УДК 811.162
DOI: 10.18384/2310-712X-2022-3-2-80-86

ЧЕШСКИЙ НАЦИОНАЛЬНЫЙ КОРПУС КАК ОСНОВНОЙ ИТОГ РАЗВИТИЯ ТРАДИЦИЙ ПРАЖСКОГО ЛИНГВИСТИЧЕСКОГО КРУЖКА В ЧЕШСКОЙ ЛИНГВИСТИКЕ В XX ВЕКЕ

Изотов А. И.
Московский государственный университет имени М. В. Ломоносова
119991, г. Москва, Ленинские горы, д. 1, Российская Федерация

Аннотация
Целью настоящего исследования является рассмотрение Чешского национального корпуса как основного инструмента лингвистического анализа в современной богемистике.
Процедура и методы. Представлен состав Чешского национального корпуса, включающий в себя, помимо 4,9-миллиардной серии SYN (современные чешские письменные тексты), также десятки иных корпусов, как, например, Пражский разговорный корпус, Брненский разговорный корпус, корпус публичных выступлений 14 чешских президентов от Эдуарда Бенеша до Милоша Земана, корпус записи школьных уроков, корпус школьных сочинений детей мигрантов и т. д. Рассмотрены результаты использования корпусного материала в области грамматики, лексикографии и также лингвокультурологии.
Результаты. Показана продуктивность обращения к материалам корпуса в исследованиях в самых разных областях лингвистического исследования.
Теоретическая и/или практическая ценность. Рассмотренный материал и полученные на его основе выводы могут быть использованы для дальнейших исследований в области грамматики, лексикологии и лексикографии, а также лингвокультурологии.

Ключевые слова: Чешский национальный корпус, корпусные исследования, грамматика, лексикография, лингвокультурология

THE CZECH NATIONAL CORPUS AS THE MAIN RESULT OF THE DEVELOPMENT OF THE TRADITIONS OF THE PRAGUE LINGUISTIC CIRCLE IN CZECH LINGUISTICS IN THE XX CENTURY

A. Izotov
Lomonosov Moscow State University
Leninskie Gory 1, Moscow 199991, Russian Federation

Abstract
Aim. The purpose of this study is to consider the Czech National Corpus as the main tool of linguistic analysis in modern Bohemian studies.
Methodology. The composition of the Czech National Corpus is presented. The Czech National Corpus includes, in addition to the 4.9 billion SYN series (modern Czech written texts), dozens of other corpora such as the Prague Colloquial Corpus, the Brno Colloquial Corpus, the Corpus of public speeches of 14 Czech presidents from Eduard Beneš to Miloš Zeman, the Corpus of recorded school lessons, the Corpus of school essays written by migrants etc. The present study considers the results of the use of Corpus material in the field of grammar, lexicography and linguoculturology.

Results. The effectiveness of referring to the corpus materials in research in various fields of linguistics is shown.

Research implications. The material presented in the paper, as well as the conclusions based on its analysis can be used for further research in the field of grammar, lexicology and lexicography, as well as linguoculturology.

Keywords: Czech National Corpus; corpora studies; grammar; lexicography; linguoculturology

Введение

По свидетельству основательницы советской школы богемистики заслуженного профессора МГУ имени М. В. Ломоносова и почётного доктора Карлова университета в Праге А. Г. Широковой, академик Б. Гавранек – один из тех пятерых, которые, встретившись 6 октября 1926 г. в английском семинаре В. Матезиуса на Велеславиновой улице, решили основать Пражский лингвистический кружок, – часто говорил, что «языкознание – наука точная». Имелось в виду, как при необходимости поясняла А. Г. Широкова, что ценность лингвистической работы – не в искусном «плетении словес» и не в рождённых «на кончике пера» оригинальных концепциях, а в убедительности эмпирического материала, в основу данной работы лёгшего. Для чешских членов Пражского лингвистического кружка всегда очень много значила опора на эмпирический материал, в частности на огромную картотеку Института чешского языка Академии наук Чехии, который считает датой своего возникновения образование в 1910 г. «Канцелярии картотеки чешского языка».

1. Чешский национальный корпус как один из наиболее успешных проектов подобного типа в мире

С учётом сказанного неудивителен интерес, проявленный чешскими лингвистами к корпусным исследованиям, ставшим возможными в результате стремительного развития в последние десятилетия XX в. вычислительной техники и интернета, когда появились сравнительно недорогие персональные компьютеры, а также Мировая паутина, способная соединить эти разбросанные по всей стране или даже по всему миру персональные компьютеры с физическим местом хранения корпуса, так что общедоступным стало то, чем ранее могли заниматься единицы.

Разрабатываемый с 1994 г. в Карловом университете Чешский национальный корпус (ЧНК)[1] сегодня бесспорно принадлежит к числу наиболее успешных проектов подобного типа в мире.

1.1. SYN2000 как первый крупный корпус, входящий в состав ЧНК

Уже к 2000 г. для всех заинтересованных лиц был открыт бесплатный онлайн-доступ к первому из входящих в его состав крупных корпусов – к 100-миллионному SYN2000, образованному современными письменными чешскими текстами (количество токенов, включая знаки пунктуации – 120 908 724, количество токенов без знаков пунктуации – 100 061 381, количество несовпадающих словоформ – 1 763 813, количество несовпадающих исходных форм слов – 891 713, количество предложений – 7 639 321, количество целых текстов – 233 797). SYN2000 – корпус **референтный**, то есть его состав уже не редактируется, что позволяет проводить верификацию полученных с его помощью данных, и **репрезентативный**, то есть включает в себя тексты различных функциональных стилей и жанров (в случае с SYN2000 входящие в него тексты были маркированы как Beletrie ‘беллетристика’ с более дробным делением: NOV ‘романы’, COL ‘рассказы и повести’, FAC ‘литература факта’, VER ‘стихи’, SON ‘песни’, SCR ‘драматургия’, IMA ‘иные художе-

[1] Český národní korpus [Электронный ресурс]. URL: https://ucnk.ff.cuni.cz/cs (дата обращения: 15.03.2022).

ственные жанры', как Odborná literatura 'специальная литература' с более дробным делением на SCI 'научные тексты', POP 'научно-популярные тексты', TXB 'учебные тексты'; ENC 'алфавитные или тематические справочники', ADM 'административные тексты'; и, наконец, как Publicistika 'публицистика' с более дробным делением; PUB 'газеты и журналы' MIS 'иные публицистические тексты').

Составители SYN2000 исходили из того, что письменный текст не только отражает языковую действительность, но и формирует её, звуча в сознании читателя всякий раз, когда им прочитывается (вслух или про себя), и тем самым непосредственно влияет на идиолект данного читателя, а через этот идиолект – и на прочие соответствующие идиомы национального языка, а потому включили в его состав то, что чаще всего читалось, в соответствии со специально для этого проведёнными социологическими исследователями, жителями Чешской республики на рубеже тысячелетий. Как оказалось, речь шла прежде всего о публицистике (60%), на втором месте была специальная литература (25%) и лишь на третьем – беллетристика (15%). В этом соотношении публицистические, специальные и художественные тексты и представлены в SYN2000, при этом доля того или иного автора, или той или иной газеты также непосредственно зависит от их популярности.

1.2. Серия корпусов современных письменных текстов SYN, входящая в состав ЧНК

В настоящее время в серию современных письменных текстов SYN входят также четыре стомиллионных референтных репрезентативных корпуса с иным процентным соотношением публицистических, специальных и художественных текстов (SYN2005, SYN2010, SYN2015, SYN2020) и три референтных корпуса публицистических текстов (трёхсотмиллионный SYN2006pub, семисотмиллионный SYN2009pub и девятисотмиллионный SYN2013pub). Поскольку тексты, входящие в названные корпуса, не совпадают, возможно их объединение в один общий корпус с совокупным объёмом **около 4,9 миллиардов** токенов. Все эти корпусы размечены, поэтому поиск в них возможен по словоформе, по лексеме, по грамматической матрице, а также по любой их возможной комбинации. Предусмотрена также возможность объединения в один любой комбинации названных корпусов, а также возможность формирования собственных корпусов из текстов, в данные корпусы входящих. Например, Б. Дичев, создав на базе Чешского национального корпуса свой собственный виртуальный корпус, состоящий из романа Я. Гашека о бравом солдате Швейке, проанализировал с помощью обслуживающего ЧНК программного обеспечения (доступного в полном объёме и в подобных случаях) особенности употребления в данном романе обсценной лексики, см. [2]. В принципе можно создавать подобные виртуальные корпуса для анализа языка самых различных авторов, гендерных и возрастных групп, функциональных стилей, жанров и т. п.

1.3. Иные корпусы, входящие в состав ЧНК

Кроме серии SYN, в состав Чешского национального корпуса входит ещё полсотни корпусов, в том числе нереферентный **6-миллиардный** корпус чешского сектора интернета ONLINE, 2-миллионный корпус школьных сочинений детей мигрантов CzeSl-plain, 2,5-миллионный корпус всех произведений К. Чапека Capek_uplny, миллионный корпус всей публицистики Карла Гавличека Боровского KH-NOVINY, 85-миллионный корпус переводов на чешский язык JEROME, 1,8-миллионный корпус лингвистических научных текстов LINK, 2,1-миллионный референтный репрезентативный корпус неформальной устной чешской речи с двухуровневым (орфо-

графический + фонетический) скриптом ORTOFON, 5,4-миллионный референтный корпус неформальной устной чешской речи на территории Чехии, Моравии и Силезии, 490-тысячный брненский корпус устной речи, включающий в себя записи 90-х годов прошлого столетия BMK, 675-тысячный пражский корпус устной речи PMK, 223-тысячный референтный корпус устной диалектной речи с двухуровневым скриптом DIALEKT, 790-тысячный устный корпус школьных уроков SCHOLA2010, 215-тысячный устный корпус публичных выступлений 14 чешских президентов от Эдуарда Бенеша до Милоша Земана SPEECHES, 38-миллионный устный корпус выступлений в чешском парламенте в 1993–2021 гг. Parlcorp, 3,4-миллионный на сегодняшний день размеченный диахронический корпус DIAKORP, формируемый чешскими текстами с XIV по XX вв., **1,8-миллиардный** в его нынешней версии многоязычный корпус InterCorp с параллельными (оригинальный + переводные) текстами на 42 языках.

Поэтому неудивительно, что корпусный материал используется современными богемистами самым активным образом, ср., например, уже опубликованный первый том подготовленной на корпусном материале «Большой академической грамматики литературного чешского языка» [7].

1.4. Корпус как источник эмпирического материала беспрецедентного объёма

Обращение к электронному корпусу позволяет немыслимым ранее образом ускорить сбор и обработку эмпирического материала. Так, словарь басен Крылова[1], составлявшийся его автором Кимягаровой Р. С. практически в течение всей её жизни, сейчас мог бы быть подготовлен (при наличии в электронном корпусе соответствующих текстов) за пару месяцев не очень напряжённой работы.

1.4.1. Корпус и грамматика

Возможность не в разы, а на порядки ускорить обработку эмпирического материала позволяет оперативно оценивать стремительно меняющуюся языковую ситуацию. Например, в работе О. И. Черчук «Отражение реалий коронавирусного мира в зеркале неологизмов чешского языка» [1] на материале Чешского национального корпуса и поддерживаемой отделением современной лексикологии и лексикографии Института чешского языка Академии наук Чехии электронной базы Neomat было проанализировано функционирование в современном чешском дискурсе 1284 чешских неологизмов, возникших в связи с пандемией КОВИД-19.

Однако дело не только в том, что какие-то исследования при опоре на корпус можно выполнить быстрее, чем без такой опоры. Немыслимый ещё недавно объём эмпирического материала (нетрудно подсчитать, что целой жизни не хватит, чтобы просто прочитать вслух тексты, формирующие названный выше 4,9-миллиардный SYN, ведь при 120 словах в минуту и 16 часах чтения в день без выходных и праздничных дней на это уйдёт более 100 лет) позволяет достоверно проанализировать те явления языковой периферии, описание которых ранее зависело от интуиции исследователя. При этом может случиться, что те или иные общепринятые положения придётся существенно корректировать или даже опровергать. Например, корпусное обследование чешского редкого партиципиального типа на -(v)ší, осуществлённое в работе М. Гигера [3], опровергает как минимум три положения, прочно вошедших в чешскую академическую и школьную традиции. Во-первых, раньше было принято считать, что данные формы, появившиеся в чешском языке в XIX веке

[1] Кимягарова Р. С. Словарь языка басен Крылова. М.: Оникс – Мир и Образование – Русские словари, 2006. 928 с.

под влиянием русских причастий типа *(с)делавший* в результате языкотворческой деятельности патриотически настроенных чешских интеллектуалов, в современном письменном дискурсе если и употребляются, то исключительно в текстах научных (и прежде всего гелертерских), где они служат целям смысловой конденсации дискурса (вместе с иными формами глагольной транспозиции), а также в текстах художественных – с целью стилизации последних под старину (анахронически, поскольку, как мы только что отметили, эти формы в чешском языке появились довольно поздно). Однако М. Гигер обнаружил полторы тысячи (!) контекстов употребления этих форм в корпусах SYN2006pub и SYN-2009pub, образуемых современными публицистическими текстами (где их, как считалось ранее, не должно было быть вообще, ведь это и не научные, и не художественные тексты). Во-вторых, рассматриваемые формы в чешской академической традиции принято обозначать как «адъективированные деепричастия». Исследование М. Гигера показало, что соотносительные с данными формами деепричастия типа (u)dělav, (u)dělavši, (u)dělavše / (při)nes, (při)nesši, (při)nesše встречаются в тех же текстах на порядок реже, а потому не могут на синхронном уровне мотивировать более частотные формы, которые тем самым следует считать производными не от деепричастия, а от инфинитива. В-третьих, ранее считалось, что в современном дискурсе данные образования если и встречаются, то это формы лексикализованные, как, например, přeživší – не просто ‘переживший’, а ‘бывший узник гитлеровского концентрационного лагеря’. Исследование М. Гигера продемонстрировало, что речь идёт хотя и о довольно редком, однако всё же парадигматическом явлении, так как было документировано употребление подобных форм у семи сотен разных глаголов.

1.4.2. Корпус и лексикография

Корпус бесценен для лексиколога и лексикографа, так как он позволяет при составлении словаря исходить не из неких умозрительных соображений, а из реальной употребительности тех или иных лексем в современных текстах. Крупнейший отечественный специалист по сербской лексикологии и лексикографии В. П. Гудков (1934–2020) любил повторять, что если в словаре есть слово «правый», то должно быть и слово «левый». Это звучит вполне логично, однако язык асимметричен и порой языковые феномены (особенно феномены языковой периферии) ведут себя не так, как мы этого от них ожидаем. В частности, в связи с «правым» и «левым» вспомним древне-русские слова, обозначающие «правую руку» (десница) и «левую руку» (шуйца). Первое слово продолжает функционировать в современном русском дискурсе (например, на уроке литературы в школе, когда проходят пушкинского «Пророка»), а второе – нет.

Чешские коллеги, издав в одном томе[1] переработанную чешско-русскую часть классического словаря семидесятых годов[2], радостно исключили из словника всё, что, по их мнению, относилось к ушедшей в прошлое «коммунистической» эпохе, например, слова **menševický** меньшевистский, **předřijnový** дооктябрьский, **trockist||a**, **-y** *m* троцкист *m* и т. д. Однако эти и им подобные «коммунистические» слова продолжают функционировать в современном чешском дискурсе, регулярно воспроизводясь в многочисленных ретро-произведениях о том, как страдали чехи при «тоталитаризме». Поэтому при составлении наших словарей[3]

[1] Velký česko-ruský slovník / M. Sádlíková a kol. Voznice: LEDA, 2005. 1408 s.

[2] Чешско-русский словарь / под редакцией Л. В. Копецкого, Й. Филипца, О. Лешки. В 2 томах. М.: Советская энциклопедия; Прага: Государственное педагогическое издательство, 1976. Т. 1. 580 с.; Т. 2. 861 с.

[3] Изотов А. И. Новый чешско-русский словарь:

мы проверяли каждую словарную статью по корпусу современных текстов, материал которого определял в конечном итоге и состав словника, и количество приводимых словарных значений, и порядок следования этих значений.

Обращение к корпусу позволяет учитывать статистический аспект при составлении не только частотных, но и иных словарей. Постулировав наличие на большей части современной Чешской республики (приблизительно две трети её западной части) языковой ситуации, близкой к классической диглоссии в понимании Ч. Фергюсона, когда в качестве «высокого идиома» выступает литературный чешский язык (spisovná čeština), а в качестве «низкого» идиома так называемый «обиходно-разговорный чешский язык» (obecná čeština), П. Сгалл и Й. Гронек предложили около тысячи соответствий «слово обиходно-разговорного чешского языка» – «слово или словосочетание литературного чешского» [6]. Использование материала серии SYN позволяет определить как абсолютную, так и относительную частотность предлагаемых П. Сгаллом и Й. Гронеком лексических единиц в современном чешском письменном и устном дискурсе, см. [5, с. 111–126].

1.4.3. Корпус и лингвокультурология

Являясь своего рода огромной энциклопедией, хотя и организованной не парадигматически, а синтагматически, корпус может быть весьма полезен и лингвокультурологу. Так, проверив встречаемость в текстах SYN2000, отмеченных в словаре культурной грамотности Э. Д. Хирша[1] (разделы «Литература», «Мифология», «Фольклор»), мы тем самым очертили зоны пересечения культурных компетенций современного американца и современного чеха (см. [4]).

Заключение

Таким образом, создание Чешского национального корпуса с полным правом можно рассматривать как основной итог развития традиций Пражского лингвистического кружка в чешской лингвистике в XX веке, выдвинувший, как и столетие назад, чешскую лингвистику на мировой уровень.

Дата поступления в редакцию 19.04.2022

около 100 000 слов и выражений. М.: Дрофа, 2012. 1024 с.; Изотов А. И. Учебный чешско-русский и русско-чешский словарь: около 40 000 слов и словосочетаний. М.: Филоматис, 2014. 832 с.

[1] Dictionary of Cultural Literacy / E. D. Hirsch, Jr., J. F. Kett, J. Trefil. Boston: Houghton Mifflin Company, 1988. 586 p.

ЛИТЕРАТУРА

1. Черчук О. И. Отражение реалий коронавирусного мира в зеркале неологизмов чешского языка // Вестник Московского университета. Серия 9: Филология. 2021. № 5. С. 71–79.
2. Dichev B. Drsný jazyk Haškova „Švejka“ z pohledu cizojazyčného bohemisty // Nová čeština doma & ve světě. 2012. № 1. S. 33–44.
3. Giger M. Participiální systém češtiny a pozice příčestí minulého činného na -(v)š- v něm // Čmejrková S., Hoffmannová J., Klímová J. Čeština v pohledu synchronním a diachronním. Stoleté kořeny Ústavu pro jazyk český. Praha: Nakladatelství Karolinum, 2012, S. 567–574. DOI: 10.5167/uzh-80476.
4. Izotov A. I. American Cultural Literacy Phenomena in the Mirror of Czech National Corpus: Literature, Mythology, Folklore. Moscow: Azbukovnik, 2010. 200 p.
5. Lingvistika – korpus – empirie / eds. J. Bílková, I. Kolářová, M. Vondráček. Praha: Ústav pro jazyk český AV ČR, v. v. i., 2020. 244 s.
6. Sgall P., Hronek J. Čeština bez příkras. Praha: H&H, 1992. 182 s.
7. Velká akademická gramatika spisovné češtiny. I., Morfologie. Druhy slov, tvoření slov. 2 svazky / Štícha Fr., Kolářová I., Vondráček M., Bozděchová I., Bílková J., Osolsobě K., Kochová P., Opavská Z., Šimandl J., Kopáčková L., Veselý V. Praha: Academia, 2018. 1148 s.

REFERENCES

1. Cherchuk O. I. [Realities of the coronavirus world reflected in the mirror of neologisms in Czech]. In: *Vestnik Moskovskogo universiteta. Seriya 9: Filologiya* [Moscow University Philology Bulletin], 2021, no. 5, pp. 71–79.
2. Dichev B. Drsný jazyk Haškova „Švejka" z pohledu cizojazyčného bohemisty. In: *Nová čeština doma & ve světě*, 2012, no. 1, S. 33–44.
3. Giger M. Participiální systém češtiny a pozice příčestí minulého činného na -(v)š- v něm. In: Čmejrková S., Hoffmannová J., Klímová J. *Čeština v pohledu synchronním a diachronním. Stoleté kořeny Ústavu pro jazyk český*. Praha, Nakladatelství Karolinum, 2012, S. 567–574. DOI: 10.5167/uzh-80476.
4. Izotov A. I. American Cultural Literacy Phenomena in the Mirror of Czech National Corpus: Literature, Mythology, Folklore. Moscow, Azbukovnik Publ., 2010. 200 p.
5. Bílková J., Kolářová I., Vondráček M., eds. Lingvistika – korpus – empirie. Praha, Ústav pro jazyk český AV ČR, v. v. i., 2020. 244 s.
6. Sgall P., Hronek J. Čeština bez příkras. Praha, H&H, 1992. 182 s.
7. Štícha Fr., Kolářová I., Vondráček M., Bozděchová I., Bílková J., Osolsobě K., Kochová P., Opavská Z., Šimandl J., Kopáčková L., Veselý V. Velká akademická gramatika spisovné češtiny. I., Morfologie. Druhy slov, tvoření slov. 2 svazky. Praha, Academia, 2018. 1148 s.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Изотов Андрей Иванович – доктор филологических наук, профессор кафедры славянской филологии Московского государственного университета имени М. В. Ломоносова;
e-mail: a.i.izotov@mail.ru);

INFORMATION ABOUT THE AUTHOR

Andrey I. Izotov – Dr. Sci. (Philology), Prof., Department of Slavic Philology, Lomonosov Moscow State University;
e-mail: a.i.izotov@mail.ru

ПРАВИЛЬНАЯ ССЫЛКА НА СТАТЬЮ

Изотов А. И. Чешский национальный корпус как основной итог развития традиций пражского лингвистического кружка в чешской лингвистике в XX веке // Вестник Московского государственного областного университета. Серия: Лингвистика. 2022. № 3. Т. 2. С. 80–86.
DOI: 10.18384/2310-712X-2022-3-2-80-86

FOR CITATION

Izotov A. I. The Czech national corpus as the main result of the development of the traditions of the Prague linguistic circle in Czech linguistics in the XX century. In: *Bulletin of the Moscow Region State University. Series: Linguistics*, 2022, no. 3, vol. 2, pp. 80–86.
DOI: 10.18384/2310-712X-2022-3-2-80-86