Рыжова Д.А. (Москва)

Ершов И.А. (Москва)

Мельник А.А. (Москва)

**Автоматический сбор данных для исследований по лексической типологии**

**Аннотация:**

В статье описывается ряд пилотных экспериментов по разработке алгоритма автоматического перевода и заполнения анкет для типологических исследований признаковой лексики. Перевод анкеты предлагается осуществлять с помощью машиночитаемых переводных словарей и затем заполнять переведенные анкеты материалами доступных одноязычных корпусов.

**Ключевые слова**: лексическая типология, синонимы, фрейм, контекст, автоматический перевод, корпусные исследования, автоматизация

**Automatic data collection for lexical typological research**

**Annotation:**

The paper describes a range of pilot experiments on the development of an algorithm that would automatically collect relevant data for research in lexical typology. The task comprises two stages: questionnaire translation and its further population with lexical data. We suggest to complete the first task with the help of machine-readable dictionaries, and the second task - on the basis of monolingual text corpora.

**Keywords**: lexical typology, synonyms, frame, context, automatic translation, corpus, automatization.

**0. Введение**

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. Известно, что значения слов не обсуждаются в грамматиках, а в толковых и переводных словарях представляются недостаточно структурированно и, в большинстве случаев, не очень подробно. Это вынуждает исследователей разрабатывать специальные анкеты и собирать команду экспертов по различным языкам, способных провести работу с носителями и проанализировать полученный материал. Трудоемкость всего процесса не позволяет проводить подробный анализ обширных семантических зон в большом количестве языков. Поэтому, в большинстве случаев, приходится серьезно ограничивать либо количество языков в выборке, либо степень подробности их анализа.

Одним из возможных способов решения этой методологической проблемы может стать хотя бы частичная автоматизация процесса сбора и последующей обработки лексического материала. И действительно, задача разработки компьютерных методов анализа лексики становится все более популярной. Намечено по меньшей мере два различных методологических направления, первое из которых предполагает автоматическую обработку словарных данных [Youn et al. 2016], а второе основывается на компьютерном анализе параллельных корпусов [Wälchli, Cysouw 2013].

Наиболее яркое исследование, направленное на сопоставление лексики разных языков на основе автоматического анализа данных переводных словарей, представлено в недавней статье [Youn et al. 2016]. В работе выявляется степень семантической близости между базовыми предметными концептами из списка Сводеша (ср. ‘трава’, ‘небо’, ‘солнце’, ‘день’ и под.). Семантическая близость определяется через частоту колексификации, т.е. близкими считаются те концепты, которые во многих языках обозначаются одним и тем же словом (например, ‘день’ и ‘солнце’ семантически значительно ближе друг к другу, чем ‘день’ и ‘трава’, ср., в частности, венгерское *nap*, которое объединяет два этих значения). Частота объединения вычисляется по переводным словарям на основе аккуратно составленной репрезентативной выборки из 81 языка. Выявленные связи между концептами представляются в виде взвешенного графа. Тем самым, в работе проводится полностью автоматизированный типологический анализ сразу нескольких предметных полей на материале значительного количества ареально и генетически разнородных языков. В качестве базы для сопоставления слов из разных языков (т.е. своеобразного аналога лексико-типологической анкеты) используется набор словарных подзначений, составляемый автоматически в процессе обработки словарей.

Исследование [Wälchli, Cysouw 2012] посвящено типологическому анализу основных глаголов движения (‘идти’, ‘ехать’, ‘приходить’, ‘приезжать’ и под.) путем автоматической обработки параллельного корпуса текстов. Здесь в качестве базы для сравнения слов используются не словарные (под)значения, а закономерности дистрибуции, т.е. сопоставляются наборы контекстов, в которых употребляются изучаемые глаголы. Исследование проводится на материале корпуса переводов Евангелия от Марка на 100 языков. Выделяется 360 контекстов, в которых могут быть употреблены базовые глаголы движения, и этот набор служит аналогом лексико-типологической анкеты, которая автоматически заполняется по тексту каждого из переводов Евангелия от Марка. Полученное в результате пространство визуализируется с помощью метода многомерного шкалирования, позволяющего уменьшить размерность исходного пространства, минимизировав при этом количество потерянной информации (о методе многомерного шкалирования см., например, [Cox and Cox 2001]). Таким образом получается ряд проекций исходного пространства на плоскость, которые отображают степень типологической близости рассматриваемых контекстов: чем меньше расстояние между контекстами (т.е. пунктами анкеты) на карте, тем выше вероятность того, что в них в том или ином языке будет употреблено одно и то же слово.

Несомненное достоинство каждого из этих методов анализа лексики заключается в том, что они полностью автоматические, что и позволяет им учитывать материал обширных и разнообразных языковых выборок. Однако оба они связаны с рядом методологических проблем другого характера. Так, например, словарные данные, как уже было сказано выше, не очень надежны в силу своей плохой сопоставимости: все включаемые в исследование словари оказываются разного объема и разного качества. Вероятно, это не очень существенно для анализа базовой предметной лексики (будучи базовой и частотной, она должна быть представлена в словаре практически любого объема и степени подробности), но для исследования любой другой семантической зоны применить этот метод будет крайне затруднительно. Аналогично, параллельный корпус переводов Евангелия может быть достаточной базой для сопоставления основных глаголов движения, но для многих других классов лексики такого материала будет заведомо мало.

Метод, который предлагаем мы, основывается на Фреймовом подходе к лексической типологии (см. [Рахилина, Резникова 2013], а также раздел 1 настоящей статьи), который, в отличие от двух только что рассмотренных методик, подразумевает комплексную обработку вручную доступных ресурсов всех типов: словарей, одноязычных и параллельных корпусов и результатов опросов носителей. Эта методология, направленная на анализ лексики любого типа (не только предметной, не только базовой и не только частотной) оказывается на порядок сложнее, чем методики, основывающиеся либо только на словарях, либо только на параллельных корпусах. Исследование в рамках этой парадигмы включает в себя несколько этапов:

1. Разработку анкеты;
2. Заполнение анкеты;
3. Составление семантических карт на основе собранных материалов.

Каждая из этих задач, традиционно выполняемых вручную большой командой экспертов, может быть по крайней мере частично переведена на автоматическую основу. В настоящей статье мы будем обсуждать возможные алгоритмы автоматизации только одного из этапов исследования в рамках этой парадигмы, являющегося, однако, наиболее трудоемким – процесса сбора материала по заданной заранее анкете[[1]](#footnote-1). Тем самым, мы существенно сужаем задачу по сравнению с рассмотренными выше методиками, подразумевающими полный лексико-типологический анализ, но надеемся, что это позволит нам создать инструмент для более детального анализа лексики, в том числе значительно менее базовой и частотной, чем предметные имена из списка Сводеша или основные глаголы движения.

Статья имеет следующую структуру. В разделе 1.1 мы введем основные принципы и понятия Фреймового подхода к лексической типологии, который мы используем в нашем исследовании в качестве теоретической базы. Раздел 1.2 посвящен более подробному описанию типологических анкет, которые подлежат заполнению; здесь же формулируются задачи, которые должен решать предлагаемый нами алгоритм. В разделе 1.3 дается описание наших экспериментальных данных и золотого стандарта, с которым мы сравниваем результаты работы компьютерных моделей. Раздел 2 посвящен описанию алгоритмов автоматического перевода и заполнения исходных анкет и оценке качества работы этих алгоритмов. В Заключении подводятся итоги: обсуждаются достоинства и недостатки предложенного метода и возможные пути его дальнейшего развития.

**1 Фреймовый подход к лексической типологии**

**1.1 Основные принципы и понятия**

Фреймовый подход к лексической типологии ([Рахилина, Резникова 2013]), который мы используем в качестве теоретической базы для нашего исследования, разрабатывается Московской лексико-типологической группой (MLexT) и восходит к традициям Московской Семантической Школы (см. [Апресян 1974]). Этот подход подразумевает сравнительный анализ внутри- и межъязыковых квазисинонимов на основе сопоставления их сочетаемостных свойств. Так, например, в зоне признаков фактуры поверхностей в русском языке различие в семантике слов *гладкий* и *скользкий*, говоря очень грубо, сводится к тому, что *гладкий* обозначает отсутствие неровностей, ощущаемое тактильно, а *скользкий* описывает практически то же свойство, но подчеркивает, что оно является потенциальной причиной тех или иных неудобств (на скользкой поверхности можно поскользнуться, скользкий предмет может выскочить из рук). Это противопоставление хорошо заметно при анализе контекстов употребления соответствующих слов, ср. примеры 1 и 2 из Национального корпуса русского языка. В некоторых языках (например, в хантыйском[[2]](#footnote-2)) это различие не лексикализовано, т.е. в обоих типах контекстов употребляется одна и та же лексема (в случае с хантыйским языком это прилагательное *woł’ǝk* ‘гладкий/скользкий’).

1. *Днем мужики выкатывали из куги бочонки и кадки с капустой, по* ***гладким доскам*** *спускали их в подвал.* [Виктор Астафьев. Последний поклон (1968-1991)]
2. *Она все-таки упала, не удержавшись на* ***скользком полу****, и больно ударилась коленкой.* [Наталья Александрова. Последний ученик да Винчи (2010)]

Минимальные группы контекстов (т.е. минимальные типы ситуаций), на которые может реагировать язык, выделяя для них специальную лексему, называются фреймами. Возвращаясь к нашему примеру из поля фактуры поверхностей, можно сказать, что в нем выделяются, в числе прочих, фреймы ‘отсутствие неровностей, воспринимаемое наощупь’ и ‘отсутствие неровностей, воспринимаемое наощупь и причиняющее неудобство’ (подробнее о фреймовой структуре этого поля см. [Кашкин 2013]). В русском языке для каждого из этих фреймов есть отдельное слово, а в говоре села Теги хантыйского языка эти фреймы объединяются в рамках одного лексического средства. Результатом исследования в рамках этой парадигмы является выявление набора фреймов для изучаемого поля и определение степени близости между ними, т.е., говоря упрощенно, установление вероятности объединения тех или иных фреймов в рамках одного лексического средства в некотором языке L.

Поскольку анализ сочетаемости – основной метод анализа семантических особенностей слов в рамках фреймового подхода к лексической типологии, основным инструментом исследования в этой парадигме становится анкета, состоящая из наиболее показательных, прототипических контекстов, в которых могут употребляться лексемы изучаемого поля. Как правило, анкеты составляются по материалам подробного анализа трех-пяти языков и содержат по несколько иллюстраций на каждый фрейм. Такие анкеты используются при сборе данных, и они же являются базой для сравнения полученного материала. Поскольку они состоят преимущественно из лингвистических контекстов, а не из экстралингвистических стимулов (картинок, видеоклипов и проч.), они переводятся на каждый язык, включаемый в выборку, и заполняются по результатам опросов носителей, а также анализа словарных и корпусных данных.

**1.2 Типологические анкеты и задача их автоматического заполнения**

Контекстные анкеты могут быть разной степени сложности. Наш опыт работы с лексикой показывает, что для разных классов слов требуется разный объем «диагностического» контекста - контекста, по которому можно было бы однозначно определить, к какому фрейму он относится. Если для глаголов направленного движения необходимо учитывать много факторов (свойства субъекта перемещения, начальной и/или конечной точки и др., см. [Кузьменко, Мустакимова 2015], [Reznikova and Vyrenkova 2015]), то для прилагательных в большинстве случаев достаточно учитывать минимальный контекст, а именно, семантические особенности определяемого существительного, т.е. носителя данного признака (см., например, [Рахилина и др. 2010]). По этой причине именно признаковые поля являются удобным материалом для пилотных экспериментов в области автоматизации процесса сбора лексико-типологических данных.

Анкета для исследования той или иной признаковой зоны может быть представлена в виде таблицы, в строках которой располагаются существительные, а в столбцах – изучаемые прилагательные (см. Таблицу 1). При заполнении анкеты в те ячейки, которые соответствуют допустимым (грамматичным, осмысленным и естественным) сочетаниям «прилагательное + существительное», ставится «1» (или «+»), а ячейки, которые соотносятся с незасвидетельствованными сочетаниями, получают значение «0» (или «-»). Например, в анкете для поля фактуры поверхностей, заполненной материалом русского языка, фрагмент которой представлен в Таблице 1, отмечено, что русск. *дорога* сочетается с прилагательными *скользкий, ровный, гладкий* и *прямой*, а *ступенька* -- с прилагательными *скользкий, ровный* и *гладкий*.

Таблица 1. Фрагмент анкеты для признакового поля фактуры поверхностей.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | скользкий | ровный | плоский | гладкий | прямой |
| дорога | + | + | - | + | + |
| ступенька | + | + | - | + | - |
| ... |  |  |  |  |  |

Заполнение таких анкет требует очень много усилий со стороны исследователя. Задача нашей работы – предложить набор алгоритмов, который будет позволять обследовать по крайней мере часть источников автоматически и, тем самым, расширять эмпирическую базу таких исследований, одновременно сокращая количество времени, потраченное на сбор релевантных данных.

Формально эта задача сводится к двум подзадачам:

1. Переводу анкеты на нужный язык;
2. Заполнению этой анкеты сведениями о том, могут ли рассматриваемые прилагательные и существительные в данном конкретном языке сочетаться друг с другом.

Поскольку в нашем случае анкета состоит из списков прилагательных и существительных, задача их перевода на первый взгляд кажется очень простой: возникает впечатление, что достаточно найти основной переводной эквивалент для каждого слова из анкеты. Однако на деле эта работа сопряжена с целым рядом сложностей, связанных, в первую очередь, с переводом прилагательных. С одной стороны, в разных языках одно и то же семантическое поле может обслуживаться разным количеством лексем, и при переводе важно выявить все слова, относящиеся к данной зоне в целевом языке. С другой стороны, включать в анкету все обнаруженные переводные эквиваленты исходных прилагательных нам бы тоже не хотелось, поскольку существенная их часть обычно относится только к переносным употреблениям слов из изначального списка, т.е., строго говоря, не входит в интересующее нас семантическое поле. Так, например, в число переводных эквивалентов русского прилагательного *острый* в английском языке[[3]](#footnote-3) входят, наряду с лексемами *sharp* и *pointed*, такие слова как *critical, quick* или *strong*, являющиеся подходящими переводами для слова *острый* в контекстах *острое замечание, острый ум/зрение/слух* или *острые ощущения* соответственно, но при этом явно не относящиеся к семантическому полю ‘острый’. Таким образом, задача перевода анкеты осложняется проблемой аккуратного отсева нерелевантных лексем.

**1.3 Материал исследования**

В качестве материала для экспериментов мы используем уже созданные вручную и апробированные на достаточно обширном языковом материале анкеты для четырех признаковых полей: ‘острый’, ‘толстый’, ‘тонкий’ и ‘гладкий’ (см. Кюсева 2012, Козлов и др. 2016, Кашкин 2013). Изначально эти анкеты состояли из списков русских прилагательных и существительных. Мы дополнительно перевели их вручную на итальянский и английский языки и для этих языков заполнили их традиционным способом: по материалам словарей, корпусов и опросов носителей. Таким образом мы получили золотой стандарт, с которым в дальнейшем сравнивали качество работы наших алгоритмов. Там, где это было возможно, мы использовали в качестве золотого стандарта в том числе и данные, собранные ранее также вручную авторами исследований этих семантических полей и представленные в едином формате Базы данных признаковой лексики (см. [Кюсева и др. 2013]). В результате англоязычные анкеты мы использовали в качестве входных данных наравне с русскоязычными, а для итальянского языка реализовали полный цикл работы алгоритма и провели оценку полученных результатов. Материалы других языков привлекались в качестве дополнительных в процессе тестирования тех или иных этапов работы алгоритма.

**2 Алгоритмы автоматического заполнения типологической анкеты**

Как мы уже сказали в разделе 1.2, задача автоматического заполнения типологической анкеты включает в себя две подзадачи: перевод анкеты с русского (или английского) языка на язык L и собственно заполнение переведенной анкеты. В разделе 2.1 мы предлагаем обзор нескольких способов решения задачи перевода и обсуждение их достоинств и недостатков. В разделе 2.2 приводится описание используемого нами метода автоматического заполнения анкеты. Оценка результатов применения описываемых методов к реальным данным приводится в разделе 3.

**2.1 Перевод анкеты**

Мы опробовали два метода решения задачи перевода: с помощью онлайн-переводчиков, предлагаемых Yandex и Google, и через машиночитаемые двуязычные словари. Рассмотрим последовательно каждый из них.

**2.1.1 Перевод через онлайн-переводчики**

В рамках нашего исследования мы провели ряд экспериментов по переводу анкеты для нескольких признаковых полей на основе онлайн-переводчиков, реализованных компаниями Google и Yandex. Особое преимущество этого метода заключается в том, что он допускает не только пословный перевод прилагательных и существительных из наших списков, но и перевод целых строк типологической анкеты. Потенциально реализация этого метода могла бы позволить решить сразу обе наши подзадачи: и перевода анкеты, и ее заполнения.

Мы начали наши эксперименты с более простой задачи перевода отдельных слов, без опоры на сочетаемость. В таком случае при переводе прилагательных мы сталкиваемся с определенным затруднением. Как уже было сказано выше, количество признаковых лексем целевого языка может отличаться от количества слов, обслуживающих данную зону в исходном языке, поэтому у нас нет необходимости выбирать ровно один переводной эквивалент для каждого прилагательного. Однако общее количество синонимов, предлагаемых онлайн-переводчиками, оказывается очень большим, и отнюдь не очевидно, что каждый из них нам подходит. Мы пробовали решить проблему отсева ненужных слов с помощью обратного перевода. В переводчике с целевого языка мы переводим на русский все синонимы к переведенным с русского на целевой язык словам из ряда и оставляем в анкете целевого языка только те прилагательные, в первом значении которых присутствует хотя бы одно из прилагательных исходного языка.  
 Например, мы хотим узнать, какие прилагательные составляют поле ‘острый’ в финском языке. Тогда, задав онлайн-переводчику компании Yandex в качестве исходного слова русское прилагательное *острый*, мы получаем такие переводы (8 финских слов): *akuutti, äkillinen, terävä, pistävä, kirpeä, pureva, veitsenterävä, tarkka*. Далее мы переводим каждое из полученных слов обратно на русский язык и оставляем только те, у которых указано значение *острый* в качестве первого значения. Таким образом мы выберем 5 финских прилагательных *akuutti, terävä, pistävä, kirpeä, veitsenterävä* и исключим из рассмотрения 3 оставшихся *(äkillinen, pureva* и *tarkkea)*, поскольку в первом значении у них указаны переводы *внезапный*, *кусающийся* и *точный, аккуратный, пунктуальный* соответственно). Согласно золотому стандарту, в финском языке в поле *острый* входят прилагательные *terävä, pistävä*, и мы, конечно, нашли их оба, но также зафиксировали три лишних прилагательных *akuutti* (сравнимо с русск. *острый* в *острый взгляд*)*, kirpeä* (*острый* в значении *пикантный*)*, veitsenterävä* (встречаются окказиональные употребления).

В качестве другого возможного способа отсева лишних прилагательных мы протестировали метод перевода прилагательного в контексте с существительным. Для этого мы попробовали перевести все допустимые в русском языке сочетания прилагательного и существительного из анкеты с помощью онлайн-переводчиков. Чтобы сравнить результаты, полученные этим способом, с результатами, приведенными выше, переведем снова русское прилагательное *острый* на финский язык, но уже в сочетании с каждым из рассматриваемых существительных (*острый нож*, *острый каблук, острая стрела*). Из каждого словосочетания на финском выберем прилагательное и с помощью обратного перевода аналогичным способом отберем только те, у которых указано прилагательное *острый* в качестве основного переводного эквивалента. В результате мы получаем только два финских прилагательных, одно из которых оказывается нужным (*terävä*), а другое (*akuutti*) – ненужным. Таким образом мы уменьшим количество ненужных прилагательных, но в то же время и уменьшим количество нужных.

**2.1.2 Машиночитаемые словари freedict и verdict**

Наш второй метод перевода анкет основан на использовании машиночитаемых двуязычных словарей, а именно, переводных словарей группы Freedict[[4]](#footnote-4) и электронных словарей Verdict, подготовленных компанией Yandex[[5]](#footnote-5). В этом случае возможен исключительно пословный перевод каждого существительного и прилагательного из заданных списков.

Двуязычные словари Freedict хранятся в формате .tei[[6]](#footnote-6) (который является подмножеством языка xml), структура документа определяется парными тегами. Каждое вхождение выделяется тегом <entry>. В него вложены теги <form> и <sense>. В зависимости от количества возможных переводов, может быть несколько вложений с тегом <sense>: в таком случае рядом с тегом записывается показатель n="…", в котором в кавычках указывается номер возможного перевода. Таким образом создается структура, отображающая связь между запрашиваемым словом данного языка и его возможными переводными эквивалентомы. В тег <form> вложены теги <orth> и <pron>. Первый отображает слово, записанное в орфографии данного языка. Второй – его транскрипцию. В тег вкладывается тег <cit> (цитирование другого источника), внутри которого с помощью парного тега <quote> записывается переводной эквивалент в орфографии целевого языка. Произношение переводных эквивалентов не указывается.

Мы предполагаем, что перевод с показателем n="1" является основным, соответствующим прямому употреблению заданного слова, а остальные переводы связаны с его переносными значениями. Именно поэтому в качестве перевода для исходного прилагательного мы выбираем эквивалент с показателем n="1". Далее, как и в случае с онлайн-переводчиками, производится проверка методом обратного перевода. В качестве исходного прилагательного задается полученный переводной эквивалент и осуществляется перевод на стартовый язык. Если слово входит в начальный список прилагательных, то данный переводной эквивалент включается в итоговый список слов. Для перевода существительных проводится та же операция. Однако, если искомого слова нет в словаре Freedict, то используется соответствующий двуязычный словарь Verdict. Он имеет более простую структуру. Одно вхождение занимает одну строку и имеет всего четыре позиции, разделенные табуляцией: исходное слово, его часть речи, его переводной эквивалент и ссылка на словарь, из которого данная информация получена. Переведенные слова записываются в csv-таблицу: в первый столбец таблицы записываются существительные на языке, с которого осуществлялся перевод, во вторую - на том языке, на который мы переводим анкету, а в первую строку - прилагательные. В клетки пересечения между прилагательным и существительным мы впоследствии записываем информацию о сочетаемости соответствующих единиц (см. Таблицу 2).

Таблица 2. Пример анкеты для признаковых полей ‘толстый’ и ‘тонкий’, переведенной на французский.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | maigre | gros | mince | dense | étroit | ... |
| лист | feuille |  |  |  |  |  |  |
| слой | poussière |  |  |  |  |  |  |
| книга | livre |  |  |  |  |  |  |
| ... |  |  |  |  |  |  |  |

**2.1.3 Онлайн-переводчики или машиночитаемые словари: анализ результатов**

Мы рассмотрели два возможных метода перевода анкеты. Метод онлайн-переводчиков хорош тем, что он позволяет сразу охватить большое количество языков, на вход можно сразу подавать как отдельные слова, так и целые словосочетания из исходных анкет, и он довольно прост в использовании. Тем не менее, у этого метода есть ряд существенных недостатков. Во-первых, он непрозрачен и непостоянен: технологии перевода у компаний Yandex и Google непрерывно меняются, и мы не можем быть уверены, что в любой момент времени сможем получить результаты, сопоставимые с теми, что мы получили в рамках наших пилотных экспериментов. Во-вторых, метод перевода целых словосочетаний оказывается бесполезным для нашей задачи: почти все пары «прилагательное + существительное» переводятся с помощью одного, основного элемента признакового поля, и тем самым сразу несколько периферийных прилагательных рассматриваемого поля оказываются неучтенными. Наконец, в-третьих, разные стратегии перевода с помощью онлайн-переводчиков дают для разных полей разные результаты.

Второй метод – метод машиночитаемых словарей – лишен этих недостатков. Более того, словари группы Freedict обладают дополнительным преимуществом: эксплицитно противопоставляют переводные эквиваленты для прямых и переносных значений исходных слов. Конечно, данный метод имеет свои недостатки, связанные, в первую очередь, с небольшим количеством таких словарей и ограниченностью их объёма. Так, например, не очень частотное прилагательное из рассматриваемого поля в такой словарь не попадет, а значит, предлагаемый нами алгоритм не сможет его выявить. Как видно, у обоих методов есть свои достоинства и свои недостатки. Первый метод подлежит дальнейшей проверке на материале большего количества анкет, второй позволяет получить приемлемые результаты (см. раздел 3 настоящей статьи), но только для тех языков, по которым мы имеем необходимые словари. В рамках данного исследования мы остановились на методе 2: для выбранных нами языков он кажется более надежным и перспективным, особенно с учетом того, что база переводных словарей может в будущем пополняться.

**2.2 Заполнение анкеты**

Одна из основных задач исследования в области лексической типологии – это оценка возможности употребления слов описываемого ряда в различных контекстах. Имея исходные русскоязычные списки прилагательных и существительных, с которыми эти прилагательные потенциально могут сочетаться, а также полученные на первом шаге работы нашего алгоритма соответствующие им переводы на другие языки, мы можем автоматически определять дистрибутивные свойства каждого признакового слова по отношению к выделенным предметным именам. Для этого мы используем одноязычные корпуса серии WaC. Предложения в корпусах этого типа хранятся в формате .xml и структурированы по принципу xml-разметки. Каждое предложение выделяется тегом <s>. Внутри тега прописываются словоформы, которые содержит данное предложение. Информация о каждой словоформе занимает одну строку и содержит словоформу, лексему, морфологическую информацию о данной словоформе, номер данной словоформы в предложении, номер синтаксической вершины, к которой относится данное слово, и его синтаксическую категорию, разделенные табуляцией.

Поскольку в выбранных нами полях изучаемые слова -- прилагательные, а их диагностические контексты -- существительные, для того, чтобы оценить вероятность их совместной встречаемости, мы ищем вхождения всех комбинаторно возможных биграмм “прилагательное + существительное” из анкеты в корпусе исследуемого языка. Для этого мы ищем такие лексемы, леммы которых совпадают с искомыми (поскольку искомое существительное может встретиться в тексте как, скажем, во множественном числе, так и в единственном), которые находятся на расстоянии один друг от друга[[7]](#footnote-7).

Для того, чтобы исключить окказиональные и в целом нехарактерные для языка сочетания, мы вычисляем для каждого биграмма характеризующую его меру взаимной информации (mutual information, далее MI), показывающую, насколько совместная встречаемость двух слов статистически значима в данном языке. Введение порога для MI позволяет оставить контекстно-связанные слова и исключить такие, которые приравниваются к случайно встретившимся рядом в тексте. Значение этой метрики вычисляется по следующей формуле:

, где

p(x,y) - вероятность встретить прилагательное с существительным рядом в корпусе,

p(x) - вероятность встретить прилагательное в корпусе,

p(y) - вероятность встретить существительное в корпусе.

К примеру, если мы хотим оценить допустимость употребления в итальянском языке существительного *corridoio* ‘коридор’ в сочетании с прилагательным *stretto* ‘узкий’, то мы будем искать в корпусе два идущих подряд слова с леммами *corrido* и *stretto*. В корпусе ItWaC находится 47 таких вхождений с мерой MI равной приблизительно 8.7, достаточной для того, чтобы считать сочетание *corridoio stretto* допустимым при взятом нами пороге равном нулю[[8]](#footnote-8).

**3 Оценка результатов**

Как уже было сказано выше, мы реализовали полный цикл работы нашего алгоритма на материале итальянского языка и оценили результаты, сравнив автоматическое заполнение анкет с золотым стандартом. Объединив лучший метод перевода с помощью электронных словарей и автоматическое заполнение анкет с использованием корпусов при пороге MI, равном нулю, в конечном счете мы получаем достаточно высокие показатели для метрик полноты (доля сочетаний, отмеченных допустимыми среди действительно допустимых) и точности (доля действительно допустимых сочетаний среди нами отмеченных допустимыми): 0.93 и 0.79 соответственно. Высокое значение метрики полноты позволяет говорить о том, что мы находим практически все допустимые значения, а достаточно высокая точность указывает на относительно небольшое количество сочетаний, которые мы считаем допустимыми, но которые таковыми не являются.

**4 Заключение**

Итак, мы провели серию пилотных экспериментов в области автоматизации процесса сбора лексико-типологических данных с семантическими полями качественных признаков ‘острый’, ‘гладкий’, ‘толстый’ и ‘тонкий’, которые уже были проанализированы вручную участниками группы MLexT, см. [Кюсева 2012], [Кашкин 2013], [Козлов и др. 2016]). Мы показали, что задача заполнения анкеты состоит из двух этапов: перевода анкеты и ее заполнения. В рамках настоящего исследования мы опробовали два метода перевода анкет: с помощью онлайн-переводчиков компаний Yandex и Google и с помощью машиночитаемых переводных словарей Freedict и Verdict. С помощью первого метода можно переводить как отдельные слова, так и целые словосочетания, что теоретически могло бы позволить использовать его для решения обеих подзадач сразу (и перевода, и заполнения анкеты), однако такое решение приводит к потере большого количества признаковых слов, относящихся к данному полю в целевом языке. Пословный перевод прилагательных и существительных дает сопоставимые результаты для двух методов, поэтому на данном этапе мы останавливаемся на методе перевода с помощью машиночитаемых словарей, поскольку он кажется нам более прозрачным и надежным (по крайней мере, для тех языков, для которых есть словари приемлемого объема). Вторая подзадача решается на основе анализа сочетаемости полученных прилагательных и существительных по одноязычному корпусу серии WaC.

Результаты наших пилотных экспериментов показывают, что лексико-типологические анкеты могут заполняться автоматически. Так, например, точность и полнота заполнения анкеты для семантических полей ‘толстый/тонкий’ для итальянского языка составляют 0.79 и 0.93 соответственно.

Безусловно, наш алгоритм не предполагает полной автоматизации процесса лексико-типологического исследования, в отличие от методик, предложенных в работах [Youn et al. 2016] и [Wälchli and Cysouw 2013] и проанализированных во Введении. Однако, как нам кажется, на данном этапе это является его достоинством, а не недостатком. По-видимому, до тех пор, пока в лингвистике не будет накоплено достаточного количества надежных ресурсов (в частности, сопоставимых переводных словарей и объемных и сбалансированных параллельных корпусов), разработать методику полностью автоматического и при этом достаточно подробного типологического анализа произвольного класса лексики будет крайне затруднительно. Исходя из этого, мы проектируем алгоритм, который будет заменять работу лексического типолога не полностью, а только частично, позволяя, во-первых, расширить эмпирическую базу исследований путем учета максимального количества доступных ресурсов, и, во-вторых, упростить и ускорить процесс анализа электронных источников данных.

Таким образом, в соответствии с намеченными целями нашей работы, в дальнейшем мы предполагаем увеличить охват обрабатываемых словарей и корпусов, наладив в том числе процесс перевода и заполнения анкет материалами ряда других языков (в частности, немецкого, финского и японского), а также апробировать метод перевода анкет с помощью параллельных корпусов и разработать алгоритмы сбора типологических данных для глагольных и предметных семантических полей.

**Библиография**

Bell A. Language samples. Universals of Human Language, Method and Theory. Greenberg J.H., Ferguson C.A., Moravcsik E.A. (eds.). Stanford Univ Press, 2. Palo Alto, CA. – 1978 – Vol 1: 123–156.

Cox T.F., Cox M.A.A. Multidimensional Scaling. Chapman and Hall, 2001.

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

Reznikova T., Vyrenkova A. Semantics of falling: a cross -linguistic approach. NRU HSE. Series WP BRP “Linguistics”. – 2015 – №40.

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

Ryzhova D., Obiedkov S. Formal concept lattices as semantic maps // Proceedings of the 1st International Workshop on Computational Linguistics and Language Science (CLLS 2016), CEUR-WS.org, Eds. D. Ilvovsky, E. Chernyak, A. Vybornova, D. Skorinkin. 2017.

Wälchli, B., Cysouw, M. Lexical Typology through Similarity Semantics: Toward a Semantic Map of Motion Verbs. De Gruyter – 2012 – 50.3: 671–710.

Youn H., Suttond L., Smith E., Moore C., Wilkins J. F., Maddieson I., Croft W., and Bhattacharya T. On the universal structure of human lexical semantics. Proceedings of the National Academy of Sciences of the United States of America – 2016 – 113: 1766–1771.

Апресян Ю.Д. Лексическая семантика: синонимические средства языка. М.: Наука, 1974.

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

Козлов А.А., Кухто А.В., Привизенцева М.Ю. 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, 2013. С.: 522-533.

Кузьменко Е.А., Мустакимова Э.Г. Глаголы падения в лексико-типологической перспективе // Е.А. Лютикова, А.В. Циммерлинг, М.Б. Коношенко (ред.). – Типология морфосинтаксических параметров: материалы Междунар. конф. М.: МПГУ, 2015.

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

Кюсева М.В., Резникова Т.И., Рыжова Д.А. Типологическая база данных адъективной лексики// Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая – 2 июня 2013 г.) В 2-х томах /под ред. В.П.Селегей. Т.1: Основная программа конференции. Вып. 12 (19). М.: РГГУ, 2013: 407-419.

Орехов Б.В., Резникова Т.И. Компьютерные перспективы лексико-типологических исследований // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. – 2015 – №3. С.17-23.

Рахилина Е.В., Резникова Т.И. Фреймовый подход к лексической типологии. Вопросы языкознания. – 2013 – № 2: 3-31.

Рахилина Е.В., Резникова Т.И., Карпова О.С. Семантические переходы в атрибутивных конструкциях: метафора, метонимия и ребрендинг // Рахилина Е.В. (ред.). Лингвистика конструкций. М.: Азбуковник, 2010.

Рыжова Д.А. Построение лексико-типологической анкеты с помощью моделей дистрибутивной семантики // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. – 2015 – №3: 126-131.

Рыжова Дарья Александровна, преподаватель Школы Лингвистики Национального исследовательского университета “Высшая школа экономики”

E-mail: [daria.ryzhova@mail.ru](mailto:daria.ryzhova@mail.ru)

Daria Ryzhova

National Research University “Higher Sc”

Ершов Илья Андреевич, студент 4-ого курса бакалавриата Школы Лингвистики Национального исследовательского университета “Высшая школа экономики”

E-mail: cornerstone26@yandex.ru

Мельник Анастасия Александровна, студент 4-ого курса бакалавриата Школы Лингвистики Национального исследовательского университета “Высшая школа экономики”

E-mail: melnik-a-a@mail.ru

1. О возможных методах автоматизации других этапов см. Ryzhova, Obiedkov 2017, Орехов, Резникова 2015, Рыжова 2015. [↑](#footnote-ref-1)
2. говор села Теги [↑](#footnote-ref-2)
3. По данным словаря Multitran [URL: https://www.multitran.ru/] [↑](#footnote-ref-3)
4. Словари доступны по ссылке http://www.freedict.org/ru/. [↑](#footnote-ref-4)
5. Сервер со словарями от компании Yandex больше не существует, а сами словари переданы в школу лингвистики НИУ ВШЭ, и на их основе в настоящий момент создаётся новый сервер Вышка.Словари. [↑](#footnote-ref-5)
6. http://www.tei-c.org/index.xml [↑](#footnote-ref-6)
7. Безусловно в языках со свободным порядком слов необходимо учитывать оба варианта позиции прилагательного: до и после существительного. В то же время результаты, полученные для итальянского языка, считающегося языком с более или менее фиксированным порядком “существительное + прилагательное”, показывают незначительное влияние изменения параметра порядка слов в коллокации на итоговое качество работы алгоритма. [↑](#footnote-ref-7)
8. Порог является настраиваемым параметром. [↑](#footnote-ref-8)