А.И.Изотов
Московский государственный университет имени М.В.Ломоносова

Опыт корпусного анализа лексики чешского общенационального субстандарта (obecná čeština)

корпусные исследования, общенациональный субстандарт, обиходно-разговорный чешский язык, лексикография

Характерной особенностью языковой ситуации в современной Чехии, отличающей её, в частности, от ситуации отечественной, является наличие общенационального субстандарта, активно употребляемого большинством (если не всеми) носителями языка, см. [Изотов 2017], там же дальнейшая литература. Новые возможности, предоставленные исследователям «корпусной революцией», позволили доказательно рассмотреть также и такие аспекты проблематики, которые ранее решались в основном практически исключительно на основе интуиции.

Авторы написанной ещё в «докорпусную эпоху» монографии [Sgall 1992] предложили лингвистической общественности пробный вариант словаря наиболее важной, по их мнению, чешской субстандартной лексики (около 1000 единиц), снабдив каждую подобную лексическую единицу литературным эквивалентом (либо толкованием), а также пометой: *běž* – лексическая единица, общая и для чешского общенационального субстандарта, и для литературного чешского языка, *ob* – лексическая единица чешского общенационального субстандарта, не маркированная в качестве стилистически сниженной, *n* – лексическая единица чешского общенационального субстандарта, маркированная как стилистически сниженная, *n.sl.* [букв. ‘низкие сленги’] – лексическая единица, характерная для нелитературной речи солдат, ремесленников и т.д., *v.sl.* [букв. ‘высокие сленги’] – лексическая единица, характерная для нелитературной речи врачей, инженеров, студентов и т.д., *expr* экспрессивная лексическая единица, *pejor* пейоративная лексическая единица, *zast* устаревающая лексическая единица, *zhrub* грубая лексическая единица, ср. начало данного словаря (с нашим переводом на русский язык в квадратных скобках):

Slovo [‘слово’]	Ekvivalent [‘эквивалент’]	Vrstva [‘идиом’]	Styl [‘стиль’]
áčko	písmeno A [‘буква А’]	běž	
ahoj	nazdar [‘привет!’]	běž	
ajnclík	jednolůžkový pokoj [‘одноместный номер’]	n.sl.	
ajznboňák	železničář [‘железнодорожник’]	ob	zast

Мы попытались выявить употребительность представленных в данном словарике лексем, а также их литературных эквивалентов в материале входящего в состав Чешского национального корпуса набора подкорпусов современных письменных текстов SYN2000, SYN2005, SYN2006pub, SYN2009pub, SYN2010 общим объемом 1,3 миллиарда токенов.

Поиск однословных лексических единиц осуществлялся с помощью простых запросов типа [lemma="ajznboňák"], находящих контексты употребления заданной лексемы (в данном случае лексемы ajznboňák ‘железнодорожник’) во всех её словоформах. Поиск не по лексеме, а по словоформе, который в неизмеримо меньшей степени зависит от надежности стохастического анализа и основанной на нём автоматической разметки текстов корпуса, был бы более точен, однако он же и неизмеримо более трудоёмок, так как предполагает поиск всех теоретически возможных словоформ заданной лексемы ([word="[Aa]jznboňák] + [word="[Aa]jznboňáka] + [word="[Aa]jznboňákovi] + …) с последующим суммированием результатов либо поиск типа [word="[Aa]jznboňá.*"], далее с визуальным анализом обнаруженных контекстов

и удалением многочисленных дериватов и омографов, не относящихся к данной словоформе.

Для поиска неоднословных лексических эквивалентов или же толкований приходилось использовать более сложные запросы, представляющие собой поиск по той или иной комбинации лексемы (lemma), словоформы (word) и грамматической матрицы (tag) в пределах одного предложения (within <s id=".*"/>). Например, для поиска контекстов употребления лексического толкования *být potrestán za něco* ‘быть наказанным за что-либо’ использовался запрос ([lemma="být"] [(lemma="potrestat") & (tag="Vs.*")] [word="za"]) | ([lemma="být"] [word="za"] []{0,3} [(lemma="potrestat") & (tag="Vs.*")]) within <s id=".*"/>.

Тем самым определение частотности употребления входящих в однословные соответствия «лексема чешского общенационального субстандарта» – «лексема чешского литературного языка с той же семантикой» было наиболее простым и в то же время наиболее полезным для целей практического преподавания чешского языка как иностранного, так как оно позволяет установить иерархию приоритетов при знакомстве обучаемого с чешским общенациональным субстандартом. О том, почему обучение чешскому общенациональному субстандарту следует начинать с фонетики и лексики, а не с морфологии, см. [Hrdlička 2010: 82].

Мы дополнили таблицу из [Sgall 1992: 139-169] столбцами с полученными данными о числе обнаруженных нами контекстов употребления соответствующих лексических единиц, столбцы «Идиом» и «Стиль» разместили непосредственно после столбца «Слово» и добавили столбец «Лексическое соответствие в русском». Временно удалив из таблицы строки с дву- и многословными эквивалентами (такими как приведенные выше áčko = písmeno A [‘буква A’] или ajnclík = jednolůžkový pokoj [‘одноместный номер’]), а также строки, требующие дополнительного визуального контроля контекстов (случаи с различного рода омонимией или, наоборот, синонимией), мы получили таблицу однословных соответствий «лексическая единица чешского общенационального субстандарта» – «лексическая единица чешского литературного языка», начало которой (в порядке убывания документированной частотности субстандартных лексических единиц) выглядит следующим образом:

Слово	Идиом	Стиль	Число найденных контекстов	Эквивалент	Число найденных контекстов	Лексическое соответствие в русском
gól	*Běž*		400702	branka	450042	гол
šéf	*Běž*		359917	vedoucí	372890	заведующий
jes(t)li	*Běž*		289763	jestliže	109327	если
taky	*Běž*		189554	také	2278504	тоже
pořád	*Ob*		181478	stále	697389	постоянно
fanoušek	*Běž*		155431	přívrženec	8207	сторонник
kluk	*Běž*		151781	chlapec	138205	мальчик

Данная таблица, представленная полностью в [Изотов 2014], включает 537 однословных соответствий и позволяет определять (после перенесения её данных в Excel) не только абсолютную, но и относительную частоту употреблений как отдельной лексемы, так и сгруппированных по тому или иному принципу типов рассматриваемых лексем. Например, если мы примем общее количество зафиксированных употреблений той или иной субстандартной лексемы (A) и её литературного коррелята (B) за 100%, то относительная частотность данной субстандартной лексемы будет равняться A/(A + B) × 100%.

Мы видим, что приведенные выше лексические единицы в начале таблицы характеризуются не только высокой абсолютной, но и высокой относительной частот-

ностью употребления: для gól – 47,10% (с точностью до второго знака после запятой), šéf – 49,11%, jes(t)li – 72,60%, taky – 7,68%, pořád – 20,64%, fanoušek – 94,98%, kluk – 52,34%.

Это тем более удивительно, что в соответствии с данной таблицей и данной формулой расчета (А – сумма документированных контекстов употребления всех лексем из первой колонки; В – сумма документированных контекстов употребления всех их литературных коррелятов) средняя относительная частотность лексемы из первой колонки была на порядок ниже и составляла приблизительно 5,2%.

Средняя относительная частотность лексических единиц, снабжённых пометой *ob*, то есть характерных исключительно для чешского общенационального субстандарта, однако не маркированных в качестве стилистически сниженных, в три раза ниже относительной частотности лексических единиц, снабженных пометой *běž,* то есть общих для и для субстандарта, и для литературного языка: речь идет о 2,5% и о 7,5% соответственно. Если же мы будем учитывать также и лексические единицы чешского общенационального субстандарта, маркированные как стилистически сниженные (помета *ob* + помета *n*), то эта относительная частотность будет еще ниже – 2,1%.

Что же касается стилистических помет, то весьма высокой оказалась относительная частотность лексических единиц, маркированных как *грубые* – 9,1%, тогда как у слов, маркированных как *пейоративные, экспрессивные* или *устаревшие*, относительная частотность оказалась существенно ниже средней – 1,79%, 0,95% и 0,84% соответственно.

Полученный нами частотный словарь однословных соответствий чешской субстандартной и литературной лексики не претендует на то, чтобы быть исчерпывающим уже в силу того, что на это не претендует список лексем, изначально предложенный в [Sgall 1992]. Тем не менее употребительность в современных чешских письменных текстах (а во многих случаях – весьма высокая употребительность) тех лексических единиц чешского общенационального субстандарта, которые в нем представлены, следует считать эмпирически доказанной, что безусловно должно быть учтено как минимум в лексикологии современного чешского языка, стилистике, практической лексикографии и лингводидактике.

Литература

Изотов А.И. Корпусная революция: от «искусства» к «науке» / А.И.Изотов // Филологические науки: Вопросы теории и практики. – 2013. – № 4-1. – С. 68-71.

Изотов А.И. Эскиз частотного словаря ядра обиходно-разговорной чешской лексики / А.И.Изотов // Язык, сознание, коммуникация. Вып. 48. – М.: МАКС Пресс, 2014. – С. 5-25.

Изотов А.И. К вопросу об общенациональном субстандарте в чешском и русском языковых пространствах / А.И.Изотов // Филологические науки. Вопросы теории и практики. – 2017. – № 2-2. – С.131-134.

Sgall P. Čeština bez příkras / P.Sgall, J.Hronek. – Praha: H&H, 1992. – 182 s.

Hrdlička M. Kapitoly o češtině jako cizím jazyce / M.Hrdlička. – Plzeň: Západočeská univerzita v Plzni, 2010. – 248 s.