RepLeCon: A dataset of Russian repetition constructions and its translation[*]

In this paper, we present a new resource RepLeCon (constructions with lexical repetitions) that is being created at St Petersburg University for the study of constructions with lexical repetitions, such as equative and propositional tautologies or lexical clones.

(1) *Drug est' drug*
'A friend is a friend'

(2) *Chto budet, to budet*
'What will be, will be'

(3) *Za oknom – osen'-osen'.*
'Outside the window is an autumn autumn'

The resource under consideration is based on English, French, Spanish, German, and Finnish parallel subcorpora of the Russian National Corpus (henceforth – RNC) and the OPUS2 Russian corpus. It collects Russian repetition constructions enriched with the information described below.

The first task within the project is to develop a general annotation structure and clear instructions for annotators. The constructions in question are searched using the queries proposed in Vilinbakhova and Kopotev (2017), which are then annotated manually. The following five basic categories are taken into account in manual annotation: (1) context description, (2) structure of lexical repetitions, (3) semantics of the constructions, (4) pragmatics of the constructions, and (5) description of their translations.

Context description includes 11 fields, such as full and short context in both source and target language, the subcopus along with its metadata, subcorpus size in tokens, language of the original text and its translation. **The description of the structure** includes repeated tokens in both source and target languages along with their morphological annotation, number of repetitions. **The description of the semantics** includes the following fields: semantic overlapping between repeated elements, type of information, which the construction refers to, referential status indicators, and the idiomatic feature of the construction. **The pragmatic description** includes: speech event structure, modus (oral, written, etc.), type of passage (narrative, descriptive etc.), rhetorical relations of the construction within the context, and markers of such relations. Finally, the **description of the translations** includes one field: source-target correspondence, i.e. whether the translation has formal and/or functional correspondence with the source construction.

The manual annotations in RepLeCon is crosschecked and validated. In (Artstein & Poesio 2008) has been shown that inter-annotator agreement measured with chance-corrected coefficients, i.e. Ķ coefficient, are quite high in structural and semantic annotation, while it is expected to be lower when dealing with pragmatic and discursive phenomena (cf. Spooren & Degand 2010; Grisot 2017). Hence, pragmatic information is a subject for two-step annotation process, which includes discussion in case of disagreement. The discussion turns individual

[*] The research was financially supported by the Russian Science Foundation, project no. 19-78-10048 based at St Petersburg University.

annotators' strategies into cooperative strategies (Spooren & Degand 2010), which leads the analysis to a more consistent annotation.

We started the compilation with the description of (1) tautologies establishing identity, such as *X est' X* 'lit. X is X' (67 annotated examples), structures (2) *X kak X* (46 annotated examples) and (3) *X tak X* (111 annotated examples), (4) conditional tautologies of the type *esli X, to X* 'if X, (then) X'(25 annotated examples), (5) disjunctive structures of the type *P ili ne P* 'P or non-P'(52 annotated examples), (6) relative tautologies of the type *chto P, to P* 'lit. what P, that P' (85 annotated examples), and (7) "grammaticalized" repetitions of the type *ni X, ni Y* (138 annotated examples).

In sum, RepLeCon is expected to provide a useful tool for theoretical linguistic research and its practical applications. Research questions, which are to be investigated, include the universal or language-specific nature of Russian constructions with repetitions; strategies in translating these constructions into the indicated languages as well foreign language stimuli that make the constructions appear in Russian translations; the role of structural and discourse factors in a translator strategy. The resource will also contribute to improving machine translation algorithms, and to teaching Russian as a foreign language.

REFERENCES

Artstein, Ron & Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596.

Grisot, Cristina. 2017. A quantitative approach to conceptual, procedural and pragmatic meaning: Evidence from inter-annotator agreement. *Journal of Pragmatics* 117. 245-263.

Spooren, Wilbert & Liesbeth Degand. 2010. Coding coherence relations: reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2). 241–266.

Vilinbakhova, Elena & Kopotev, Mikhail. 2017. Does "X est' X" mean "X eto X"? Looking for an answer in synchrony and diachrony. *Voprosy Jazykoznanija* 3. 110-124.