Мультиязычные параллельные корпуса: новый источник данных для типологических исследований, перспективы использования и проблемы

© 2019 **Любовь Владимировна Нестеренко**

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия; lyu.klimenchenko@gmail.com

Аннотация: В работе рассматриваются мультиязычные параллельные корпуса с точки зрения возможностей использования их в качестве источника данных для типологических исследований. Мультиязычные параллельные корпуса открывают возможности для проведения квантитативных исследований в типологии. Однако на сегодняшний день они пока не получили широкого распространения в области типологии. Это связано с отсутствием корпусов, соответствующих требованиям, необходимым для проведения типологических исследований, а также отсутствием выработанной единой схемы построения мультиязычных параллельных корпусов. В статье обсуждаются факторы, которые препятствуют использованию мультиязычных параллельных корпусов типологами и высказываются идеи о том, какие требования следовало бы учитывать при разработке мультиязычных параллельных корпусов для типологических исследований.

Ключевые слова: корпусная лингвистика, обзоры, параллельные корпуса

Для цитирования: Нестеренко Л. В. Мультиязычные параллельные корпуса: новый источник данных для типологических исследований, перспективы использования и проблемы. *Вопросы языкознания*, 2019, 2: 111–125.

DOI: 10.31857/S0373658X0004308-7.

Благодарности: Автор выражает глубочайшую благодарность А. А. Бонч-Осмоловской за ценные советы, полученные в ходе работы над текстом, а также рецензентам за их подробные комментарии и справедливые замечания к статье.

Multilingual parallel corpora: Alternative source of language data for typological studies, applying perspectives and problems

Lyubov V. Nesterenko

National Research University Higher School of Economics, Moscow, Russia; lyu.klimenchenko@gmail.com

Abstract: In this paper, we discuss the perspectives of using multilingual parallel corpora as a source of language data for cross-linguistic studies. Multilingual parallel corpora make it possible to apply quantitative methods to cross-linguistic data. However, they have not become popular among researchers yet. The reason for that is the lack of multilingual parallel corpora that are suitable for linguistic studies and also the absence of unified guidelines for multilingual parallel corpora development. In the paper, we will analyse the factors that make it difficult to use multilingual parallel corpora for linguistic experiments and present some ideas about the features one should take into account when building multilingual parallel corpora for typological studies.

Keywords: corpus linguistics, parallel corpora, surveys

For citation: Nesterenko L. V. Multilingual parallel corpora: Alternative source of language data for typological studies, applying perspectives and problems. *Voprosy Jazykoznanija*, 2019, 2: 111–125.
DOI: 10.31857/S0373658X0004308-7.

1. Введение

В настоящей статье представлен обзор, посвященный оценке мультиязычных параллельных корпусов с точки зрения возможностей их использования в качестве исследовательского инструмента для лингвистов и степени их вовлеченности в современные типологические исследования.

Параллельные корпуса представляют собой массивы параллельных текстов, или, иными словами, коллекцию оригинальных текстов на языке L_1 с их переводами на один или более языков $L_2 \ldots L_n$. Особенность параллельного корпуса состоит в том, что тексты, включаемые в него, являются выровненными. Это означает, что текстовой единице из одного языка соответствует эквивалентный перевод на другом языке или же несколько переводов на разных языках. В первом случае мы имеем дело с двуязычным корпусом, во втором — с мультиязычным. Двуязычные параллельные корпуса оказываются наиболее распространенными, мультиязычных значительно меньше, поскольку они требуют наличия текстов, для которых имеются переводы сразу на несколько языков, а также инструментов для их обработки. Обычно тексты в параллельном корпусе выровнены по предложениям (реже по фрагментам текста, абзацам), а иногда в корпусе может быть также и пословное выравнивание. Помимо выравнивания тексты могут быть снабжены морфологической и синтаксической разметкой, однако выравнивание в параллельном корпусе является ключевой составляющей.

В ходе развития корпусной лингвистики менялись методы исследования языкового материала, полученного из корпуса. Лингвисты продвинулись от ручного анализа языковых примеров к применению квантитативных методов (в том числе машинного обучения и кластерного анализа) для исследования корпусных данных. Квантитативные исследования получили широкое распространение, существует множество работ, выполненных в данном направлении, на материале одного языка или нескольких при наличии сопоставимых корпусов.

Мультиязычные параллельные корпуса открывают возможности для проведения квантитативных исследований в типологии. Это означает переход от работы с отдельными примерами к работе со структурированными данными — массивами унифицированных контекстов с разметкой. Выравнивание в корпусе дает возможность использовать мультиязычные параллельные тексты для межъязыкового сравнения. Поскольку в таком случае мы имеем дело с переводами одного текста, это позволяет рассматривать то или иное языковое явление в разных языках в одинаковых контекстах — иными словами, использовать для сравнения языковой материал с одинаковой семантической и прагматической составляющей. Морфологическая и синтаксическая разметка в корпусе наделяет контексты характеристиками по различным параметрам. Таким образом, наличие выравнивания и разметки позволяет автоматически извлекать языковые единицы с релевантными для исследования параметрами и, что важно, применять к полученным данным квантитативные методы, позволяющие оценивать достоверность выдвигаемых гипотез или, наоборот, выдвигать новые.

В результате многочисленных разработок в области машинного перевода с конца 90-х годов появилось большое количество параллельных корпусов. Рост числа параллельных корпусов, а главное, появление мультиязычных параллельных корпусов стало стимулом к написанию работ, в которых лингвисты указывали на то, что мультиязычные параллельные корпуса могут послужить материалом для типологических исследований [Dahl 2007; Cysouw, Wälchli 2007]. В статье [Wälchli 2007] Бернард Вельхли сетовал на то, что при

наличии немалого количества переводов текстов, типологических исследований, проведенных на их материале, совсем немного. С этим утверждением нельзя поспорить, к тому же, имеется ряд факторов, затрудняющих использование мультиязычных параллельных корпусов в типологии.

Если говорить о языковом наполнении, а не о функциональных возможностях современных мультиязычных параллельных корпусов, то они, по-видимому, больше подходят для контрастивных исследований, нежели для типологических исследований в их традиционном понимании. Число мультиязычных параллельных корпусов, содержащих тексты как минимум на паре десятков языков (без учета их генетического разнообразия), довольно мало, а мультиязычных параллельных корпусов, которые могли бы стать полноценным материалом для типологических исследований, практически нет. Исключением являются переводы Библии [Mayer, Cysouw 2014], которые представлены на сотнях языков, из этого набора определенно можно сформировать типологическую выборку. Однако в текстах Библии довольно часто встречается калькирование, соответственно, используется много малоупотребительных конструкций. Это обстоятельство будет искажать лингвистические данные, например при попытке количественного анализа.

Более того, на сегодняшний день, по-видимому, не выработано критериев и единой схемы создания мультиязычных параллельных корпусов для типологических исследований. Если обратиться к современным мультиязычным параллельным корпусам, возникает ряд методологических проблем, связанных с различными аспектами их разработки. Некоторые корпуса имеют «кусочное» наполнение. В них включены различные тексты, которые представлены не на всех языках, а являются параллельными лишь для части языков. На деле получается, что корпус состоит из нескольких параллельных корпусов и текстовое наполнение для языковых пар разное. Например, если взять ресурс OPUS [Tiedemann 2009] — объемную коллекцию разнообразных параллельных текстов, то исследователю потребуется опираться на информацию о количестве и содержательной подборке текстов для каждой языковой пары отдельно, чтобы составить полноценный мультиязычный параллельный корпус. Несмотря на то, что в OPUS представлены тексты на более чем 90 языках, не все имеющиеся тексты являются переводными эквивалентами друг друга. В качестве противоположного примера можно привести корпус Европейского парламента [Koehn 2005], где все представленные тексты имеют переводные эквиваленты.

Другая проблема связана со снабжением текстов морфологической и синтаксической разметкой. Нередко создатели корпуса ограничиваются лишь выравниванием текстов, в некоторых случаях наличие разметки предусмотрено только для части языков. Кроме того, есть корпуса, в которых имеющаяся разметка оказывается неоднородной, т. е. схожая грамматическая информация, например частеречная принадлежность, кодируется поразному в разных языках. Это усложняет процесс извлечения языковых данных из корпуса и вынуждает исследователей подстраивать процесс первичной обработки материала под варианты разметки. Следовательно, важным является наличие в корпусе не только выравнивания, но и разметки в унифицированном формате. Выработка единой схемы, учитывающей все важные этапы создания мультиязычных параллельных корпусов, которой будут придерживаться разработчики, позволит сделать эти корпуса более доступными для применения в типологии.

Итак, следует признать, что пока мультиязычные параллельные корпуса не получили широкого распространения в типологии. В этой статье мы попробуем подробно разобраться, почему ситуация сложилась таким образом и что может способствовать ее улучшению. Для этого мы сначала рассмотрим возможности мультиязычных параллельных корпусов и различные области их использования: от прикладных, таких как машинный перевод или снятие семантической неоднозначности, до сугубо теоретических — использование мультиязычных параллельных корпусов для сравнительных лексико-грамматических (типологических) исследований разноструктурных языков. Затем будет представлен краткий обзор существующих мультиязычных параллельных корпусов и их характеристик. После этого

мы перечислим проблемы, связанные с использованием существующих параллельных корпусов в типологии, а в завершение предложим свое видение того, как должны быть устроены параллельные корпуса, предназначенные для типологических исследований.

2. Возможности мультиязычных параллельных корпусов

В этом разделе речь пойдет о прикладных задачах и направлениях лингвистических исследований, в которых используются мультиязычные параллельные корпуса. К первым относятся прежде всего машинный перевод и автоматическое составление переводных словарей, ко вторым — исследования лексики, в том числе лексическая типология, и филогенетика, а также контрастивные и типологические исследования грамматики.

Параллельные корпуса уже давно активно используются в качестве источника материала в области автоматической обработки естественного языка. Одной из таких задач является машинный перевод, см., например, работы [Brown et al. 1993; Koehn et al. 2003; Callison-Burch et al. 2004; Koehn 2005]. Добавление в мультиязычную систему статистического машинного перевода очередной языковой пары требует наличия корпуса параллельных текстов, выровненных пословно и/или по предложениям [Brown et al. 1993]. Параллельный корпус в данном случае необходим для построения вероятностной переводной таблицы (probabilistic translation table), которая используется в системе перевода для нахождения соответствий между языками [Koehn et al. 2003; Callison-Burch et al. 2004]. Некоторые существующие параллельные корпуса, например мультиязычный корпус Europarl [Koehn 2005], специально создавались под данную задачу. Мультиязычный корпус позволяет построить сразу несколько языковых моделей для перевода, которые основаны на текстах одного жанра и одной тематики, что является преимуществом перед использованием множества двуязычных корпусов, которые могут быть разнородными по составу.

Выровненные по предложениям мультиязычные параллельные тексты также оказываются подходящим материалом для автоматического составления разнообразных переводных (двуязычных) словарей и конкордансов, см. [Gale, Church 1991]. Такие задачи перекликаются с задачей пословного выравнивания. Для установления соответствий между словами в двух языках программа принимает на вход пары предварительно выровненных предложений, после установления соответствий определяется уровень достоверности для каждой пары слов. Подобные словари, составленные на основе мультиязычных параллельных текстов и представляющие собой списки вероятностно взвешенных соответствий, в свою очередь используются в мультиязычных приложениях обработки естественного языка [Sahlgren, Karlgren 2005]. В частности, довольно активно они применялись для перевода поисковых запросов в мультиязычных системах извлечения информации [Davis, Dunning 1995; Nie et al. 1999; Chen, Nie 2000]. Выровненный мультиязычный параллельный корпус может также быть использован для разработки и тестирования программ снятия семантической омонимии в разных языках. Для оценки работы алгоритма, цель которого, например, различать употребления слова *bank* в английском (со значением ‘банк’ и со значением ‘скамейка’), можно использовать соответствующий французский текст, где данные значения будут выражены словами *banc* и *banque* (пример взят из работы [Gale, Church 1991]).

Помимо использования мультиязычных параллельных корпусов в качестве лексического ресурса для автоматического составления переводных словарей, на материале мультиязычных параллельных корпусов можно также проводить и исследования лексики. Примерами являются, например, работы [Sharoff 2002; Добровольский 2009; Сичинава 2015]. В статье [Sharoff 2002] автор пошагово описывает процесс создания лексикографической базы данных для русского, английского и немецкого на примере прилагательных размера. Затем на основе анализа прилагательных предлагается схематическое разделение типов контекстов, в которых употребляются прилагательные размера. В [Добровольский 2009] указывается,

что особенно удачно параллельные корпуса подходят для изучения лингвоспецифической лексики. В работе также приводится сопоставительное исследование русских обращений и их реализаций в английском и немецком на материале параллельных корпусов Национального корпуса русского языка [НКРЯ] и корпуса Австрийской академии наук в Вене. Другой пример — исследование, представленное в статье [Сичинава 2015], где на основе сравнения статистических мер автор проверяет предположение о том, что «лингвоспецифическая лексика скорее появится при создании оригинального текста, чем при переводе».

К исследованиям лексики можно отнести также и работы по лексической типологии. Таких работ, выполненных с использованием мультиязычных параллельных корпусов, пока совсем немного: статьи [Wälchli, Cysouw 2012] и [Östling 2016], где в качестве мультиязычного параллельного корпуса используются выровненные тексты Библии, и статья [Бонч-Осмоловская, Нестеренко 2018] на материале выровненных книг о Гарри Поттере. В [Wälchli, Cysouw 2012] корпус Библии используется для построения вероятностной семантической карты глаголов движения в сотне языков. Карта строится на основе матрицы семантического сходства с использованием метода многомерного шкалирования. Такой подход позволяет более дробно классифицировать анализируемые языковые единицы. Работа [Östling 2016] посвящена исследованию колексификации (выражению различных концептов одной лексемой) на материале текстов Библии в 1001 языке. Списки слов, извлеченные автоматически с использованием пословного выравнивания и переноса морфологической разметки английского и шведского на другие языки, хоть и содержат ошибки, но тем не менее позволяют обнаруживать имеющиеся в языках схемы колексификации. В статье [Бонч-Осмоловская, Нестеренко 2018] исследуются глаголы семантического поля «искать — находить» в восьми индоевропейских языках на материале переводов книг о Гарри Поттере. Авторы используют метод сетевого анализа в качестве инструмента для построения семантической карты. В получившемся графе вершинами являются переводные эквиваленты глаголов поля «искать — находить», а соединяющие их ребра — указание на то, что лексемы являются переводами друг друга, — имеют вес, количественно характеризующий эту связь. Такой граф дает лингвисту возможность предварительно оценить устройство семантического поля и учесть полученный результат в планируемом исследовании.

Помимо того, что пословное выравнивание позволяет сравнивать переводные эквиваленты лексем, оно может послужить основой для определения филогенетического родства языков. Так, в исследовании [Mayer, Cysouw 2012] на материале религиозных брошюр с сайта Watchtower проводится сравнение 146 языков на предмет филогенетического родства путем оценивания их сходства на основе результатов мультиязычного пословного выравнивания. Важным показателем среди языков является частота совпадений слов в рамках выравниваний. Слова на разных языках, которые встречались в одних цепочках выравнивания, группируются в кластеры (пример кластера: *Исус* (болгарский), *Jesus* (английский), *Fia* (эве), *Yesu* (эве), *Ġesù* (мальтийский), *Jesús* (испанский), *Jesus* (немецкий)). Сходство языков определяется следующим образом. Те языки, слова из которых чаще попадают в один кластер, иначе говоря, между которыми часто устанавливается соответствие при выравнивании, считаются более похожими. Авторы предлагают свой подход в том числе в качестве альтернативы спискам Сводеша.

Мультиязычные параллельные корпуса также являются ценным материалом для лингвистов, занимающихся сравнительными и контрастивными исследованиями грамматики, и для типологов. Поскольку параллельные тексты представляют собой переводы одного текста, то их преимущество для подобных исследований состоит в том, что в них можно найти примеры предложений с одинаковой семантической и прагматической составляющей. Такие предложения можно считать своего рода минимальными парами. Если исследователь хочет сопоставить некую конструкцию из языка *L* с ее реализациями в других языках, ему необходимо полагаться на примеры с одинаковым значением и контекстным окружением. Объем данных, содержащийся в параллельных корпусах, позволяет проследить внутриязыковую вариативность [Cysouw, Wälchli 2007], а также открывает возможности

для применения в типологии квантитативных методов, в том числе кластерного анализа, сетевого анализа и машинного обучения, являющихся в области лингвистики новыми методами исследования. Использование примеров, извлеченных из параллельных текстов, позволяет задействовать в анализе информацию о контексте и расширить количество признаков, участвующих в построении статистической модели.

В качестве примера контрастивного грамматического исследования мы рассмотрим работу [Stambolieva 2011], а затем обратимся к квантитативным типологическим исследованиям [Cysouw 2014; Östling 2015; Asgari, Schütze 2017]. Работа [Stambolieva 2011], выполненная на материале двуязычного англо-болгарского корпуса, является примером того, как можно использовать переводные эквиваленты для сравнения языкового явления из одного языка с его реализациями в другом. В статье исследуются переводные эквиваленты предложений с типами ситуаций, требующими употребления различных аспектуальных форм глагола. В болгарском языке категория глагольного вида устроена иначе, чем в английском, что нередко вызывает трудности при переводе. В ходе количественного анализа автор составляет список языковых средств, используемых для выражения перфектива и имперфектива в английских переводах.

Теперь рассмотрим типологические работы, выполненные с применением квантитативных методов. Отметим, что в этих статьях, как и в работах по лексической типологии, обсуждаемых выше, в качестве материала используются два источника: Библия и религиозные брошюры. Статья [Cysouw 2014] посвящена исследованию маркирования семантических ролей в 15 языках, языковым материалом служат тексты религиозных брошюр с сайта Watchtower. Автор рассматривает падежное маркирование слова ‘Библия’ в 34 различных контекстах. Контексты отбирались вручную, немало контекстов не попали в выборку, так как в них на многих языках отсутствовало прямое упоминание Библии. Затем для каждого языка были объединены контексты, в которых слово ‘Библия’ имело одинаковое маркирование. В результате была построена сеть NeighborNet для 15 рассматриваемых языков, которая основывается на сходстве моделей падежного маркирования.

В [Östling 2015] представлено исследование порядка слов, выполненное также на материале текстов Библии для 986 языков. В работе предлагается алгоритм, позволяющий автоматически выровнять тексты пословно, а затем, используя разметку одного языка, спроецировать ее на остальные языки. Несмотря на привлекательность решения, при таком подходе сильно страдает полнота и значительная часть материала отсекается. На основании информации, полученной об устройстве порядка слов в языках, была проведена иерархическая кластеризация, а результат был автоматически оценен с использованием WALS в качестве золотого стандарта. Полученные результаты согласуются с данными WALS на 86–96 %.

Статья [Asgari, Schütze 2017] посвящена исследованию грамматической категории времени на обширном языковом материале — переводах Библии на 1000 языков. Авторы статьи предлагают новый подход к обработке данных мультиязычных параллельных корпусов, позволяющий получить информацию об устройстве языкового явления даже в тех языках, для которых лингвисты не располагают большим количеством данных. Описанное авторами решение позволяет разметить в корпусе языковое явление, проецируя информацию о нем из нескольких заданных языков на большее число языков, а затем на все языки выборки. Подход основывается на двух гипотезах: 1) Для языкового явления *f*, которое часто кодируется в языках мира, существует несколько языков, где оно явно маркировано на поверхностном уровне; 2) Для языка *l*, в котором кодируется явление *f*, проекция *f* из языков с явным маркированием в массово параллельном корпусе позволяет обнаружить способ кодирования *f* в *l*, как в случае наличия в *l* явного маркирования *f*, так и в случае неявного маркирования. Авторы работы извлекают из корпуса способы кодирования времени для 1000 языков, а затем, основываясь на полученных данных, проводят иерархическую кластеризацию временны́х показателей в разных языках.

Итак, мы рассмотрели ряд работ, выполненных на материале мультиязычных параллельных корпусов; в этих исследованиях предлагаются новые подходы к анализу

различных языковых явлений. При этом наиболее популярным источником параллельных текстов, по-видимому, является Библия ввиду доступности текстов, отсутствия у них авторских прав и наличия переводов на редкие, малораспространенные языки. Чтобы понять, что тормозит использование мультиязычных корпусов в грамматических и типологических исследованиях, обратимся прежде всего к существующим на сегодняшний день ресурсам и проанализируем их возможности для применения в лингвистических исследованиях.

3. Современные мультиязычные параллельные корпуса

На данный момент существует немало мультиязычных параллельных корпусов, каждый из которых обладает своими отличительными характеристиками, связанными с его языковым и текстовым наполнением, а также общим устройством. В этом разделе мы сначала рассмотрим мультиязычные параллельные корпуса, содержащие специфические тексты, такие как Библия, тексты докладов с заседаний Европарламента и субтитры, затем речь пойдет о параллельных корпусах в составе национальных корпусов (НКРЯ и Чешский национальный корпус), и в завершение обратимся к параллельным корпусам с синтаксической разметкой.

Для типологических исследований представляют интерес прежде всего параллельные корпуса, основой которых являются мультиязычные тексты. Самый известный пример таких текстов — это переводы Библии: она хорошо изучена и была переведена на большое количество современных языков, также имеются ее версии и на многих древних языках. Подобные тексты называют массово параллельными (massively parallel texts). Помимо Библии к массово параллельным текстам относят статьи из Википедии, книги о Гарри Поттере, повесть «Маленький принц», доклады с заседаний Европарламента и др. Большинство этих текстов находится в открытом доступе и используется исследователями. Для Библии существует корпус, созданный учеными из Университета Марбурга. Параллельный корпус включает в себя 1169 переводов с выравниванием по предложениям, тексты доступны для скачивания в текстовом формате [Mayer, Cysouw 2014].

Другим объемным мультиязычным параллельным корпусом, созданным на основе массово параллельных текстов, является корпус Европейского парламента (Corpus of the European Parliament, DCEP), он содержит 1,5 миллиона документов (общим объемом в 1,37 миллиарда токенов), опубликованных на официальном сайте Европейского парламента. Тексты доступны пользователям как в исходном виде, так и с выравниванием по предложениям.

Обширная коллекция текстов с переводными эквивалентами содержится на ресурсе OPUS Open Source Parallel Corpus [Tiedemann 2009]. Среди текстов имеются, например, субтитры с ресурса opensubtitles.org, тексты докладов с заседаний Европейского парламента, фрагменты Корана, файлы локализации операционной системы Linux. Тексты, разумеется, неоднородны в отношении языкового разнообразия: доклады Европарламента представлены на заведомо фиксированном количестве языков, а, например, коллекция субтитров продолжает пополняться новыми языками. Тексты предобработаны, автоматически выровнены, а некоторые имеют морфологическую разметку. Обработка текстов производилась исключительно автоматически, ручные правки не вносились. В OPUS можно найти тексты на более чем 90 языках, общий объем корпуса составляет 40 миллиардов токенов и, соответственно, 2,7 миллиарда параллельных единиц (предложений, текстовых фрагментов). Однако объем коллекций по языкам сильно варьирует: так, для 100 наиболее представленных языковых пар объем находится в диапазоне от 100 до 500 миллионов токенов.

Теперь рассмотрим, как устроены параллельные корпуса, входящие в состав НКРЯ и Чешского национального корпуса. В НКРЯ имеется раздел, в котором представлены 15 двуязычных параллельных корпусов, а также мультиязычный параллельный корпус на 12 языках, преимущественно славянских: русском, украинском, польском, чешском,

словацком, словенском, хорватском, сербском, македонском, болгарском, а также английском и нидерландском. По данным [Сичинава 2015], в состав мультиязычного корпуса входит девять текстов: «Алиса в стране чудес» и «Алиса в Зазеркалье» Л. Кэррола, «Собака Баскервилей» А. Конан Дойля, «Винни-Пух» А. А. Милна, «Код да Винчи» Д. Брауна, «Маленький принц» А. де Сент-Экзюпери, «Пиноккио» К. Коллоди, «Алхимик» П. Коэльо и «Мастер и Маргарита» М. А. Булгакова. Все тексты выровнены и снабжены морфологической разметкой. Согласно работе [Sitchinava 2012], мультиязычные корпуса НКРЯ создавались при сотрудничестве с проектами ParaSOL под руководством Р. фон Вальденфельса [Waldenfels 2006], ASPAC под руководством А. Барентсена (https://spraakbanken.gu.se/eng/resources/corpus) и InterCorp, о котором мы расскажем далее.

В Чешском национальном корпусе также есть раздел с мультиязычным параллельным корпусом, созданный в рамках проекта InterCorp [Vavřín, Rosen 2008; Čermák, Rosen 2012]. В последней версии корпуса, опубликованной в 2016 г., представлены 39 языков, преимущественно европейских, общий объем словоупотреблений около 5,5 миллионов, для 23 языков имеется морфологическая разметка, для 20 доступна лемматизация. Объем представленных текстов различается от языка к языку, при этом нет такого фрагмента материала, который бы охватывал все языки. Наиболее «параллельными» составляющими корпуса являются три книги Дж. К. Роулинг о Гарри Поттере, первая часть «Властелина колец» Дж. Р. Р. Толкина, романы М. Кундеры «Невыносимая легкость бытия» и «Бессмертие» и «Похождения бравого солдата Швейка» Я. Гашека. Пользователь может осуществлять поиск по корпусу через веб-интерфейс, подкорпус задается выбранным множеством языков.

Существуют также мультиязычные параллельные корпуса с синтаксической разметкой: англо-шведско-турецкий и англо-итальяно-французский. Англо-шведско-турецкий корпус [Megyesi et al. 2010] содержит 300 000 токенов на шведском, 160 000 на турецком и 150 000 на английском, на интернет-странице проекта доступна демо-версия, основная часть корпуса находится в закрытом доступе. Англо-итальяно-французский корпус ParTUT[1] содержит в общей сложности 167 000 токенов, в среднем по 2100 предложений на язык, и доступен для скачивания.

На сегодняшний день исследователи располагают небольшим количеством ресурсов, содержащих мультиязычные параллельные тексты. Разнообразие языков в выборке, наличие синтаксической разметки, наличие текстов различных жанров, удобный формат представления данных — все это важные особенности мультиязычных параллельных корпусов, хотя отметим, что эти характеристики присущи не каждому из корпусов, упомянутых выше. Однако, несмотря на наличие преимуществ, имеется ряд факторов, которые ставят под вопрос возможность их использования в типологических исследованиях, — эти факторы мы обсудим ниже.

4. Проблемы использования мультиязычных параллельных корпусов в типологических исследованиях

Несмотря на разнообразие современных ресурсов, многие из существующих мультиязычных параллельных корпусов сложно использовать в качестве материала для лингвистических исследований ввиду ряда недостатков. К ним относится отсутствие разнообразия в языковой выборке, неоднородность переводных пар, отсутствие разметки, неоднородность имеющейся разметки, отсутствие информации о точности разметки, калькирование и вольный перевод. Рассмотрим перечисленные недостатки современных мультиязычных параллельных корпусов более подробно.

[1] Все части корпуса доступны в репозитории https://github.com/UniversalDependencies. Имеется разметка по стандарту Universal Dependencies.

1. Отсутствие языкового разнообразия.

При проведении типологических исследований важную роль играет языковая выборка, поэтому прежде всего следует сказать о проблеме, связанной с генетическим разнообразием языков, представленных в корпусе. Существующие ресурсы содержат тексты преимущественно на индоевропейских языках, из других языковых семей часто бывают представлены финский, эстонский, китайский. Лингвисты получают довольно скудный языковой набор и оказываются ограниченными в выборе языков для исследования.

2. Неоднородность переводных пар.

Иногда в корпусах, например в упомянутых выше OPUS и мультиязычном параллельном корпусе Чешского национального корпуса InterCorp, текстовое наполнение для разных языковых пар совпадает не полностью. Некоторые тексты могут быть представлены одновременно на языках *L1*, *L2* и *L3* (т. е. переводные эквиваленты на трех языках), а еще какое-то количество текстов может быть только для пары *L1* — *L2*, и другая часть для пары *L2* — *L3*. То есть не все имеющиеся в корпусе тексты являются переводными эквивалентами друг друга, как, например, в случае Библии или собрания книг о Гарри Поттере на разных языках.

3. Отсутствие разметки.

Другой проблемой является отсутствие в корпусе какой-либо разметки: например, в OPUS, где содержатся тексты для множества языковых пар, разметка имеется лишь частично. В отличие от обучения статистических моделей для машинного перевода, где разметка в корпусе не требуется, для типологических исследований необходимо наличие в мультиязычном параллельном корпусе морфологической и/или синтаксической разметки. Сложность состоит в том, что если выравнивание мультиязычных параллельных текстов можно осуществить при помощи алгоритмов, не зависящих от языка, то для морфологической и синтаксической разметки нужны инструменты, ориентированные на конкретные языки.

4. Различия в разметке.

Наличие в корпусе разметки, несомненно, является преимуществом, однако и здесь имеются спорные моменты. При создании мультиязычных параллельных корпусов могут быть использованы разные морфологические и синтаксические анализаторы, в которых системы тэгов различаются. Пример такой ситуации — корпус InterCorp. В описании ресурса перечислены различные морфологические анализаторы, использованные при его разработке[2]. Также для каждого анализатора указана ссылка на список тэгов, которые в нем используются, однако наборы тэгов не совпадают. Сравним тэги для обозначения рода существительных в русском и исландском.

Таблица 1

Тэги для обозначения рода в русском и исландском

Род	Тэг в русском	Тэг в исландском
мужской	m	k
женский	f	v
средний	n	h
не определен	—	x

Из таблицы видно, что в русском и исландском тэги для обозначения рода существительных абсолютно разные. Использование разных тэгов в рамках одного корпуса затрудняет не только ручной поиск по корпусу, но и автоматическую обработку и извлечение информации из корпусных данных, ведь исследователю необходимо учитывать, что в некоторых языках разметка различается.

[2] http://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze9#fnt__2

5. Отсутствие информации о качестве разметки.

При наличии в корпусе автоматической разметки всегда встает вопрос о ее качестве, в особенности, если заходит речь об исследовании некоторого языкового явления на материале этого корпуса. Извлекая из корпуса языковые данные, обладающие релевантными параметрами для построения математических моделей, мы опираемся на разметку. Параметры для модели, получаемые через разметку, непосредственно влияют на выводы, которые мы делаем, и оказывается невозможным оценить, насколько точными являются наши результаты. В работе [Östling 2015] используется процедура трансфера разметки, при которой разметка, сделанная для одного языка, переносится на данные других языков через пословное выравнивание. Однако такой подход дает много погрешностей (низкая полнота); учитывая это, автор тщательно фильтрует полученный результат аннотирования, чтобы минимизировать попадание некорректно размеченного материала в рабочую версию корпуса. Таким образом, важно иметь представление о качестве разметки в корпусе, в частности о качестве разметки исследуемого явления. Ошибки и неточности, так или иначе содержащиеся в автоматической разметке, могут серьезно повлиять на результаты исследования. Кроме того, при работе с мультиязычными параллельными корпусами следует учитывать, что уровень качества разметки может различаться у разных языков. Для того, чтобы убедиться в этом, рассмотрим в качестве примера корпус ParTUT и проведем в нем оценку качества разметки одного грамматического явления.

Корпус ParTUT содержит тексты на английском, итальянском и французском[3], данных о точности разметки в корпусе нет, хотя разметка производилась автоматически. Для оценивания мы выбрали эксплетивы, грамматическое явление, присутствующее как в английском, так и в итальянском. Эксплетивами называют местоимения, которые имеют в предложении синтаксическую роль, но не вносят никакого конкретного значения:

(1) Английский
Everywhere one looks, it seems, ***there are*** *deep-seated problems.*

Итальянский
Ovunque si guardi, sembrano ***esserci*** *problemi profondi.*

При оценивании качества разметки эксплетивов из корпуса был отобран массив примеров с данными конструкциями. Сперва были извлечены примеры, где имелся тэг маркирования связи эксплетива и глагола «expl», а затем при помощи поиска по местоимениям *there*, *ci* и др. были добавлены примеры, являющиеся потенциальными эксплетивами и не помеченные при разметке как таковые. В полученном материале нами было подсчитано число ложно-положительных, ложно-отрицательных и верно размеченных примеров и вычислена точность и полнота. Оценка проводилась на материале корпуса ParTUT, скачанного в текстовом формате, работа с примерами и тэгами разметки производилась при помощи написанных нами скриптов на Python, определение ошибок и верно размеченных эксплетивов сделано вручную. Результаты нашей оценки представлены в таблице 2.

Таблица 2

Качество разметки эксплетивов в корпусе ParTUT

Язык	Точность	Полнота	F_1-мера
английский	0,83	0,86	0,84
итальянский	0,99	0,94	0,96

Из полученных результатов видно, что в качестве разметки эксплетивов для английского и итальянского в корпусе ParTUT имеется расхождение. Это следует учесть при

[3] Французский язык был исключен из сравнения, поскольку для него в корпусе содержится лишь около 500 предложений, что непропорционально меньше, чем для английского и итальянского (по 2100 предложений на язык).

работе с материалом корпуса. Следует также иметь в виду, что при работе с другими корпусами и грамматическими явлениями результаты разметки в разных языках могут различаться сильнее.

Информация о качестве разметки является важным атрибутом мультиязычного параллельного корпуса для типологических исследований. Предоставить информацию о точности разметки множества языковых явлений в корпусе довольно затруднительно, однако средняя точность качества морфологической (отдельно можно указать точность лемматизации и частеречной разметки) и синтаксической разметки все же необходима. Информация, указывающая на расхождения в средней точности для разных языков (даже если низшее значение точности является удовлетворительным), как минимум, послужит исследователю поводом для поиска причин такого расхождения, а также для проверки качества разметки исследуемого явления в соответствующем языке.

6. Наличие калькирования и вольного перевода.

Если говорить о проблемах, связанных с качеством текстов в мультиязычных параллельных корпусах, то следует упомянуть, что эквивалентные переводы параллельных текстов могут содержать кальки, а также фрагменты с вольным переводом [Сичинава 2015]. Эти факторы будут искажать результаты лингвистических исследований, а наличие вольных переводов может также повлиять на результаты автоматического выравнивания.

Возможности современных мультиязычных параллельных корпусов не делают их доступными для использования в типологии. Это связано, по-видимому, с тем, что их разработчики не ориентировались на создание корпуса для типологических исследований. Некоторые, например, корпус Европейского парламента, изначально создавались совсем под другую задачу — машинный перевод. Разработка корпусов, обладающих соответствующими характеристиками, позволит решить проблему использования мультиязычных параллельных корпусов в типологических исследованиях. О том, каким нам представляется устройство мультиязычного параллельного корпуса для типологических исследований, будет сказано в следующем разделе.

5. Разработка мультиязычного параллельного корпуса для типологических исследований

Попробуем определить, какими характеристиками должен обладать мультиязычный параллельный корпус для типологических исследований. Типология в традиционном понимании требует наличия разнообразия в языковой выборке. Мультиязычный параллельный корпус должен по возможности включать в себя языки разных семей и разного грамматического строя, а не быть ограничен европейскими языками, если, конечно, речь не идет о создании корпуса языков какой-то конкретной группы, например параллельного корпуса славянских языков [Waldenfels 2006].

Что касается технических требований к мультиязычным параллельным корпусам, то, с одной стороны, они не сильно отличаются от требований к обычным корпусам, за исключением необходимости иметь возможность выбрать некоторое подмножество языков. С другой стороны, в последнее время исследователи все больше заинтересованы в применении квантитативных методов к языковым данным, а значит, привычный корпус с пользовательским интерфейсом становится уже не таким удобным ресурсом. Важным аспектом становится наличие возможности скачать материал корпуса в текстовом формате, это также позволит исследователям использовать свои программные инструменты для обработки языковых данных. Необходимость скачивания материала корпуса также связана с тем, что не представляется возможным обеспечить проведение квантитативных исследований с применением методов кластерного анализа и/или машинного обучения в рамках того же компьютерного ресурса, где располагается корпус с поисковым интерфейсом

(то есть добавить соответствующую опцию в интерфейсе). Такая возможность потребует от ресурса больших вычислительных мощностей. Тексты для скачивания могут быть представлены как в обработанном виде с разметкой и выравниванием, так и в виде выровненного текста без разметки, а также в совершенно необработанном виде, чтобы при необходимости можно было обработать корпус другим анализатором.

Особенно важным при разработке корпусов является стандарт разметки. Наборы языковых признаков и обозначающих их тэгов, используемые в корпусах и различных парсерах, нередко сильно различаются. Если в мультиязычном параллельном корпусе для каждого языка используется свой стандарт разметки, то могут возникать трудности при автоматической обработке такого корпуса и поиске по нему. Это можно исправить, сделав конвертацию тэгов в единый формат. Данная манипуляция носит чисто технический характер и необходима для обеспечения удобства работы с материалом. Однако при попытке конвертации может оказаться, что не все тэги можно конвертировать друг в друга однозначно. Например, в одном стандарте разметки деепричастие маркируется как деепричастие, а в другом как глагол. В таком случае невозможно будет конвертировать второй стандарт в систему тэгов, где деепричастие маркируется как деепричастие.

Следовательно, при создании мультиязычных параллельных корпусов и парсеров важную роль играет наличие единого стандарта разметки. С похожей проблемой типологи сталкивались ранее, когда отсутствовали унифицированные требования к оформлению примеров с глоссами. В связи с этим были разработаны Лейпцигские правила глоссирования [Bickel et al. 2008], которые на данный момент считаются общепризнанным стандартом.

Единый стандарт разметки для мультиязычных параллельных корпусов прежде всего упростит процесс автоматической обработки языковых данных, а именно извлечение морфологических и синтаксических признаков. Также унифицированная разметка позволит ученым, проводящим исследования на мультиязычных параллельных корпусах, иметь единую «систему координат», в рамках которой можно изучать грамматические явления без необходимости самостоятельно решать, как адаптировать данные разных языков для сравнения. Отметим, что для типологических исследований предпочтительно, чтобы разметка в корпусе была адаптирована к морфологическому строю языка. Создание стандартов разметки для языков с богатой морфологией является отдельной исследовательской задачей.

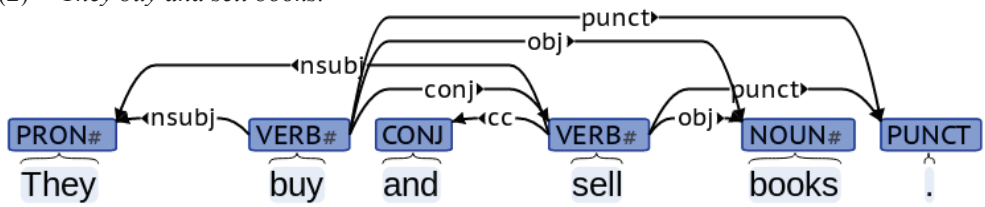
Хотя разработка единого стандарта разметки на данный момент не является до конца завершенной, неверно было бы утверждать, что такого стандарта в принципе не существует. Последние несколько лет ведется активная работа над проектом Universal Dependencies (далее UD), нацеленным на создание унифицированных принципов аннотирования, которые учитывали бы типологические особенности языков, оставаясь при этом в рамках одной идеи об устройстве разметки [Nivre et al. 2016]. Разметка UD предназначена прежде всего для синтаксического разбора в терминах структуры зависимостей, она также содержит информацию о морфологических характеристиках слов. Приведем пример UD разбора в виде схемы (2) и в формате CoNLL-U[4] [Buchholz, Marsi 2006].

На примере разметки в формате CoNLL-U видно, что токены в предложении пронумерованы, во втором столбце указана лемма, столбцы с третьего по пятый содержат морфологическую информацию, в шестом указан номер токена-вершины, в седьмом и восьмом — информация о типах зависимостей, последний столбец предназначен для дополнительной информации.

Силами разработчиков UD большое количество существующих корпусов для разных языков было конвертировано в формат UD, и затем на этих данных были обучены языковые модели для использования в парсерах. В 2017 г. на соревновании CoNLL Shared Task как основные (baseline) были заявлены два парсера: UDPipe [Straka, Straková 2017], созданный учеными Пражского университета, и SyntaxNet от компании Google [Alberti et al.

[4] Пример взят со страницы http://universaldependencies.org/format.html#sentence-boundaries-and-comments.

(2) *They buy and sell books.*



They buy and sell books.

1	They	they	PRON	PRP	Case=Nom|Number=Plur	2	nsubj	2:nsubj|4:nsubj	_
2	buy	buy	VERB	VBP	Number=Plur|Person=3|Tense=Pres	0	root	0:root	_
3	and	and	CONJ	CC	_	4	cc	4:cc	_
4	sell	sell	VERB	VBP	Number=Plur|Person=3|Tense=Pres	2	conj	0:root|2:conj	_
5	books	book	NOUN	NNS	Number=Plur	2	obj	2:obj|4:obj	_
6	.	.	PUNCT	.	_	2	punct	2:punct	_

2017]. Оба используют языковые модели, обученные на корпусах UD, на данный момент доступны модели для 40 языков, среди них немало индоевропейских: английский, немецкий, французский, испанский, итальянский, хинди и др., а также в списке присутствуют языки из других семей — турецкий, индонезийский, арабский, казахский, финский, баскский и др. Средняя точность получаемой разметки варьирует от 64 % до 93 % в зависимости от языка и объема материала, использованного для обучения языковой модели.

На сегодняшний день в стандарте UD имеются лакуны, например не выработаны окончательные правила отражения эллиптических связей, нет решения, которое бы позволило отражать в разметке структуры словоформ в языках с богатой морфологией. Несмотря на это, имеющийся стандарт разметки является уже в значительной мере проработанным и может применяться для аннотирования корпусов. Версии стандарта разметки время от времени обновляются, однако выход новой версии не станет проблемой, поскольку вместе с новым стандартом разметки выходят обновленные языковые модели для анализаторов.

Выше мы перечислили ряд аспектов, на которые следует обратить особое внимание при разработке мультиязычных параллельных корпусов для типологических исследований. На данный момент исследователи обладают необходимыми инструментами для создания таких корпусов. Остается лишь определиться, какие коллекции параллельных текстов, представленных на множестве разноструктурных языков, могут быть использованы для создания мультиязычных параллельных корпусов.

6. Выводы

Мультиязычные параллельные корпуса заслуживают внимания лингвистов, поскольку такой формат данных позволит расширить возможности для типологических исследований. Прежде всего это касается корпусных исследований и применения в лингвистическом анализе квантитативных методов, которые становятся все более распространенными. На сегодняшний день имеется небольшое количество типологических исследований, выполненных с использованием мультиязычных параллельных корпусов. Предпочтительным материалом в этих работах оказываются переводы Библии и тексты религиозных брошюр ввиду своей доступности и представленности на большом количестве языков. Однако при использовании этих текстов исследователям так или иначе приходится решать

проблему отсутствия разметки, например, используя процедуру трансфера или ориентируясь на формы одного слова (в [Cysouw 2014] слово ‘Библия’), которое однозначно извлекается из текстов. Разработка унифицированной разметки, ориентированной на нужды типологов, и применение ее в обработке массово параллельных текстов позволит создать мультиязычные параллельные корпуса, подходящие для типологического сравнения. Наличие соответствующего языкового материала должно способствовать появлению большего количества квантитативных корпусных исследований в области типологии.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Бонч-Осмоловская, Нестеренко 2018 — Бонч-Осмоловская А. А., Нестеренко Л. В. Сети как инструмент поиска и находок в мультиязычных параллельных корпусах. *ЕВРика! Сборник статей о поисках и находках к юбилею Е. В. Рахилиной.* Рыжова Д. А., Добрушина Н. Р., Бонч-Осмоловская А. А., Выренкова А. С., Кюсева М. В., Орехов Б. В., Резникова Т. И. (ред.). М.: Лабиринт, 2018, 305–320. [Bonch-Osmolovskaya A. A., Nesterenko L. V. Networks as an instrument for search and findings in multilingual parallel corpora. *EVRika! Sbornik statei o poiskakh i nakhodkakh k yubileyu E. V. Rakhilinoi.* Ryzhova D. A., Dobrushina N. R., Bonch-Osmolovskaya A. A., Vyrenkova A. S., Kyuseva M. V., Orekhov B. V., Reznikova T. I. (eds.). Moscow: Labirint, 2018, 305–320.]

Добровольский 2009 — Добровольский Д. О. Корпус параллельных текстов в исследовании культурно-специфичной лексики. *Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы.* СПб.: Нестор-История, 2009, 383–401. [Dobrovol’skij D. O. Corpus of parallel texts in research of culture-specific vocabulary. *Natsional’nyi korpus russkogo yazyka: 2006–2008. Novye rezul’taty i perspektivy.* St. Petersburg: Nestor-Istoriya, 2009, 383–401.]

НКРЯ — *Национальный корпус русского языка*. [*Natsional’nyi korpus russkogo yazyka* [Russian National Corpus].] URL: http://www.ruscorpora.ru.

Сичинава 2015 — Сичинава Д. В. Параллельные тексты в составе Национального корпуса русского языка: новые направления развития и результаты. *Труды Института русского языка им. В. В. Виноградова*, 2016, 6: 194–235. [Sitchinava D. V. Parallel texts within the Russian National Corpus: New development paths and results. *Trudy Instituta russkogo yazyka im. V. V. Vinogradova*, 2016, 6: 194–235.]

Alberti et al. 2017 — Alberti C., Andor D., Bogatyy I., Collins M., Gillick D., Kong L., Koo T., Ma J., Omernick M., Petrov S., Thanapirom C., Tung Z, Weiss D. *SyntaxNet Models for the CoNLL 2017 Shared Task.* 2017. URL: http://arxiv.org/abs/1703.04929.

Asgari, Schütze 2017 — Asgari E., Schütze H. *Past, Present, Future: A computational investigation of the typology of tense in 1000 languages.* URL: http://arxiv.org/abs/1704.08914. 2017.

Bickel et al. 2008 — Bickel B., Comrie B., Haspelmath M. *The Leipzig Glossing Rules. Conventions for interlinear morpheme by morpheme glosses* (Revised version of February 2008). URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.

Brown et al. 1993 — Brown P., Della Pietra S., Della Pietra V., Mercer R. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 1993, 19(2): 263–311.

Buchholz, Marsi 2006 — Buchholz S., Marsi E. CoNLL-X shared task on multilingual dependency parsing. *Proc. of the 10th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2006, 149–164.

Callison-Burch et al. 2004 — Callison-Burch C., Talbot D., Osborne M. Statistical machine translation with word- and sentence-aligned parallel corpora. *Proc. of the 42nd Annual Meeting of Association for Computational Linguistics*. Association for Computational Linguistics, 2004, 175–182.

Chen, Nie 2000 — Chen J., Nie J. Y. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. *Proc. of the 6th Conference on Applied Natural Language Processing*. Association for Computational Linguistics, 2000, 21–28.

Cysouw 2014 — Cysouw M. Inducing semantic roles. *Perspectives on semantic roles*. Luraghi S., Narrog H. (eds.). Amsterdam: Benjamins, 2014, 23–68.

Cysouw, Wälchli 2007 — Cysouw M., Wälchli B. Parallel texts: using translational equivalents in linguistic typology. *STUF-Sprachtypologie und Universalienforschung*, 2007, 60(2): 95–99.

Čermák, Rosen 2012 — Čermák F., Rosen A. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 2012, 17(3): 411–427.

Dahl 2007 — Dahl Ö. From questionnaires to parallel corpora in typology. *STUF-Sprachtypologie und Universalienforschung*, 2007, 60(2): 172–181.

Davis, Dunning 1995 — Davis M. W., Dunning T. Query translation using evolutionary programming for multi-lingual information retrieval. *Evolutionary Programming*, 1995: 175–185.

Gale, Church 1991 — Gale W. A., Church K. W. Identifying Word Correspondences in Parallel Texts. *HLT*, 1991, 91: 152–157.

Koehn et al. 2003 — Koehn P., Och F. J., Marcu D. Statistical phrase-based translation. *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Vol. 1. Association for Computational Linguistics, 2003, 48–54.

Koehn 2005 — Koehn P. Europarl: A parallel corpus for statistical machine translation. *MT Summit*, 2005, 5: 79–86.

Mayer, Cysouw 2012 — Mayer T., Cysouw M. Language comparison through sparse multilingual word alignment. *Proc. of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Association for Computational Linguistics, 2012, 54–62.

Mayer, Cysouw 2014 — Mayer T., Cysouw M. Creating a massively parallel Bible corpus. *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA), 2014, 3158–3163.

Megyesi et al. 2010 — Megyesi B., Dahlqvist B., Csato E., Nivre J. The English-Swedish-Turkish Parallel Treebank. *Proc. of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA), 2010, 3393–3397.

Nie et al. 1999 — Nie J. Y., Simard M., Isabelle P., Durand R. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 1999, 74–81.

Nivre et al. 2016 — Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Haji J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D. Universal Dependencies v1: A multilingual treebank collection. *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, 1659–1666.

Östling 2015 — Östling R. Word order typology through multilingual word alignment. *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Vol. 2: Short papers. 2015, 205–211.

Östling 2016 — Östling R. Studying colexification through massively parallel corpora. *The lexical typology of semantic shifts*, 2016, 58: 157.

Sahlgren, Karlgren 2005 — Sahlgren M., Karlgren J. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 2005, 11(03): 327–341.

Sharoff 2002 — Sharoff S. Meaning as use: Exploitation of aligned corpora for the contrastive study of lexical semantics. *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. European Language Resources Association (ELRA), 2002, 447–452.

Sitchinava 2012 — Sitchinava D. Parallel corpora within the Russian National Corpus. *Prace Filologiczne*, 2012, 63: 271–278.

Stambolieva 2011 — Stambolieva M. Parallel corpora in aspectual studies of non-aspect languages. *Proc. of The Second Workshop on Annotation and Exploitation of Parallel Corpora.* Association for Computational Linguistics, 2011, 39–42.

Straka, Straková 2017 — Straka M., Straková J. Tokenizing, POS-tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* 2017, 88–99.

Tiedemann 2009 — Tiedemann J. News from OPUS — a collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing*, 2009, 5: 237–248.

Vavřín, Rosen 2008 — Vavřín M., Rosen A. Intercorp: A multilingual parallel corpus. *Proc. of the International Conference "Corpus Linguistics".* St. Petersburg: Saint Petersburg State Univ., 2008, 156–162.

Waldenfels 2006 — von Waldenfels R. Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment. *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, 9. Brehmer B., Zdanova V., Zimny R. (eds.). München: Otto Sagner, 2006: 123–138.

Wälchli 2007 — Wälchli B. Advantages and disadvantages of using parallel texts in typological investigations. *STUF — Sprachtypologie und Universalienforschung*, 2007, 60(2): 118–134.

Wälchli, Cysouw 2012 — Wälchli B., Cysouw M. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 2012, 50(3): 671–710.

Получено / received 13.11.2017 *Принято / accepted 14.06.2018*