СОПОСТАВИТЕЛЬНЫЕ КОРПУСНЫЕ ИССЛЕДОВАНИЯ СИНТАКСИЧЕСКИХ ЕДИНИЦ (НА МАТЕРИАЛЕ АНГЛИЙСКИХ И БЕЛОРУССКИХ КОНСТРУКЦИЙ ВТОРИЧНОЙ ПРЕДИКАЦИИ)

УДК (811.111+811.161.3)'367
DOI: 10.56246/18294480-2024.16-377

МОРОЗ ЕВГЕНИЙ
Старший преподаватель кафедры современных технологий перевода Минского государственного лингвистического университета, Беларусь
электронная почта: mail4maroz@yandex.by

Цель настоящей статьи – раскрыть специфику проведения сопоставительных корпусных исследований грамматических конструкций. Возможности и ограничения таких исследований описываются с применением методов контрастивного и количественного анализа на примере объектной инфинитивной конструкции вторичной предикации в каузативных высказываниях в английском и белорусском языках. В корпусных исследованиях данные представляются в терминах классов слов безотносительно к лексическому наполнению конструкций, что дает возможность делать выводы об особенностях концептуализации и языкового кодирования знаний представителями разных языковых коллективов. В практическом плане полученные данные могут применяться в переводоведении, в частности, для усовершенствования систем машинного перевода. На материале Британского национального корпуса и Белорусского N-корпуса в статье выделяются факторы сопоставимости как требования временно´й и жанровой соотнесенности текстов корпусов, а также их репрезентативности, которая зависит не столько от объема выборки, сколько от изменения относительной частоты рассматриваемого явления в ее сегментах. Приведенный алгоритм работы с корпусами английского и белорусского языков может использоваться для создания сопоставимого репрезентативного корпуса для проведения исследования синтаксических единиц неблизкородственных языков.

Ключевые слова: *корпусная лингвистика, корпус, репрезентативность, синтаксис, вторичная предикация*

Современные лингвистические исследования естественным образом проводятся с использованием информационных технологий. Это справедливо главным образом для корпусной лингвистики – направления, которое начало свое бурное развитие в 1960-1980-х гг. с изобретением и внедрением компьютеров, позволивших значительно ускорить процесс обработки информации. Совершенствование технологий дает исследователям дополнительные возможности, хотя использование понятия «корпус» как собрания текстов имеет в лингвистике достаточно долгую «докомпьютерную» традицию. Однако в современном значении оно соотносится именно с компьютерными технологиями и обозначает «представленный в электронном виде, как правило, размеченный для анализа в лингвистических целях, обеспеченный сравнительно простой в использовании поисковой системой репрезентативный массив неотредактированных текстов, репрезентирующих максимальное множество вариантов языка»[1]. Примечательно, что корпусную лингвистику характеризует определенный циклический путь развития от теоретического к прикладному и далее опять к теоретическому направлению: размеченный корпус текстов как таковой является практическим результатом теоретических исследований и классификаций языковых явлений, но также сам становится основой для дальнейших теоретических лингвистических построений и прикладных исследований.

Составление лингвистического корпуса требует объединения технической компетенции для обработки массива текстов, установления классов и классификаций с усилиями исследователя как интерпретатора полученных данных, его языковой интуицией и умением предвосхищать результат. В то же время актуальным для любого языковедческого научного исследования остается требование его объективности, нарушение которого при опоре лишь на интуицию может привести к тому, что исследователь рискует получить заранее запрограммированный им результат. Применение выверенных и зарекомендовавших себя методик в сочетании с тщательным анализом и интерпретацией наблюдаемых явлений, однако, приводит к получению объективных данных.

Синтаксические исследования основываются на современной корпусной базе и позволяют проводить лингво-статистический анализ без привязки к лексической реализации синтаксических позиций, а также определять особенности функционирования рассматриваемых единиц. Такое направление в лингвистике получило название экспликативного синтаксиса. Предметом его исследования

---

[1] Чилингарян К.П., Корпусная лингвистика: теория vs методология // Вестник Российского университета дружбы народов, Серия: Теория языка, Семиотика. Семантика., 2021. т. 12. № 1. С. 213. DOI: 10.22363/2313-2299-2021-12-1-196-218.

являются «определяемые в терминах грамматических классов синтаксические структуры, рассматриваемые как репрезентации исходных семантических пропозиций, представляющих конфигурации вершинного (матричного) предиката и его комплементов, обусловленных заключенным в предикате лексическим понятием»[1]. Особый интерес также представляют сопоставительные корпусные исследования на материале как родственных, так и неблизкородственных языков, например, английского и белорусского. Их результаты позволяют устанавливать сходства и различия в формах языкового кодирования и представления знаний. Такие данные вносят вклад в описание языковых картин мира носителей языков, но также имеют и практическое применение, в частности, могут усовершенствовать системы машинного перевода. Работа, однако, может осложняться некоторыми ограничениями. В частности, первое из них – это проблема сопоставимости корпусов, второе – проблема их репрезентативности. Рассмотрим возможности проведения сопоставительных синтаксических исследований на примере британского национального корпуса английского языка[2] и корпуса белорусского языка[3]. В качестве синтаксической единицы для сопоставления возьмем конструкции вторичной предикации, употребляемые в каузативных высказываниях.

Вторичная предикация возникает при языковом кодировании сложных информационных структур, когда в одном предложении описывается сразу несколько положений дел. Говорящий, основываясь на своих коммуникативных установках, выделяет одно из положений дел как основное, другие – как дополнительные, фоновые. Такое основное положение находит свое выражение в предикативном центре предложения, где глагол-сказуемое в личной форме отличается грамматически выраженными категориями темпоральности, модальности и персональности. Дополнительные, фоновые ситуации (положения дел) также представляют собой пропозицию (имеют свою логическую схему), но их предикативные категории не имеют полного грамматического выражения, а зависят от основного. Такие конструкции в лингвистическом понимании представляют собой единство формы и смысла. Несводимость значения конструкции к сумме значений ее составляющих можно продемонстрировать на примере каузативных конструкций в сопоставляемых языках. Так, в английском высказывании *I let John speak freely* ‘Я позволил Джону высказаться открыто’

---

[1] Киклевич А., Корпусные исследования в синтаксисе: возможности и ограничения // Przegląd rusycystyczny. 2022, No. 1 (177). С. 105. DOI: 10.31261/pr.12755.

[2] British National Corpus (BNC). URL: https://www.english-corpora.org/bnc/

[3] Беларускі N-корпус. URL: https://bnkorpus.info/

вторично-предикативная объектная конструкция с инфинитивом *John speak* выражает каузированную ситуацию. Деление же конструкции и выведение субъектного или предикатного элементов приводит к полной потере смысла предложения: *?I let John freely* ‘букв. Я позволил Джону открыто’ или *?I let speak freely* ‘букв. Я позволил говорить открыто’. То же справедливо и для белорусского эквивалента рассматриваемого высказывания *Я даў Джону казаць адкрыта* с тем же значением. Устранение субъектного или предикатного членов конструкции изменяют смысл предложения: *Я даў Джону адкрыта* ‘Я позволил Джону открыто’ или *Я даў казаць адкрыта* ‘Я позволил высказаться открыто’. И хотя с грамматической точки зрения такие предложения могут существовать, они не передают тех смыслов, которые заложены в оригинальном высказывании с использованием конструкции.

Для проведения корпусных синтаксических исследований конструкций необходимо выделить их прототипические черты, т.е. те свойства, которые позволят обеспечить максимально независимое от контекста и лексического наполнения построение схемы ситуации. Схемы «содержат некоторое число переменных или свободных мест, предназначенных для того, чтобы быть заполненными специфическими элементами ситуации, которая будет репрезентирована этой схемой»[1]. Применительно к конструкциям вторичной предикации как языковым знакам, денотатом которых выступает ситуация, прототипическая схема может быть представлена в виде «субъектный член конструкции – предикатный член конструкции». Субъектный член выражает семантический субъект описываемой ситуации, предикатный – приписываемый ему признак. Следующим этапом является морфологическое описание прототипической схемы. Применительно к рассматриваемой каузативной конструкции в английском языке субъектный член конструкции выражается существительным в общем падеже или местоимением в объектном падеже, предикатный – инфинитивом без частицы *to*. В белорусском языке субъектный член конструкции также выражается существительным или местоимением, чей падеж зависит от согласования с каузативным глаголом (дательный или винительный), а предикатный также выражается инфинитивом.

На основе выделенных морфологических свойств, а также принципа заполнения позиций комплементов матричного каузативного глагола, формируется поисковый запрос для отбора языкового материала из корпуса. Для рассматриваемой каузативной конструкции запрос в корпусе английского языка

[1] Ришар Ж. Ф., Ментальная активность. Понимание, рассуждение, нахождение решений (сокр. пер. с франц. Т. А. Ребеко), М., Институт психологии РАН, 1998, с. 15.

может быть представлен в виде *let* _n _v?i*. Синтаксис запроса отражает порядок следования элементов конструкции, использование знака астерикс (*) после каузативного глагола позволяет включить в поиск все формы, включая 3-е лицо единственного числа. Формат запроса *_n* позволяет получить в выдаче любые формы существительных (как имена собственные, так и нарицательные) в единственном или множественном числе. Для поиска по местоимениям в запросе элемент *_n* следует заменить на *_p*. А элемент *_v?i* указывает на запрос инфинитивной формы глагола. Интерфейс корпуса позволяет не запоминать условные обозначения, а указать необходимые частеречные значения через форму *POS* (Part of Speech). В корпусе белорусского языка формула поискового запроса для необходимой конструкции представлена в виде *даць ND V<0>*. Однако интерфейс программы также позволяет не запоминать синтаксис запроса, а задавать необходимые значения через предложенную форму. Так, *ND* представляет любое существительное в дательном падеже, а *V<0>* – инфинитив. Выбор опции «Все словоформы» для каузативного глагола позволяет получить в выдаче глаголы в любой из временных или родовых форм.

Ограничением в использовании такого подхода в белорусском корпусе является омонимия грамматических форм. Так, в рассматриваемой конструкции при наличии отрицания с глаголом *даць* существительное может употребляться как в дательном, так и в родительном падеже, например, *Ён не даў яму дагаварыць* ‘Он не позволил ему договорить’ (объектная инфинитивная конструкция вторичной предикации) и *Слабое здароўе не дало магчымасці атрымаць сістэматычнай адукацыі* ‘Слабое здоровье не дало возможности получить систематическое образование’ (дополнение с зависимой атрибутивной группой). Подобные примеры могут быть исключены из выборки в ручном режиме при последующем ее анализе. Решение такой проблемы представляется возможным за счет более точной грамматической разметки при создании корпуса.

Следующей задачей при проведении исследования на базе двух языков является установление сопоставимости корпусов. Данный фактор помимо сравниваемого объекта учитывает также «качественные характеристики текстов, образующих исследовательские корпусы, – их временну́ю и жанровую соотнесенность»[1]. Анализ представленных данных показывает, что в случае с рассматриваемыми корпусами сопоставимый исследовательский корпус можно получить, отбирая контексты употребления конструкции вторичной предикации в текстах художественного стиля, поскольку они созданы англоязычными и

[1] Тарасавич Л. А., Пространственные предлоги в немецком и русском языках: семантика и функционирование : дисс. … докт. филол. наук, Минск, МГЛУ, 2016, с. 62.

белорусскими авторами в примерно один и тот же период времени. Выбор других стилей и жанров текстов также возможен, однако они представлены в рассматриваемых корпусах непропорционально, таким образом, требование сопоставимости может быть нарушено.

На следующем этапе определяется объем выборки для проведения исследования. Решение этой задачи может вызывать определенные трудности, поскольку предполагает с одной стороны получение объема данных, достаточного для того, чтобы сделать объективные выводы относительно объекта и его свойств и продемонстрировать достоверность полученных результатов, а с другой – экономию усилий исследователя, поскольку обработка очень большого массива данных может быть достаточно трудозатратной, но не всегда целесообразной. В связи с этим более значимым является не столько объем выборки, сколько репрезентативность корпуса. Причем, если для общеязыкового корпуса параметрами репрезентативности могут считаться его объем и пропорциональная представленность текстов разных периодов, авторов, стилей и жанров[1], то для исследовательского корпуса решающим фактором является не объем, а относительная частота рассматриваемого явления при увеличении выборки. И если она «от прибавления каждого последующего фрагмента текста будет изменяться все меньше и меньше, то это означает, что корпус в целом репрезентативен»[2]. Такую идею определения репрезентативности корпуса можно проиллюстрировать, применив методику пропорциональной организации данных при накоплении и коррекции относительной частоты явления А. Н. Баранова (цит. по 7)[3]. Суть метода продемонстрируем на примере рассматриваемой каузативной конструкции, в которой вторично-предикативный элемент семантически может выражать каузированное действие, состояние или качество. Проведем замер относительной частоты распределения выражения указанных значений в равных фрагментах примеров употребления конструкции по 50 единиц. В первом замере примеры выражения каузируемого действия составляют 66% употреблений, каузированное состояние описывается в 31% контекстов, изменение качества – в 3% примеров. При добавлении в выборку еще 50 примеров употреблений, соотношение контекстов представлено следующим образом: действие – 66%,

---

[1] Богоявленская Ю. В., Репрезентативность лингвистического корпуса: метод верификации достоверности полученных данных // Политическая лингвистика, № 4, 2016, С. 163.

[2] Кибрик А. Е., Брыкина М. М., Леонтьев А. П., Хитров А. Н., Русские посессивные конструкции в свете корпусно-статистического исследования // Вопросы языкознании, 2006, Вып. 1., с. 21.

[3] Баранов А. Н., Проблема репрезентативности корпуса данных (на примере политической метафорики) // Труды Междунар. семинара «Диалог 2001, М., Наука, 2001.

состояние – 30%, качество – 4%. При следующем замере с еще 50 примерами употреблений соотношение остается примерно таким же: 67% – 29% – 4%. Таким образом, при увеличении выборки относительная частота исследуемых явлений в целом сохраняется, что свидетельствует о том, что исследовательский корпус может быть признан репрезентативным уже при объеме в 150 контекстов. Примерно такая же относительная частота распространения характерна и для корпуса английских контекстов, следовательно, для выделения и описания языковых корреляций и закономерностей полученный исследовательский корпус можно признать сопоставимым и репрезентативным.

Таким образом, сопоставительные корпусные исследования синтаксических структур могут проводиться с помощью автоматизированного отбора языкового материала на основе следующего алгоритма: выбор объекта исследования – синтаксической единицы, выражающей одинаковую семантику, определение ее прототипических черт и способов их языковой морфологической реализации для составления поискового запроса в корпусах, элиминация омонимичных грамматических форм в поисковой выдаче, составление сопоставимого по жанрам и времени создания текстов репрезентативного многоязычного исследовательского корпуса, анализ полученных данных и интерпретация результата.

Список использованной литературы

1. Баранов А. Н., Проблема репрезентативности корпуса данных (на примере политической метафорики) // Труды Междунар. семинара «Диалог 2001», М., Наука, 2001.

2. Беларускі N-корпус. URL: https://bnkorpus.info/

3. Богоявленская Ю. В., Репрезентативность лингвистического корпуса: метод верификации достоверности полученных данных // Политическая лингвистика, № 4, 2016, с. 163–166.

4. Кибрик А. Е., Брыкина М. М., Леонтьев А. П., Хитров А. Н., Русские посессивные конструкции в свете корпусно-статистического исследования // Вопросы языкознания, 2006, Вып. 1, с. 16–45.

5. Киклевич А., Корпусные исследования в синтаксисе: возможности и ограничения // Przegląd rusycystyczny. 2022, No. 1 (177), P. 98–114, DOI: 10.31261/pr.12755.

6. Ришар Ж. Ф., Ментальная активность. Понимание, рассуждение, нахождение решений (сокр. пер. с франц. Т. А. Ребеко). М., Институт психологии

РАН, 1998, 232 с.

7. Тарасавич Л. А., Пространственные предлоги в немецком и русском языках: семантика и функционирование : дисс. … докт. филол. наук, Минск, МГЛУ, 2016. 232 с.

8. Чилингарян К. П., Корпусная лингвистика: теория vs методология // Вестник Российского университета дружбы народов. Серия, Теория языка. Семиотика. Семантика. 2021. Т. 12. № 1, с. 196–218. doi: 10.22363/2313-2299-2021-12-1-196-218.

9. British National Corpus (BNC). URL: https://www.english-corpora.org/bnc/

ՇԱՐԱՀՅՈՒՍԱԿԱՆ ՄԻԱՎՈՐՆԵՐԻ ՀԱՄԱԴՐԱԿԱՆ ԿԱՌՈՒՑՎԱԾՔԱՅԻՆ ՈՒՍՈՒՄՆԱՍԻՐՈՒԹՅՈՒՆ /ԱՆԳԼԵՐԵՆ ԵՎ ԲԵԼԱՌՈՒՍԵՐԵՆ ՍՏՈՐՈԳԵԼԻ ԵՐԿՐՈՐԴԱԿԱՆ ՆԱԽԱԴԱՍՈՒԹՅՈՒՆՆԵՐԻ ՀԻՄԱՆ ՎՐԱ/

ՄՈՐՈԶ ԵՎԳԵՆԻ

Մինսկի պետական լեզվաբանական համալսարանի
ժամանակակից թարգմանչական տեխնոլոգիաների
ամբիոնի ավագ դասախոս,
Բելառուսի Հանրապետություն
էլփոստ: mail4maroz@yandex.by

Հոդվածում նկարագրվում են անգլերենում և բելառուսերենում հանդիպող պատճառական կոնտեքստների երկրորդային ստորոգումային կառույցներում օբյեկտային ինֆինիտիվով կազմված շարահյուսական միավորների համեմատական կորպուսային հետազոտության հնարավորություններն ու սահմանափակումները: Շարահյուսական ուսումնասիրություններում լեզվաբանական վերլուծության համար տվյալները ներկայացվում են խոսքի մասերի տերմինաբանությամբ կառույցների լեքսիկական բովանդակությունից անկախ, ինչը թույլ է տալիս եզրակացություններ անել տարբեր լեզվական խմբերի ներկայացուցիչների կողմից իրականության մասին գիտելիքների հայեցակարգային և լեզվական կոդավորման առանձնահատկությունների վերաբերյալ: Գործնական առումով ստացված տվյալները կարող են կիրառվել թարգմանագիտության մեջ՝ մասնավորապես մեքենայական թարգմանության համակարգերի կատարելագործման համար: Օգտագործելով բրիտանական ազգային կորպուսի և բելառուսական N-կորպուսի նյութերը, հոդվածում առանձնացվում են թե՛ համադրելիության գործոնը՝ որպես կորպուսների տեքստերի ժամանակի և ժանրային փոխկապակցվածության պահանջ, թե՛

դրանց ներկայացուցչական գործոնը, որը կախված է ոչ այնքան նմուշի ծավալից, որքան դրա սեգմենտներում դիտարկվող երևույթի հարաբերական հաճախության փոփոխությունից: Քերականական ձևերի համանունությունը կարող է լինել կորպուսային հետազոտությունների համար լեզվական նյութի ավտոմատացված ընտրության տեխնիկական սահմանափակում: Անգլերեն և բելառուսերեն կորպուսների հետ աշխատելու վերը նշված ալգորիթմը կարող է կիրառվել համադրելի ներկայացուցչական կորպուս ստեղծելու համար՝ իրականացնելու ոչ հարակից լեզուների շարահյուսական միավորների ուսումնասիրություն:

Բանալի բառեր՝ *կորպուսային լեզվաբանություն, կորպուս, ներկայացուցչականություն, շարահյուսություն, երկրորդական սփորոգում:*

STUDY OF COMPARATIVE CORPUS OF SYNTACTIC UNITS (A CASE STUDY OF ENGLISH AND BELARUSIAN CONSTRUCTIONS OF SECONDARY PREDICATION)

MAROZ YAUHENI
Senior Lecturer, Department of Modern Translation Techniques,
Minsk State Linguistic University, Belarus
e-mail: mail4maroz@yandex.by

The article aims to explore the peculiarities of studies of comparative corpus of grammatical constructions. It describes the potential and restrictions of these studies through contrastive and quantitative analysis of the objective infinitive construction of secondary predication in causative contexts in English and Belarusian. Corpus studies analyze data in terms of word classes, disregarding the lexical content of the constructions. This allows to draw conclusions about how knowledge is conceptualized and linguistically coded by speakers of different languages. The findings of such studies have practical applications in translation, particularly in improving machine translation systems. The article also emphasizes the factors that allow for data comparison, such as the temporal and genre similarities of the corpus texts and their representativeness. The article uses the British National Corpus and the Belarusian N-corpus as examples, and suggests that the used algorithm can be applied to create comparable representative corpora for studying syntactic units in remotely-related languages.

Keywords*: corpus linguistics, representative corpus, syntax, secondary*

predication.

Հոդվածը ներկայացվել է խմբագրական խորհուրդ 07.09.2023թ.:
Հոդվածը գրախոսվել է 11.09.2023թ.:
Ընդունվել է տպագրության 17.05.2024թ.: