Автоматический сбор данных для исследований по лексической типологии

Результаты исследования в области лингвистической типологии считаются достаточно надежными в том случае, если они получены на материале репрезентативной языковой выборки, включающей не менее 100 языков, максимально разнородных ареально и генетически ([Bell 1978], [Rijkhoff, Bakker 1998]). Однако для лексической типологии такой идеал практически недостижим, что связано с особой трудоемкостью процесса сбора надежных сведений о значении и сочетаемости лексических единиц. В настоящем докладе мы предлагаем алгоритм частичной автоматизации этого этапа лексико-типологического исследования, внедрение которого в работу лексического типолога позволило бы проводить анализ большего количества языков в более сжатые сроки.

В качестве теоретической базы мы выбираем фреймовый подход к лексической типологии ([Рахилина, Резникова 2013]), разрабатываемый Московской лексико-типологической группой MLexT и восходящий к традициям Московской Семантической Школы (см. [Апресян 1974]). Этот подход подразумевает сравнение лексики разных языков через сопоставление синтагматических характеристик рассматриваемых лексических единиц. Основной инструмент исследования в рамках данной парадигмы - анкета, состоящая из прототипических контекстов, в которых могут употребляться лексемы изучаемого поля. Анкеты переводятся на каждый язык, включаемый в выборку, и заполняются по результатам опросов носителей, а также анализа словарных и корпусных данных.

Мы предлагаем алгоритмы автоматизированного решения двух задач:

1. перевода анкеты с русского (или английского) языка на язык L;
2. заполнения анкеты данными языка L по доступным одноязычным корпусам текстов.

Мы провели серию пилотных экспериментов с семантическими полями качественных признаков ‘острый’, ‘гладкий’, ‘толстый’ и ‘тонкий’, которые уже были проанализированы вручную участниками MLexT на материалах значительных языковых выборок (от 10 до 30 языков, см. [Кюсева 2012], [Кашкин 2013], [Козлов и др. 2016]). Для этих полей уже были разработаны типологические анкеты, в большинстве случаев сводящиеся к спискам прилагательных, относящихся к рассматриваемому полю, и спискам существительных, в сочетании с которыми эти прилагательные могут употребляться.

Наш метод автоматизации процесса решения задачи (1) опирается на машиночитаемые словари группы FreeDict (http://www.freedict.org/ru/). Преимущество этих словарей заключается в том, что возможные переводы в них размечены по тому, к какому значению исходного слова они относятся. Наш алгоритм выбирает переводные эквиваленты только для первого значения исходного прилагательного, а потом проводит дополнительную проверку по методу обратного перевода: найденное слово переводится на исходный язык и включается в итоговый список только в том случае, если прилагательное, являющееся его эквивалентом в первом значении, входит в стартовый список признаковых слов. Существительные переводятся по той же методологии, но с небольшой модификацией: в случае, если существительного нет в словаре FreeDict, перевод для него подбирается по соответствующему машиночитаемому словарю компании Яндекс.

Далее, для решения задачи (2), анкета переводится в табличный формат, где в качестве столбцов выступают прилагательные, а в качестве строк – существительные. Полученная таблица заполняется по материалам доступных корпусов: если прилагательное встречается в комбинации с тем или иным существительным в корпусе, для этой пары подсчитывается значение метрики взаимной информации ([Pothos, Juola 2007]). Комбинации с отрицательным значением метрики взаимной информации считаются случайными и исключаются из итоговой заполненной анкеты.

В докладе мы предполагаем обсудить преимущества и недостатки данного алгоритма и предложить возможные пути его дальнейшей оптимизации.

Литература

Bell A. 1978. Language samples. *Universals of Human Language, Method and Theory*. Greenberg J.H., Ferguson C.A., Moravcsik E.A. (eds.). Stanford Univ Press, Palo Alto, CA. Vol 1, pp. 123–156.

Pothos E.M., Juola P. 2007. Characterizing linguistic structure with mutual information. British Journal of Psychology, Vol 98, issue 2, pages 291–304.

Rijkhoff J., Bakker D. 1998. Language sampling. *Linguistic Typology* 2(3):263–314.

Кашкин Е.В. 2013. Языковая категоризация фактуры поверхностей (типологическое исследование наименований качественных признаков в уральских языках). Дисс.канд. филол. наук. М.: МГУ

Козлов А.А., А.В. Кухто, М.Ю. Привизенцева. 2016. О реальности семантического поля: лексико-типологический подход // В кн.: ACTA LINGUISTICA PETROPOLITANA. Труды Института лингвистических исследований РАН / Отв. редактор Н. Н. Казанский. Т. XII. Ч. 2. Материалы Десятой конференции по типологии и грамматике для молодых исследователей (2013 г.) / Отв. ред. тома Д.В. Герасимов. СПб.: Наука. С. 522-533.

Кюсева М.В. 2012. Лексическая типология семантических сдвигов названий качественных признаков ‘острый’ и ‘тупой’. Дипломная работа. М.: МГУ.

Рахилина Е.В., Резникова Т.И. 2013. Фреймовый подход к лексической типологии. [*Вопросы языкознания*](https://publications.hse.ru/articles/?mg=61287017)*.* № 2. С. 3-31.