Предложные конструкции русского языка:
автоматическая обработка ошибок иностранцев

Екатерина Уэтова
НИУ ВШЭ, Россия, euetova@gmail.com

Юлия Перова Нувело
Université Côte d'Azur, BCL (UMR 7320), Франция, juliaperova@gmail.com

Аннотация

Доклад посвящён результатам обработки данных по предложным конструкциям русского языка на материале французского подкорпуса учебных текстов *Russian Learner Corpus* (www.web-corpora.net/RLC).

Ключевые слова: русский язык, корпусная лингвистика, анализ ошибок, русский как иностранный.

Процесс овладения как родным языком, так и иностранными языками, ещё до конца не изучен. С появлением в XXI-м веке электронных корпусов исследования на эту тему входят в новую стадию, открываются новые перспективы в изучении этой области. Несмотря на большое количество существующих учебных корпусов, в которых собран материал для изучения языка иностранных носителей, автоматическая обработка данных пока ещё развита недостаточно хорошо.

Проблема автоматической обработки данных отчасти объясняется отсутствием теоретических работ по аннотированию ошибок. Так, если нормативные тексты, отвечающие всем нормам одного языка, уже почти больше десяти лет свободно размечаются автоматически, – как при помощи орфографической, так и грамматической и отчасти семантической разметки –, то тексты с отклонениями от нормы вызывают немало трудностей при аннотировании, поскольку сам репертуар возможных отклонений кажется безграничным.

Мы решили проверить это предположение на примере использования предложных конструкций студентами-французами, изучающими русский язык как во Франции, так и в России, и постараться выявить при помощи автоматической обработки ошибок, какие ошибки являются наиболее частотными и какие факторы на них влияют. Материалом нашего исследования стали письменные работы с разными уровнями владения русским языком, от начинающих A1 до свободного владения C1-C2. Эти работы собраны во французском подкорпусе *Russian Learner Corpus* (34.755 слов).

Для эффективной обработки все предложные конструкции были разделены на два типа: конструкции с управлением (Gov) и независимые предложные конструкции (Constr). После этого было принято отмечать только три случая отклонений с предлогами: это i) Extra – лишний предлог, ii) Miss – пропущенный предлог, iii) Subst – заменённый неправильный предлог. Кроме этого, при аннотировании показалось релевантным сразу отмечать возможные факторы, влияющие на ошибку в предложной конструкции. Среди таких факторов мы чаще всего встретили случаи калькирования французского словарного эквивалента предлога (Transfer) и случаи наложения, когда носитель путал, или совмещал, две конструкции (Fusion). Кроме тэгов, для более эффективной обработки ошибок использовались комментарии, где отмечалась дополнительная информация, например, о типе конструкций (конструкция времени, места, причины или конструкция с управлением существительным, глаголом) и о самих предлогах, количество которых, как известно, лексически ограничено.

Кроме аннотирования ошибок, были привлечены к исследованию дополнительные данные о самих носителях языка. Эти данные позволили проследить, как влияют на использование конструкций с предлогами такие факторы, как уровень владения языком, пол, языковой background (эритажный или иностранный язык).

Результаты такого исследования показали, что больше всего ошибок в конструкциях с глагольным управлением (типа "Часто я также **езжу в море**.") и в конструкциях времени (например, "**На вечером** я смотрю телевизор или я учюсь."). Студенты чаще неверно используют предлоги *на, в, для, с, о*. И чаще используют не тот предлог там, где нужны предлоги *в, на, по, о* и *за*. Оказалось, что предлоги *в, на, о* и падежи локатив, аккузатив, генетив и датив являются наиболее трудными для студентов-французов.

Эти и другие первые результаты этого исследования будут наглядно проиллюстрированы на графиках и примерах полученных в результате автоматической обработки деревьев решений и случайных лесов. Несмотря на возрастающее количество учебных корпусов, подобные исследования, по нашим данным, пока ещё не проводились. Результаты такого рода могут быть отправной точкой для компаративных русско-французских исследований по предлогам, а также могут служить ценным материалом для преподавателей русского как иностранного, особенно во франкоязычной аудитории. С развитием корпуса станется только выяснить, насколько окажутся универсальны такие ошибки.

Библиография

1. Рахилина, Е. В. *Грамматика ошибок и грамматика конструкций: «эритажный» («унаследованный») русский язык* / Е. В. Рахилина, А. С. Выренкова, М. С. Полинская // Вопросы языкознания. — 2014. — № 3. — С. 3–19.
2. Díaz-Negrillo, Ana, and Salvador Valera. *A learner corpus-based study on error associations 1*, Procedia-Social and Behavioral Sciences 3, 2010: 72-82.
3. Lee, Lung Hao, et al., *A tagging editor for learner corpora annotation and error analysis,* 22nd International Conference on Computers in Education, ICCE 2014. Asia-Pacific Society for Computers in Education, 2014.
4. Hana, Jirka, et al., *Error-tagged learner corpus of Czech*, Proceedings of the Fourth Linguistic Annotation Workshop. Association for Computational Linguistics, 2010.
5. Sinclair, John McHardy, ed., *How to use corpora in language teaching*, Vol. 12. John Benjamins Publishing, 2004.
6. López, Willelmira Castillejos, *Error analysis in a learner corpus. What are the learners' strategies*, 1998.
7. Granger, Sylviane, Anne Vandeventer, and Marie-Josée Hamel. *Analyse de corpus d'apprenants pour l'ELAO basé sur le TAL.*, Revue. Volume 1.1 (1998).
8. Thewissen, Jennifer, *Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus*, The Modern Language Journal 97.S1 (2013): 77-101.
9. Corder, Stephen Pit., *The significance of learner's errors*, IRAL-International Review of Applied Linguistics in Language Teaching 5.1-4 (1967): 161-170.
10. Yang, Wenxing, and Ying Sun., *Dynamic development of complexity, accuracy and fluency in multilingual learners' L1, L2 and L3 writing*, Theory and Practice in Language Studies 5.2 (2015): 298.
11. Dagneaux, Estelle, Sharon Denness, and Sylviane Granger, *Computer-aided error analysis*, System 26.2 (1998): 163-174.
12. Facchinetti, Roberta, *Corpus linguistics 25 years on*., BRILL, 2007.
13. Hawkins, John A., and Paula Buttery, *Criterial features in learner corpora: Theory and illustrations*, English Profile Journal 1.1 (2010).
14. Granger, Sylviane, *Error-tagged learner corpora and CALL: A promising synergy*, CALICO journal (2003): 465-480.