ANALYZING GRAMMATICAL EQUIVALENCES IN A RUSSIAN-FRENCH MULTIVARIANT PARALLEL CORPUS

Sylvain Loiseau[1], Dmitri V. Sitchinava[2], Anna A. Zalizniak[3], Igor M. Zatsman[4]

The Russian-French parallel corpus as a part of the Russian National Corpus (http://ruscorpora.ru) is being transformed into a so-called **multivariant corpus** where for each Russian text a set of different alternative French translations is provided. The multivariant corpus will likewise be available online. Concurrently, a Database of functionally equivalent lexico-grammatical verbal forms has been created using the multivariant corpus. One of the main objectives of the database creation is to obtain statistical estimates of different types of equivalences between the Russian and French verbal forms, and, in particular, the impersonal verbal constructions of the kind of 'it seems to me' (Russian *mne kazhetsja,* French *il me semble*).

The technology for creating the multivariant corpus and the database should support the following basic functions:

1) alignment of parallel texts with several variants of translation;
2) morphological annotation and lemmatization of parallel texts;
3) including parallel texts with several variants of translation into the corpus;
4) constructing the database of equivalent lexico-grammatical verbal forms (Russian - French);
5) calculating the statistical estimates of equivalences between Russian and French verbal forms.

The existing works on parallel corpora usually use only one option of translation per language. Lexical and grammatical studies performed on parallel corpora may use alternative options of translations, envisaging all possible correspondences. They can reflect objective variability in the target language and can be a valuable resource for contrastive grammatical description of both languages. For instance, one can expect to observe greater variability amongst the translations when there are structural differences between the two languages (a Russian «signifié» with no clear correspondence in the French grammar and/or lexicon is more likely to exhibit variation in its translations).

Quantitative methodologies for the analysis of mono-, poly- and hyperequivalences and, more generally, for contrastive grammatical description based on aligned corpora has been elaborated. Quantitative methods may be used for identifying the most frequent cases of variability, including an analysis of the regular correspondences between the two linguistic systems. When a variability of translations is observed for a given lexical or grammatical feature, statistical analysis will allow for discovering which contextual features are correlated with each variants. These contextual features help describing the values of each variant. Factorial analysis, and in particular Correspondences analysis, is well suited for this task of identifying the stable groups of contextual features across numerous instances of variants. Statistical analysis is used as a mean for providing a summary of numerous translations (polyequivalence) under scrutiny. While each single context is of no use for drawing conclusions, and while it is also impossible to figure out the big picture when dealing with numerous collected contexts, statistical analysis is useful for mapping variants.

The results of the statystical analysis of the impersonal constructions of the type 'it

[1] Université Paris 13, Sorbonne Paris Cité, Laboratoire LDI (Lexiques, dictionnaires, informatique), CNRS, UMR 7187, sylvain.loiseau@univ-paris13.fr.

[2] Institute of the Russian Language of the Russian Academy of Sciences, mitrius@gmail.com.

[3] Institute of Linguistics of the Russian Academy of Sciences, Institute of Informatics Problems of the Russian Academy of Sciences, anna.zalizniak@gmail.com.

[4] Institute of Informatics Problems of the Russian Academy of Sciences, iz_ipi@a170.ipi.ac.ru.

seems to me' should be presented in the talk.