УДК 2-3

О ПЕРСПЕКТИВАХ ИСПОЛЬЗОВАНИЯ СТАТИСТИЧЕСКОГО МЕТОДА АНАЛИЗА ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ В ЛИНГВОДИДАКТИКЕ

А.Н. Магомедова
Дагестанский государственный университет

Аннотация. Статья посвящена перспективе использования в лингводидактике – методике обучения английскому языку – результатов статистической обработки параллельных текстов, которая в настоящее время находит свое применение в новых направлениях прикладной и компьютерной лингвистики. В работе рассматриваются основные понятия корпусной лингвистики в их связи с методами и математическими моделями контекстной обработки в плане учебной деятельности по иностранному языку. Показано, что статистические методы предоставляют широкие возможности для изучения контекстов, что подтверждается полученными результатами, которые требуют анализа и обобщения.

Ключевые слова: корпусная лингвистика, статистический метод, анализ параллельных текстов, машинный перевод, обработка контекста, корпус, преподавание английского языка, лингводидактика, перспективы.

Опыт преподавательской деятельности и анализ параллельных текстов привели нас к мысли о том, что статистические методы анализа параллельных текстов, нашедшие свое применение в корпусной лингвистике, могут быть использованы и в методике обучения иностранному (английскому) языку бакалавров, магистрантов и аспирантов, а полученные результаты только подтвердили наше предположение.

Обратимся к истории вопроса.

В последние десятилетия информационные технологии предоставили возможность изучать язык не только на традиционном материале словарей, определенного круга художественных произведений и других письменных текстов, но и вводить языковой материал в компьютер и обрабатывать там большие массивы текстов, называемые корпусами текстов [6, с. 123–128]. Корпус текстов – это «*большой объем живого, «реального» языкового материала, извлеченного из разнообразных источников и сведенного в компьютеризованную систему с тем, чтобы исследователи, в особенности лексикографы, могли изучать значение и возникающие языковые закономерности*» [1, с. 46–62]. Используя обширные корпуса текстов, можно получить новейшие сведения о функционировании изучаемого языка. Значительные результаты в использовании корпусных данных уже достигнуты, в частности, в лексикографии и компьютерной лингвистике [2].

Рост внимания к статистическим методам обработки языкового материала привел к разработке целого ряда исследовательских приемов, связанных с

использованием параллельных или близких текстов на разных языках [2]. Применяя к таким текстам методы статистики, исследователи находят новые решения некоторых классических задач прикладной лингвистики. Так, взгляд на язык как на код повлек за собой применение статистических методов исследования текстов с целью выявления в них определенных лингвистических закономерностей. В результате были сформулированы основные положения дистрибутивной теории, занимающейся изучением текстового поведения языковых элементов и их последующей всесторонней характеристикой. Дистрибутивная теория, в свою очередь, стимулировала становление теоретического языкознания, в котором с этого времени стало применяться моделирование, понимаемое как построение моделей, объясняющих действие языковых законов или проверяющих работу и эффективность воспроизводящих языковые действия кибернетических устройств [4, с. 109–115].

Выполняя перевод с одного языка на другой, переводчик принимает решения, основываясь на знаниях обо всех уровнях межъязыковых соответствий – лексическом, грамматическом, идиоматическом и т.д., однако в его обязанности не входит, например, составление учебной литературы. Создание словарей и учебников, формализация процесса перевода для уровня автоматических систем происходят отдельно от собственно перевода. Если бы мы смогли проанализировать, формализовать и документировать мысль переводчика, воплощенную в паре «оригинал-перевод», мы смогли бы достичь иного, более высокого, уровня эффективности использования переводческого труда.

Системы статистической обработки параллельных текстов делают шаг вперед в этом направлении. Возможность проведения подобных работ ограничивают: а) недостаточное количество материала в машинном формате, б) невысокое быстродействие компьютеров, в) дороговизна систем хранения информации [7]. Вместе с тем, всеобщая компьютеризация, действие закона Г. Мура и снижение стоимости хранения единицы информации позволяют сделать статистическую обработку параллельных текстов ***более*** реальной [там же]. Дело в том, что крайне простые системы, практически не использующие знания о структурной организации языка, но способные обрабатывать крупные корпуса параллельных текстов, оказались в состоянии конкурировать с программами, использующими сложный грамматический и семантический анализ по созданным вручную правилам.

Первые работы с параллельными текстами были выполнены в конце 80-х – начале 90-х годов прошлого века в рамках разработки различных систем ***статистического машинного перевода. С***пециалисты фирмы IBM Ф. Дженилек, П. Браун и другие создали первую систему машинного перевода, извлекающую знания о языке оригинала, языке перевода и правилах перевода исключительно из массива примеров переводов. Несмотря на то, что запатентованная фирмой IBM система ***Кандид*** обеспечивала требуемое качество перевода, эти работы вызвали критику со стороны представителей традиционной лингвистики, так как в них отвергались подходы к машинному переводу, считавшиеся общепринятыми в течение десятилетий. В частности, в

качестве модели языка использовались биграммы и триграммы, то есть, последовательности из двух или трех слов, на непригодность которых для целей моделирования языка Н. Хомский указывал еще в 50-е годы [7]. Кроме того, система не пыталась использовать какие-либо знания о структуре предложения [там же].

Несмотря на эти и другие, не названные здесь недостатки, статистический метод в применении к машинному переводу привлек к себе внимание колоссальным сокращением трудоемкости построения новых систем по сравнению с традиционными системами. Достоинством такого подхода является, во-первых, отказ от ручного составления переводных словарей и грамматик [7]. Во-вторых, если в логике системы обнаруживается ошибка, ее устранение в худшем случае означает необходимость повторного запуска процедуры извлечения параметров из корпуса примеров, а не ручное переписывание этих ресурсов [там же]. В-третьих, первые подобные системы перевода задумывались как несвязанные с конкретной парой языков. Чтобы добавить новую пару языков в систему, предполагалось всего лишь обработать соответствующий корпус текстов [там же].

Первые системы машинного перевода на основе контекстов послужили толчком для дальнейших исследований в области извлечения лингвистической информации из параллельных текстов. С одной стороны, развивались сами системы статистического машинного перевода – за счет разработки более сложных статистических моделей, отказа от идеи абсолютной независимости алгоритмов от обрабатываемых языков, привлечения морфологического разбора. Так, в системах статистического перевода не используются развитые формальные грамматики, однако понятие класса слов уже стало частью таких систем [7; 8; 10].

С другой стороны, произошла некоторая переоценка роли обработанных параллельных текстов. Алгоритмы обучения на параллельных корпусах стали применяться в инструментах, облегчающих труд переводчика – человека, а также в системах автоматической проверки переводов, выполненных вручную. К первым относятся системы известные как «переводная память» (translation memory), они используются при работе над большими объемами схожих текстов. Такие системы ищут в массиве уже переведенного текста предложения и обороты, совпадающие с переводимой фразой, и показывают их переводы. Промежуточное место между переводной памятью и статистическим машинным переводом занимают системы машинного перевода на основе примеров (EBMT – example-base machine translation), которые составляют законченные предложения из фрагментов, хранящихся в памяти машины. Вторые включают использование статистических алгоритмов для контроля полноты перевода и единства используемой при переводе терминологии [7].

Еще одной проблемой, решаемой с помощью анализа параллельных текстов, стала двуязычная лексикография. Автоматическое составление переводного конкорданса является одной из задач, решаемой любой системой статистического машинного перевода [9]. К настоящему времени составление переводных конкордансов выделилось в отдельную область исследований [там

же], и именно в этой области исследования по статистической обработке параллельных текстов достигли наибольших успехов. Так, некоторые системы достигают точности установления переводных соответствий на уровне слов выше 95% [там же].

В середине 90-х годов прошлого века автоматическое построение переводных словарей ***по*** параллельным текстам стало рассматриваться как отдельная задача, и к настоящему времени дало новый, многообещающий подход – установление лексических переводных соответствий по несвязанным текстам [там же].

Статистические методы обработки параллельных текстов стали частью систем поиска информации на разных языках (cross-language information retrieval).

Наконец, двуязычные корпуса могут представлять интерес и для исследований в рамках одного языка, например, в области автоматического разрешения многозначности.

За годы работы с параллельными текстами исследователями было предложено значительное количество подходов и конкретных алгоритмов их анализа. Некоторые из подходов и алгоритмов специфичны для решения отдельных задач, другие применимы во всех областях обработки параллельных текстов. Ниже рассматриваются некоторые общие решения проблемы обработки параллельных текстов, а также основные подходы в области статистического машинного перевода.

Первое, что необходимо для статистических исследований параллельных текстов, – это собственно корпус параллельных текстов. В зависимости от сферы исследования этот корпус используется как: а) материал для тренировки системы, б) источник для построения конкорданса, в) объект статистических исследований.

Первые эксперименты в области статистического машинного перевода были проведены ***на*** так называемом корпусе Хансарда, представляющем собой стенограммы заседаний парламента Канады на английском и французском языках. Канадские парламентарии выступают на любом из двух официальных языков страны, а после заседания все выступления переводятся и публикуются – в частности, в электронном виде. Объем корпуса Хансарда составляет сотни миллионов слов [9]. Кроме того, крупный корпус тестов был собран в рамках европейского проекта АРКАДА: более 10 миллионов слов на девяти языках, то есть, примерно по 1,1 миллиону слов на каждый язык [там же]. Основой этого корпуса являются запросы членов Европейского парламента и ответы на них, опубликованные в Журнале Европейского Союза [там же]. В других экспериментах также использовались фрагменты многоязычного корпуса ООН, документы различных учреждений Швейцарии и «Гонконгского Хансарда». Привлекались также выдержки из текстов российских законов, переведенные на английский язык компаниями Гарант и Кодекс [там же].

Подавляющее большинство существующих параллельных текстов не дают возможности непосредственного извлечения информации о переводных соответствиях. Этому препятствует: 1) отсутствие однозначного и линейного

соответствия на уровне слов между текстом оригинала и переводом. В любой паре языков существуют различия в лексико-грамматической структуре и в идиоматике; 2) разный уровень профессиональной подготовки конкретных переводчиков.

Тем не менее, разработка различных аспектов перспективного направления использования статистических методов при анализе параллельных текстов продолжается. Перспективу этого направления мы видим в привлечении результатов подобных исследований к решению задач лингводидактики, в частности, в сфере обучения неродным или иностранным языкам.

Литература

1. *Гвишиани Г.Б., Герви О.Ю*. Корпусная лингвистика и грамматика речи // Вестник МГУ. - Серия 9. Филология. - Вып. 2. - М.: Изд-во МГУ, 2001.

2. *Магомедова А.Н.* Анализ параллельных текстов как один из методов корпусной лингвистики // МГУ, лаборатория общей и компьютерной лексикологии и лексикографии, 2002. URL: www.philol.msu.ru (Дата обращения: 01.12.2016).

3. *Магомедова А.Н.* Слово в коммуникативном аспекте // Голоса молодых ученых. - Вып. 13. - М.: МГУ, 2003.

4. *Магомедова А.Н*. Исходный корпус текстов и методы его обработки // Вестник ДНЦ РАН. - Вып. 1 (12). - Махачкала: Изд-во ДНЦ, 2002.

5. *Магомедова А.Н.* Проблема анализа параллельных текстов и машинного перевода в корпусной лингвистике // Вестник ЦМО МГУ. Филология. Культурология. Педагогика. Методика. - 2013. - № 2. – М.: ЦМО МГУ, 2013.

6. *Марчук Ю.Н., Магомедова А.Н.* Корпусная лингвистика и контекст // Теоретические и практические аспекты лингвистики и лингводидактики. - Сургут: Изд-во СурГУ, 2002.

7. Методы работы с параллельными текстами. URL: http://semantic-evolution.narod.ru/Thesis2003_2_3_1.htm (Дата обращения: 03.12.2016).

8. *Brown P., Cocke J., Della Pietra S., Della Pietra V., Jenilek F., Lafferty J., Mercer R., Roossin P.* A statistical approach to Machine Translation in Computational Linguistics, 16 (2), 1990.

9. *Macklovitch E.* Can Terminological consistency be validated automatically? Lexicommatique et dictionariques // Proceedings of the 4-th journees scientifiques. Laval, Canada, 1995.

10. *Rapp R.* Automatic identification of word translations from unrelated English and German Corpora // Proceedings of the 37-th annual meeting of the Association for Computational Linguistics. Maryland, USA, 1999.

References

1. *Gvishiani G.B., Gervi O.Y.* Korpusnaya lingvistika i grammatika rechi // Vestnik MGU. - Seriya 9. Filologiya. - Vyp. 2. - M.: Izd-vo MGU, 2001.

2. *Magomedova A.N.* Analiz parallel'nyh tekstov kak odin iz metodov korpusnoj lingvistiki // MGU, laboratoriya obshchej i komp'yuternoj leksikologii i leksikografii, 2002. URL: www.philol.msu.ru (Data obrashcheniya: 01.12.2016).

3. *Magomedova A.N.* Slovo v kommunikativnom aspekte // Golosa molodyh uchenyh. - Vyp. 13. - M.: MGU, 2003.

4. *Magomedova A.N.* Iskhodnyj korpus tekstov i metody ego obrabotki // Vestnik DNC RAN. - Vyp. 1 (12). - Mahachkala: Izd-vo DNC, 2002.

5. *Magomedova A.N.* Problema analiza parallel'nyh tekstov i mashinnogo perevoda v korpusnoj lingvistike // Vestnik CMO MGU. Filologiya. Kul'turologiya. Pedagogika. Metodika. - 2013. - № 2. – M.: CMO MGU, 2013.

6. *Marchuk Y.N., Magomedova A.N.* Korpusnaya lingvistika i kontekst // Teoreticheskie i prakticheskie aspekty lingvistiki i lingvodidaktiki. - Surgut: Izd-vo SurGU, 2002.

7. Metody raboty s parallel'nymi tekstami. URL: http://semantic-evolution.narod.ru/Thesis2003_2_3_1.htm (Data obrashcheniya: 03.12.2016).

8. *Brown P., Cocke J., Della Pietra S., Della Pietra V., Jenilek F., Lafferty J., Mercer R., Roossin P.* A statistical approach to Machine Translation in Computational Linguistics, 16 (2), 1990.

9. *Macklovitch E.* Can Terminological consistency be validated automatically? Lexicommatique et dictionariques // Proceedings of the 4-th journees scientifiques. Laval, Canada, 1995.

10. *Rapp R.* Automatic identification of word translations from unrelated English and German Corpora // Proceedings of the 37-th annual meeting of the Association for Computational Linguistics. Maryland, USA, 1999.

ABOUT THE PROSPECTS OF USING STATISTICAL METHOD OF THE ANALYSIS OF PARALLEL TEXTS IN LANGUAGE TEACHING

A.N. Magomedova
Dagestan State University

Abstract: The research under study sheds light upon the prospect of the use in language teaching methods of teaching the English language, presenting results of statistical processing of parallel texts, currently finds its application in new trends in Applied and Computational Linguistics. The paper discusses the main concepts of Corpus Linguistics in their relationship with the methods and mathematical models of contextual treatment in terms of learning activities in a foreign language. It is shown that statistical methods provide opportunities to study the contexts, as evidenced by the results, requiring analysis and synthesis.

Keywords: Corpus Linguistics, statistical method, analysis of parallel texts, machine translation, processing of context, corpus, teaching English, Linguistics, perspectives.

Сведения об авторе

Магомедова Адигат Нурахмагаджиевна, кандидат филологических наук, доцент кафедры иностранных языков для гуманитарных факультетов, Дагестанский государственный университет (Махачкала, Россия), член редакционной коллегии журнала «Дидактическая филология».

Рецензент

Нифанова Татьяна Сергеевна, доктор филологических наук, профессор кафедры общего и германского языкознания, Северодвинский филиал Северного (Арктического) федерального университета имени М.В. Ломоносова (Северодвинск, Россия).