O‘ZBEKISTON RESPUBLIKASI
OLIY VA O‘RTA MAXSUS TA’LIM VAZIRLIGI

ALISHER NAVOIY NOMIDAGI
TOSHKENT DAVLAT
O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

“O‘ZBEK TILINING MILLIY KORPUSI: MUAMMO VA VAZIFALAR”

mavzusidagi
xalqaro ilmiy-amaliy konferensiya materiallari

(2022-yil 31-may)

Tashkent-2022

2-sho‘ba. O‘ZBEK TILINING MILLIY KORPUSINI YARATISH MASALALARI

ПАРАЛЛЕЛЬНЫЕ КОРПУСЫ, КАК ЭФФЕКТИВНЫЙ МЕТОД ОБУЧЕНИЯ ИНОСТРАННОГО ЯЗЫКА.

Нигматова Лолахон Хамидовна*

Аннотация. В статье рассматривается потенциал включения материалов корпусной лингвистики в методику обучения иностранным языкам. Приводятся и анализируются примеры использования информационной базы корпусных интернет-технологий в преподавании иностранных языков.

Annotation. The article discusses the potential of including corpus linguistics materials in foreign language teaching methods. It provides the analysed examples of using the information base of corpus Internet technologies in teaching foreign languages.

Ключевые слова: интернет-технологии, корпусная лингвистика, электронный корпус, обучение иностранному языку.

Key words: Internet technologies, corpus linguistics, electronic corpus, teaching a foreign language.

Введение

Информационно-коммуникативные технологии являются одним из основных помощников в процессе обучения иностранному языку у учащихся и студентов. [Ю.Ю.Маркова, 2011]. Корпусная лингвистика–является одной из технологий, на основе которой можно формировать лексические и грамматические навыки речи.

Одной из главных задач разработчиков лингвистического корпуса подбор большого количества текстов, помогающих для изучения иностранного языка. В этой связи можно заключить, что корпус – это уменьшенная модель конкретного языка. Раз это уменьшенная копия языка, то и как любой язык, корпус – репрезентативен. т. е. обязательное пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т. д. Таким образом, ***лингвистический корпус*** – это система текстов, расположенных на машинном источнике и объединенных по различным параметрам (язык, жанр, стиль, отрасль). Нам всем известно, что корпусная лингвистика непосредственно связана с

*д.ф.н. Бухарского госуниверситета.

компьютерными технологиями и всю обработку проводит машина. Одной из программ, помогающих в выполнении этих услуг является конкорданс, позволяющих обрабатывать и анализировать большие объемы текстов и находить в них лингвистические связи. Конкорданс помогает при поиске слова, являясь поисковой командой программа выдает фрагменты и количество слов и словосочетаний с запрашиваемой лексемой, а также рассмотреть все связи этой единицы в языковом пространстве. [В.П.Захаров,2005:47].

В центре внимания нашего исследования выступает корпус параллельных текстов, поэтому видится целесообразным подробнее остановиться на таком типологическом признаке, как «параллельность». Отметим, что основная часть методических исследований была посвящена изучению вопросов формирования грамматических и лексических навыков речи на основе монолингвальных одноязычных корпусов [П.В. Сысоев,2010:99]. Однако кроме одноязычных лингвистических корпусов существуют еще и двуязычные и многоязычные корпусы, т. е. корпусы, в которых представлены тексты на двух и более языках. Многоязычные корпусы можно условно разделить на несколько типов:

корпус, в котором представлен текст оригинала и его перевод;

одноязычный корпус, в котором представлен текст оригинала и его толкование на этом же языке

Необходимо отметить, что о дни ученые считают первый тип корпуса – корпусом переводов, а второй тип корпуса – корпусом параллельных текстов [Aijmer K,1996]. Другие же, на оборот, первый тип корпуса – корпусом параллельных текстов, а второй – сравнительным корпусом [Baker M,1999:281]. Для третьей категории ученых оба типа являются корпусами параллельных текстов [Johansson S,1999:3]. В нашем понимании **корпусом параллельных текстов** это тип **лингвистического корпуса, состоящий из исходного текста на одном языке и его перевода на другой или другие языки**.[А.А.Кокорева, 2013:57]

Так как параллельный корпус узбекско- русского языка находится на начальной стадии разработки, наши примеры несут чисто теоретический характер. Корпусы параллельных текстов по своим типовым особенностям могут также разделяться на двуязычные и многоязычные.

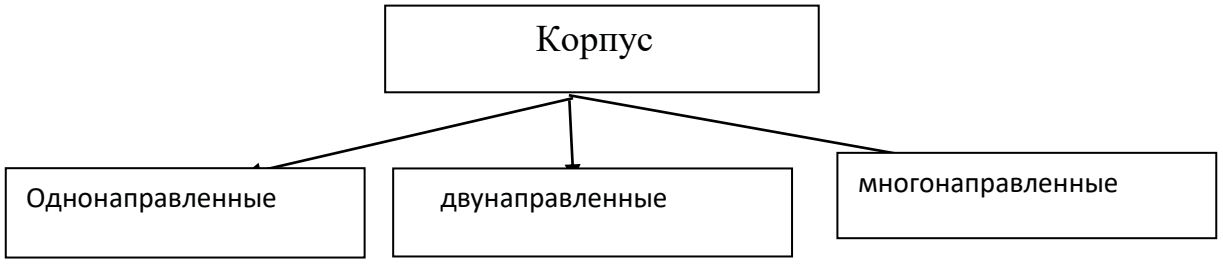
Двуязычные – это текст – оригинал и тест- перевод на другом языке.

Узбекский язык(текст – оригинал)	Русский язык (текст-перевод, М.Сафаров)
-Аҳли мажлис Отабекни кўкларга кўтариб мақтар эди,	– Собравшиеся дружно превозносили Атабека, однако

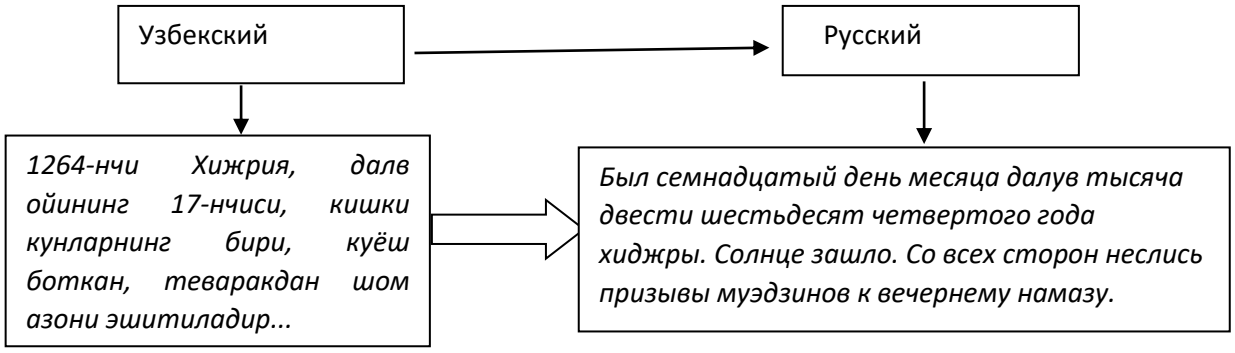
лекин Ҳомид бу мақташларга иштирок этмас ва нимадандир ғижингандек кўринар эди. Шу орада қутидорнинг “уйланганми?”деб Хасаналидан сўраши Ҳомидга яна бошқача ҳолат берди.	Хамид в этом не учавствовал, казалось чем-то раздосадованным и уже совсем изменился в лице, когда кутидор спросил у Хасанали, не женат ли бек.

Многоязычные корпусы – это текст – оригинал и его тексты переводы на другие языки, а тексты имеющие переводы разных авторов называются поливариантными. Например, роман А.Кодири «Минувшие дни» переведен на русский, английский, китайский, немецкий и др. языки.

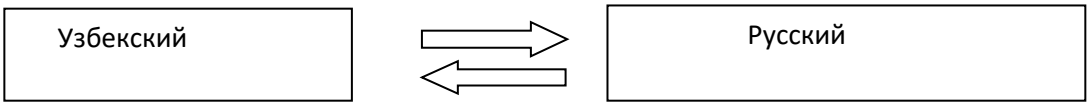
По направленности



Однонаправленные корпусы – это перевод с одного языка на другой например, с узбекского на русский, без возможности обратного перевода.



Двунаправленные корпусы включают параллельные тексты на двух языках узбекском и русском и позволяют осуществлять перевод как с узбекского на русский, так и наоборот.



Многонаправленный корпус – состоит из текстов на более чем двух языках и позволяют осуществлять перевод с любого языка на любой (в рамках существующих языков корпуса), например роман А. Кодири «Минувшие дни» был переведен Марком Эдвардом Ризом , основываясь на русский перевод М.Сафарова.

Основываясь на вышеуказанные языковые типы и направленность корпусов параллельных текстов, можно составить примерный план обучения студентов иностранной профессионально-ориентированной лексике на основе которого можно обучать студентов профессиональной лексике.

Типы	Характеристика
Литературность	Использование корпуса письменных текстов профессиональной направленности, так как студенты в процессе работы часто пользуются академической литературы .
Перевод	Двунаправленный, позволяющие осуществлять перевод как с узбекского на русский, так и наоборот
Языковые	двуязычные корпусы, включающие исходные тексты на одном языке и их переводы на другом.
Динамичность	использование корпуса, включающего современные тексты по научной проблематике
Разметка	В процессе обучения должны использоваться корпусы с морфологической, синтаксической, семантической разметкой
Объем текстов	Полнотекстовые и фрагментотекстовые корпусы

Очень часто при переводе с узбекского на русский или обратно слова переводятся на другой язык рядом синонимов например

ЗАРҲАЛ 1 золоченый, позолоченный, ~ бериш [зарҳал бериш, деб ўқилади] [Акобиров, 1988: 25]

АДАШМОҚ 1 блуждать, плутать, заблудиться, сбиваться с пути; ўрмоида адашиб қолмоц заблудиться в лес\; адашиб юрвшк блуждать:

2 бродить, странствовать- Тоғлар ошди, қир ошди. Ойлар бўйи адашди (Ҳ. Олим- жон, «Семурғ») Прошёл он горы, прошёл он долы. Месяцами скитался; адашиб улоқиб юрмок странствовать; бродить; 3 заблуждаться, ошибаться, сбиваться (с толку, со счёта), ҳисобдан —

сбиваться со счёта; терять счёт; 4 потерять (кого-л.); отбиться (от кого-л); ёридан — потерять возлюбленную (возлюбленного); подадан адашган бузоқ телёнок, отбившийся от стада.[С.Акобиров, 1988: 10]

Благ/о, -а с.–яхшилик, эзгулик, бахт, бахт--саодат; манфаат, фойда; трудиться на благо Родины Ватан бахт-саодати учун мехнат килмок, на благо народа халқ манфаатини кўзлаб. [Т.Аликулов,1982:34]

Как видим из приведенных примеров, Такой перевод может привести к употреблению синонимичных слов в несвойственном им контексте. При использовании параллельного корпуса и проанализировав примеры параллельного корпуса, можно чётко определить грань между словами ***бахт*** ва ***манфаат***. Мен бахтли ман .– Я счастливый.

Менинг манфаатим – Мои интересы. Как видим слова- синонимы слова ***благо*** при переводе контекстуально несут в себе другую информацию. При пользовании же правильно составленного параллельного корпуса, студенты обучающие иностранный язык смогут выбрать правильный перевод. Параллельные корпуса – это лучший способ показать правильное употребление предлогов. Так, в русском языке предлог направления “в” и “на ” и случаи их употребления с определенными словами нужно запоминать отдельно например – *в Узбекистане, в Ташкенте*, но *на Кавказе, на Урале*. *Сел в электричку. – но Приехал на электричке.*

Таким образом, с помощью национального корпуса процесс интегрирования в незнакомую языковую среду и запоминания подобных случаев происходит быстрее. Лингвисты не отрицают роль интуиции в изучении языка, знания, основанные на самостоятельном анализе и интуиции, значительно отличаются от основанных на явных доказательствах, предлагаемых корпусом. [Т.Ю.Антонова,2020:45.]. Использование лингвистических корпусов в преподавании иностранного языка – это новый эффективный и продвинутый способ работы с классом. В настоящее время использование корпусов перестало быть интересным только для небольшой группы лингвистов. Работа с национальными корпусами, несомненно, улучшает качество образовательного процесса, при этом повышая интерес как студентов, так и самих преподавателей. Есть все основания полагать, что корпусное языкознание узбекского языка, Национальный корпус узбекского языка будет дальше развиваться и в скором будущем повлияет на каждый аспект того, как языки преподаются, изучаются и исследуются, а корпуса узбекско-иностранных параллельных текстов, не зависимо от того английский, немецкий, русский , китайский или арабский языки, позволяют формировать у студентов лексические навыки речи посредством перевода при изучении иностранного языка для специальных целей, а также более

точно определять значения новых слов и показывать особенности словоупотребления в реальных языковых ситуациях.

СПИСОК ЛИТЕРАТУРЫ

1. Т.Ю.Антонова, Л.Ф.Шангараева. Корпусная лингвистика и обучение иностранным языкам. TERRA LINGUAE, выпуск 8 (2020) [56] Сборник научных статей, 02.03.2020-01.05.2020
2. Акобиров С. Узбекско-русский словарь . Ташкент. 1982. С.10
3. Т.Аликулов. Русско-узбекский учебный словарь . Ташкент. 1982. С.34
4. Aijmer K., Altenberg B., Johansson M. Language in contrast: Papers from a symposium on textbased cross-cultural studies. Lund, 1996.
5. Baker M. The role of corpora in investigating the linguistic behavior of professional translators // International Journal of Corpus Linguistics. 1999. № 4. P. 281-298
6. Johansson S. On the role of corpora in crosslinguistic research // Corpora and cross-linguistic research. Amsterdam, 1999. P. 3-24.
7. В.П.Захаров. Корпусная лингвистика. 2005. С.47
8. А.А.Кокорева. Корпус параллельных текстов в обучении иностранному языку . Вестник ТГУ, выпуск 2 (118), 2013
9. Ю.Ю.Маркова Методика развития умений письменной речи студентов на основе викитехнологии: автореф. дис. канд. пед. наук. М., 2011.

COMPUTATIONAL LINGUISTICS AND ITS APPLICATIONS

Zamira Salimgerey *

Abstract. The article deals with the issue of computational linguistics, the task of which can be formulated as the development of methods and means of constructing linguistic processors for various applied tasks for automatic processing of texts in natural language.

Key words: computational linguistics, artificial intelligence, information retrieval, text clustering, text mining.

The emergence of the Internet and the rapid growth of available textual information has significantly accelerated the development of a scientific field that has existed for many decades and is known as Natural Language Processing and Computational Linguistics. Within the framework of this field, many promising ideas for automatic processing of natural language texts have been proposed, which have been implemented in many application systems, including commercial ones. The scope of computational linguistics applications is constantly expanding, new tasks are emerging that are successfully solved, including involving the results of related scientific fields.

The scientific achievements of the field can be viewed on the website of ACL (Association of Computational Linguistics) [1] - the international Association for Computational Linguistics, which aggregates the work of numerous scientific conferences in this field.

Computational linguistics (CL) is an interdisciplinary field that arose at the intersection of such sciences as linguistics, mathematics, Computer Science, Artificial Intelligence.

In its development, it still chooses and applies (if necessary, adapting) the methods and tools developed in these sciences. The origins of CL go back to the research of the famous American linguist N. Chomsky on the formalization of the structure of natural language [2], to the first experiments on machine translation performed by programmers and mathematicians, as well as to the first programs developed in the field of artificial intelligence for understanding natural language (for example, [3]).

Since natural language texts are the object of processing in the CL, its development is impossible without basic knowledge in the field of general

* Senior Lecturer, Master of Arts the Department of World Languages Aktobe Regional University named after K. Zhubanov Aktobe, Republic of Kazakhstan, zamirasalimgerey26@gmail.com

linguistics (linguistics) [4]. Linguistics studies the general laws of natural language — its structure and functioning, and includes such areas:

*phonology — studies the sounds of speech and the rules of their connection in the formation of speech;

*morphology — deals with the internal structure and external form of speech words, including parts of speech and their categories;

*syntax — studies the structure of sentences, the rules of compatibility and the order of words in a sentence, as well as its general properties as a unit of language;

*semantics and pragmatics are closely related fields: semantics deals with the meaning of words, sentences and other units of speech, and pragmatics — the peculiarities of expressing this meaning in connection with specific communication goals;

*lexicography describes the lexicon of a particular word — its individual words, their grammatical and semantic properties, as well as methods for creating dictionaries. Computational Linguistics is most closely related to the field of Artificial Intelligence (AI) [5], within which software models of individual intellectual functions are developed. Despite the obvious intersection of research in the field of Computational Linguistics and AI (since language proficiency refers to intellectual functions), AI does not absorb the entire CL, since it has its own theoretical basis and methodology. Common to these sciences is computer modeling as the main method and the final goal of research, the heuristic nature of many of the methods used.

In a somewhat simplified way, the task of computational linguistics can be formulated as the development of methods and means of constructing linguistic processors for various applied tasks of automatic text processing in its language. The development of a linguistic processor for some applied task involves a formal description of the linguistic properties of the text being processed (at least the simplest), which can be considered as a text model (or a language model).

The scope of CL applications is constantly expanding, so here we describe the most well-known applied problems solved by its tools. Machine Translation [6] is the earliest application of CL, with which this field itself arose and developed. The first translation programs were built in the middle of the last century and were based on the simplest strategy of word-by-word translation. However, it was quickly realized that machine translation requires a much more complete linguistic model.

Such a model was developed in the domestic STAGE system [7], as well as in several other systems that translate scientific texts. Currently, there is a whole range of computer systems of machine translation (of different quality), from large international research projects to commercial

automatic translators. Projects of multilingual translation using an intermediate language in which the meaning of the translated phrases is encoded are of significant interest. The modern direction is statistical translation, based on the statistics of translated pairs of words and phrases. Despite many decades of research on this task, the quality of machine translation is still far from perfect. A significant breakthrough in this field is associated with the use of machine learning and neural networks (originated and studied in the framework of AI).

Another rather old application of computational linguistics is Information Retrieval [8] and related tasks of indexing, abstracting, classifying and categorizing documents. Full-text document search in large databases of text documents involves indexing texts, requiring their simplest linguistic preprocessing, and the creation of special index structures. Several models of information search are known, the most well-known and used is the vector model, in which an information request is presented as a set of words, and suitable (relevant) documents are determined based on the similarity of the request and the vector of words of the document.

Modern Internet search engines implement this model by indexing texts according to the words used in them and using very sophisticated ranking procedures to issue relevant documents. The current direction of research in the field of information retrieval is multilingual document search.

Summarizing the text (Summarization) — reducing its volume and obtaining a summary of its content — an abstract, which makes it faster to search in collections of documents. The abstract can also be compiled for several documents related to the topic (for example, for a cluster of news documents). The main method of automatic abstracting is still the selection of the most significant sentences of the refereed text based on the statistics of words and phrases, as well as structural and linguistic features of texts.

The task close to abstracting is annotating the text of the document, i.e. compiling its annotation. In the simplest form, the abstract is a list of the main (key) topics of the text, for which statistical and linguistic criteria are used to highlight. When processing large collections of documents, the tasks of classification (Categorization) and clustering of texts (Text Clustering) are relevant [9].

Classification means assigning each document to a certain class with pre—known parameters, and clustering means splitting a set of documents into clusters, i.e. subsets of thematically similar documents. To solve these problems, machine learning methods are used, and therefore these applied

tasks are often referred to the Text Mining direction, considered as part of the scientific field of Data Mining (data mining) [10]. The problem of classification is becoming more widespread, it is solved, for example, when recognizing spam, classifying SMS messages, etc.

Literature:

1.Ageev M. S., Dobrov B. V., Lukashevich N. V. Automatic text categorization: methods and problems // Scientific notes of Kazan State University. Series of Physical and mathematical sciences. - 2008. — Vol. 150, No. 4. — pp. 25-40

2. Vorontsov K. V., Potapenko A. A. Regularization of probabilistic thematic models for increasing interpretability and determining the number of topics // Computational linguistics and intellectual technologies: Based on the materials of the annual International Conference "Dialog" (Bekasovo, June 4-8, 2014). — Issue 13 (20). — Moscow: Publishing House of the Russian State University, 2014. — pp. 676-687.

3.Bolelli L., Ertekin S., Giles C. L. Topic and trend detection in text collections using latent Dirichlet allocation // ECIR. — Vol. 5478 of Lecture Notes in Computer Science. — Springer, 2009. — Pp. 776–780.

4. Chien J.-T., Chang Y.-L. Bayesian sparse topic model // Journal of Signal Processessing Systems. — 2013. — Vol. 74. — Pp. 375–389.

5.Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // J. of the Royal Statistical Society, Series B. — 1977. — no. 34. — Pp. 1–38

6.Bassiou N., Kotropoulos C. Online PLSA: Batch updating techniques including outof-vocabulary words // Neural Networks and Learning Systems, IEEE Transactions on. — Nov 2014. — Vol. 25, no. 11. — Pp. 1953–1966.

7.Blei D. M., Griffiths T. L., Jordan M. I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies // J. ACM. — 2010. — Vol. 57, no. 2. — Pp. 7:1–7:30.

8.Eisenstein J., Ahmed A., Xing E. P. Sparse additive generative models of text // ICML’11. — 2011. — Pp. 1041–1048.

9.Bodrunova S., Koltsov S., Koltsova O., Nikolenko S. I., Shimorina A. Interval semisupervised LDA: Classifying needles in a haystack // MICAI (1) / Ed. by F. C. Espinoza, A. F. Gelbukh, M. Gonzalez-Mendoza. — Vol. 8265 of Lecture Notes in Computer Science. — Springer, 2013. — Pp. 265–274.

10.Blei D. M., Jordan M. I. Modeling annotated data // Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. — New York, NY, USA: ACM, 2003. — Pp. 127–134.