo nominalization through zero-derivation, cf. (10):

- (10) *It is as if the wealthier old have convinced the young*
to look away from their money...⁴

In Russian *staryi* “old” is used as a noun only idiomatically in the archaic construction *star i mlad* (“old and young”), in other contexts the noun from the adjective *staryj* “old” is derived through suffixation, cf. the noun *starik* “old man” formed with the help of *-ik* suffix:

- (11) *Ogon' gorel vseгда, dnjom i noč'ju,*
i stariki i deti spali na peči. Standard Russian
lit. “The fire was always burning, day and night,
and old men and children were sleeping
on the Russian stove”

Thus, word-formation rules present another subject for further research in the domain of OLD.

In this section we have outlined three areas for further typological research, and we are now proceeding with briefly showing how lexical errors found in the Corpus help uncover minuses in the initial research methodology.

Contexts in the scope of the MLexT research (gaps in the methodology)

This section focuses on errors in contexts prototypical for the four frames elicited in the course of our study of OLD. Based on its results, we have to assume

⁴ The example was taken from COCA (<http://corpus.byu.edu/coca/>).

that these contexts should accommodate the dominant lexeme *staryj* “old”, however, this is not the case, cf. (12):

- (12) *V lesu on vstrelil bednogo *starogo muščinu,*
kotoryj zastrjal v bolote. *L2 Russian*
lit. “In the woods he met a poor old man
who got stuck in the marsh”

People’s age is described by the first frame and it is expressed by the dominant lexeme in Russian (*staryj čelovek* “old man”, *staryje ljudi* “old people”, *staryj professor* “old professor”, *bednaja saraja ženščina* “poor old woman”). However, in combination with the word *muščina* (“man”, gender marked) it is not acceptable and should be substituted by a quasisynonym, cf. (13):

- (13) *V lesu on vstrelil bednogo požilogo muščinu,*
kotoryj zastrjal v bolote. *Standard Russian*
lit. “In the woods he met a poor aged man
who got stuck in the marsh”

Thus, the dominance of *staryj* “old” is not absolute and there are gaps in its lexical distribution within a given frame.

In the next example, the speaker uses the dominant lexeme *staryj* “old” in combination with the noun *poezd* “train”. The concept *staryj poezd* “old train” is described in our system either by the second frame (object having a fixed life-span that has become useless or/and decayed in the process of its continuous use) or the fourth frame if the speaker implies that the train is antique (object belonging to the previous age). In both interpretations the word *staryj* “old” as a dominant adjective should be the right lexical choice. However, what hasn’t been captured by our initial methodology is variations in referential status of the entire phrase, which is the clue to the error in (14).

- (14) *Kočegar periodičeski položit ugol’ v kamin, naprimer,*
*v *starom poezde čtoby ogon’ ne končilsja.* *L2 Russian*
lit. “A stoker would put coal in the fireplace
from time to time for example in an old train
to keep the fire burning”

In (14) the speaker’s intention is obviously to use the noun phrase *starom poezde* “old train [prep. case]” in non-referential interpretation and should be thus substituted either by the plural form or a construction, cf. (15):

- (15) *Kočegar periodičeski položit ugol’ v kamin,*
kak, naprimer, v staryx poezdax [PI] /

kak ran'še delali v poezdax [constr],

čtoby ogon' ne končilsja.

Standard Russian

lit. “A stoker would put coal in the fireplace
from time to time as for example in old trains /
as they used to do in old trains to keep
the fire burning”

The last lexical error to be discussed in this section is displayed in (16).

(16) *A zdes' — korova idjot mimo lodki*

*v *byvšej territorii morja!*

L2 Russian

lit. “And here — a cow is going past the boat
across the former seabed”

This sentence illustrates the third frame in the system of OLD: combined with the word *territorija* “territory” (in this case *territorija morja* “seabed”), the lexeme *byvšij* “former” refers to the situation when a given piece of land that used to be a sea bottom is no longer covered with water and is not classified as a body of water. *Byvšij* “former”, though, isn’t acceptable in this type of context and, moreover, its substitution by the dominant *staryj* “old” doesn’t solve the problem either, cf. (17):

(17) *A zdes' — korova idjot mimo lodki v *staroj territorii*

morja / tam, gde ran'še bylo more [constr]!

lit. “And here — a cow is going past the boat
across the old seabed”

Thus, (17) points to the gap in the lexical system of Russian, where the idea should be expressed with a construction/clause. Note that in English the word *seabed* can be combined with an adjective with OLD semantics, which makes two dominant systems — Russian and English — inequivalent⁵:

(18) *At least soils are relatively uniform:*

*an uplifted former seabed, with sandstone,
silt and fractured quartz.*

A closer look at the differences between dominant systems with regard to the phenomena listed above is one of the further steps of our typological research.

⁵ Note that all three types of bigrams are possible in Russian in other contexts belonging to the same frames, cf.: *vstretil eščjo ne starogo mužčinu...* (*lit.* “met a man that is not old yet”), *Začem molodym devuškam staryje mužčiny* (*lit.* “why do young girls need old men”) [negation, contrast]; *na vystavke byl predstavlen staryj avtobus* (*lit.* “an old bus was displayed in the exhibition”) [referent noun]; *byvšaja territorija SSSR* (*lit.* “former territory of the USSR”), *staraja territorija Kremlja* (*lit.* “former territory of the Kremlin”) [geographical terms].

Conclusions

Tying together two linguistic paradigms — second language acquisition and lexical typology, this paper introduces a new tool for the studies in the latter field. Containing texts produced by speakers who have access to at least two language systems, learner corpora provide new data and pose new problems for lexical typology. To illustrate this, we have addressed a semantic domain of OLD and chosen two languages (English and Russian) where lexical systems of OLD are of the same dominant type. Though the distribution of lexical items belonging to the domain of OLD should have been proven identical, L2 learners of Russian made lexical errors in producing attributive collocations. In the majority of erroneous data, attributive bigram combinations were potentially possible, all restrictions determined by broader context. It has also been found that according to corpus data, lexical items in the domain compete not only with other lexical items, but also with multiple-word constructions. These assumptions may prove valuable for further research within the OLD domain, as well as in other domains within the scope of lexical typology.

References

- Apresjan Ju.D.* Issledovanija po semantike i leksikografii. T. I: Paradigmatika [Research in Semantics and Lexicography. Vol. 1: Paradigmatics]. Moscow: Jazyki Slavjanskoj Kul'tury, 2009. (Studia philologica).
- Cutting and Breaking Events: A Crosslinguistic Perspective / A. Majid, M. Bowerman (eds). Cognitive Linguistics [Spec. Iss.]. 2007. Vol. 18. No. 2. P. 133–152.
- Glagoly dviženija v vode: leksičeskaja tipologija [Verbs of AQUA-Motion: A Lexical Typology] / T.A. Majsak, E.V. Rakhilina (eds). Moscow: Indrik, 2007.
- Glagoly zvukov životnyx: Tipologija metaphor [Verbs of Animal Sounds: Typology of Metaphors] / T.I. Reznikova, E.V. Rakhilina, D.A. Ryzhova, A.S. Vyrenkova, B.V. Orekhov (eds). Moscow: Jazyki Slavjanskoj kul'tury, 2015.
- Kashkin E., Vinogradova O.I.* The Domain of Surface Texture: Working Papers by NRU HSE. 2015. Ser. WP BRP “Linguistics”. No. 30.
- Kruglyakova V.A.* Semantika glagolov vraščeniya v tipologičeskoj perspektive [Semantics of Rotation Verbs in a Typological Perspective]: PhD thes. Moscow: Russian State Univ. for Humanities, 2010.
- Kyuseva M.V.* Leksičeskaja tipologija semantičeskix sdvigov nazvanij kačestvennyx priznakov “ostrj” i “tupoj” [Lexical Typology of Semantic Shifts in Adjectives “Sharp” and “Blunt”]: diploma paper. Moscow State Univ., 2012.
- Levinson S.C.* Space in Language and Cognition. Singapore: World Publ. Co./CUP, 2008.

Rakhilina E.V. Aspectual Classification of Nouns: A Case Study of Russian // Tense-Aspect, Transitivity and Causativity / W. Abraham, L. Kulikov (eds). Amsterdam: Benjamins, 1999. P. 341–350.

Rakhilina E.V., Prokofjeva I.A. Russkie i pol'skie glagoly kolebatel'nogo dvizhenija: semantika i tipologija [Russian and Polish Verbs of Oscillation: Semantics and Typology] // Jazyk. Ličnost'. Tekst: sb. k 70-letiju T.M. Nikolaevoj / V.N. Toporov (ed.). Moscow: Inst. Slavjanovedenija RAN, 2005. P. 304–314.

Rakhilina E.V., Reznikova T.I. Frejmovyj podxod k leksičeskoj tipologii [Frame Approach to Lexical Typology] // Voprosy jazykoznanija. 2013. No. 2. P. 3–31.

Reznikova T.I., Rakhilina E.V., Bonch-Osmolovskaya A.A. Towards a Typology of Pain Predicates // Linguistics. 2012. Vol. 50. No. 3. P. 421–465.

Semantic and Lexical Universals — Theory and Empirical Findings / C. Goddard, A. Wierzbicka (eds). Amsterdam: John Benjamins, 1994.

Shapiro M.M. Glagoly kolebatel'nogo dviženija v ural'skix jazykax (na materiale finskogo, komi-zyrjanskogo i neneckogo jazykov): semantika i tipologija [Verbs of Oscillation in Uralic Languages (Based on Finnish, Komi-Zyrian and Nenets Languages): Semantics and Typology] // Uralo-altajskie issledovanija. 2015. No. 1 (16). P. 29–52.

The Senses in Language and Culture / A. Majid, S.C. Levinson (eds) // The Senses a. Society. 2011. Vol. 6. No. 1. P. 5–18.

Words and Meanings: Lexical Semantics across Domains, Languages, and Cultures / C. Goddard, A. Wierzbicka (eds). Oxford Univ. Press, 2014.