„Bohemistyka” 2021, nr 4, ISSN 1642–9893

Milena HEBAL-JEZIERSKA
Uniwersytet Warszawski

DOI: 10.14746/bo.2021.4.8

Terminologia lingwistyki korpusowej w języku czeskim, polskim i słowackim – wybrane problemy

Keywords: terminology, corpus linguistics, Czech language, Slovak language, Polish language

Słowa kluczowe: terminologia, terminologia korpusowa, język czeski, język słowacki, język polski

Abstract

The source for the terminology of corpus linguistics in West Slavic languages is undoubtedly English. In each of the languages studied here, we deal with both the process of acquiring loans denoting terms used in corpus linguistics, and attempts to create terminology based on native vocabulary. The meaning of the terms is either consistent with the original meaning or modified. In some cases, there is a divergence of meaning not only on the inter-linguistic level but also within one language. The explanation of this phenomenon is not related to the nature of corpus linguistics, but to the specificity of terminology.

Źródłem terminologii językoznawstwa korpusowego w językach zachodniosłowiańskich jest niewątpliwie język angielski. W każdym z omawianych tu języków mamy do czynienia zarówno z procesem pozyskiwania zapożyczeń oznaczających terminy używane w językoznawstwie korpusowym, jak i próbami tworzenia terminologii opartej na słownictwie rodzimym. Znaczenie terminów jest albo zgodne z pierwotnym znaczeniem, albo zmodyfikowane. W niektórych przypadkach istnieje rozbieżność znaczeń nie tylko na poziomie międzyjęzykowym, ale także w obrębie jednego języka. Wyjaśnienie tego zjawiska nie jest związane z naturą językoznawstwa korpusowego, ale ze specyfiką terminologii.

Celem artykułu jest przyjrzenie się terminom stosowanym w lingwistyce korpusowej w języku czeskim, słowackim i polskim. Są one obecne stosunkowo od niedawna w językach zachodniosłowiańskich.

Ich występowanie w języku jest z pewnością jednym z przykładów odzwierciedlenia postępu technologicznego[1] w leksyce. Lingwistyka korpusowa, choć obecna w świecie zachodnim od lat 60., w krajach zachodniosłowiańskich swoje początki miała dopiero pod koniec lat dziewięćdziesiątych XX wieku[2]. Nowa technologia, nowy paradygmat badawczy[3], nowe terminy mające swoje źródło w oryginalnej lingwistyce korpusowej w różny sposób zadomowiły się w języku czeskim, słowackim i polskim. W tym tekście chciałabym omówić zarówno ich strukturę, poziom dostosowania form zapożyczonych w językach docelowych, jak i rozbieżności terminologiczne. Wydawać by się mogło, że zauważone różnice terminologiczne nie tylko międzyjęzykowe, ale również mające miejsce w obrębie jednego języka, są związane z młodością dziedziny, w której występują. Tymczasem okazuje się, że rozbieżności terminologiczne dotykają też dyscyplin lingwistyki, które na dobre zadomowiły się w danym kraju. Przykładem może być teoria zapożyczeń mająca prawie stuletnią tradycję w polskim językoznawstwie. Zwróciła uwagę na to Alicja Witalisz (2012, s. 107–114). Ponadto jej spostrzeżenia są w niektórych as-

[1] Inne przykłady terminów będące egzemplifikacją postępu technologicznego podaje Zabawa 2017.

[2] Pierwszy elektroniczny korpus językowy powstał 1964 r. Był to tzw. Brown Corpus, autorstwa Francisa i Kučery. W latach 80-tych nastąpiła ekspansja zachodnich korpusów językowych: BNC (Brytyjski Korpus Narodowy) i rozwój lingwistyki korpusowej (Grabowski, Hebal-Jezierska 2016). Korpusy języków słowiańskich czekały dłużej na swoją kolej. Ich powstawanie przypadło na koniec XX oraz początek XXI w. Pierwszy był korpus języka słoweńskiego (1999 r.), drugi – korpus języka czeskiego (2000 r.), następnie – korpus języka słowackiego (2002 r.), korpus języka rosyjskiego (2004 r.), korpus języka polskiego (2007–2010 r.), korpus języka dolnołużyckiego (2009 r.), korpus języka górnołużyckiego (2009 r.), korpus języka chorwackiego (2005 r.), korpus języka serbskiego, korpus języka bułgarskiego (2009 r.), korpus języka ukraińskiego (2011 r.), korpus języka białoruskiego, korpus języka macedońskiego (Hebal-Jezierska 2014).

[3] Lingwistyka korpusowa w zależności od jej rozwoju w danym kraju jest różnie postrzegana. Może być widziana tylko jako nowa technologia (budowanie korpusów) lub również jako paradygmat badawczy (por. Čermák 2017).

pektach uniwersalne, dotyczą zarówno młodych, jak i starych dyscyplin nauki.

1. Źródła oryginalnej terminologii lingwistyki korpusowej

Oryginalna terminologia lingwistyki korpusowej powstała w języku angielskim. Oprócz terminów właściwych lingwistyce korpusowej (np. terminy oznaczające rodzaje korpusów) jest w niej mnóstwo terminów zapożyczonych z innych dziedzin, takich jak: informatyka (np. *tag*, *token*), matematyka, szczególnie statystyka (nazwy testów statystycznych, np.: *informacja wzajemna*, *logarytm Dice'a*), filologia (nazwy części mowy itp., gatunków literackich), socjologia. Warto podkreślić, że w wielu terminach występujących w oryginalnej literaturze przedmiotu dochodzi do rozbieżności definicyjnych, np. *corpus-based* 'corpus-based', *corpus-driven* 'corpus-driven', *semantic prosody* 'prozodia semantyczna'. W przypadku pierwszych dwóch terminów ze względu na ich niejednoznaczność związaną właśnie z chaosem terminologicznym podjęto próbę zastąpienia ich nowym podziałem *corpus-as-theory* i *corpus-as-method*. (McEnery, Hardie 2002, s. 150). Natomiast w przypadku terminu *semantic prosody* mamy do czynienia z tak dużą liczbą definicji (por. Stewart 2013), że nie pozwala to na jego ujednoznacznienie.

W języku angielskim powstało mnóstwo publikacji dotyczących lingwistyki korpusowej. W swoim artykule za podstawę wykorzystaną w analizie uznałam słownik *A Glossary of Corpus Linguistics* oraz *Corpus Linguistics – Some Key Terms.* Materiał ten uzupełniłam innymi publikacjami anglojęzycznymi. Badanie terminologii lingwistyki korpusowej w językach zachodniosłowiańskich oparłam natomiast na analizie publikacji leksykograficznych i encyklopedycznych (*Výberový slovník termínov z korpusovej lingvistiky*, *Nový encyklopedický slovník češtiny*) oraz publikacjach dotyczących lingwistyki korpusowej języków słowiańskich, np. *Podstawy językoznawstwa korpusowego*. Ponadto posłużyłam się kwerendą internetową oraz literaturą przedmiotu**.**

2. Terminologia lingwistyki korpusowej w językach zachodniosłowiańskich

Terminy lingwistyki korpusowej, jak już wspomniałam, mają swoje źródło w lingwistyce angielskiej. Możemy zatem wyróżnić tu następujące procesy dotyczące terminologii:

– przyswajanie zapożyczeń z języka angielskiego w językach zachodniosłowiańskich,
– próby tworzenia własnego nazewnictwa na bazie słownictwa rodzimego,
– definiowanie znaczenia w języku docelowym zgodnie ze znaczeniem oryginalnym,
– definiowanie znaczeń na podstawie wieloznaczności leksemów wyrażających termin,
– definiowanie znaczeń na podstawie istniejącej tradycji językoznawczej w języku docelowym,
– definiowanie terminu w języku docelowym na podstawie deleksykalizacji wyrazów wyrażających termin.

2.1. Struktura terminów

Struktura terminów stosowanych w lingwistyce korpusowej w językach zachodniosłowiańskich jest zróżnicowana. Większość stanowią zapożyczenia z angielszczyzny, w tym zapożyczenia właściwie, cytaty, hybrydy, neosemantyzmy. Część terminów powstała na bazie leksyki rodzimej. W użyciu w zależności od języka oraz od terminu są bądź wyłącznie struktury obce/rodzime, bądź obydwa warianty (por. tab. 1). Większość zapożyczeń wyrażających terminy stosowane w lingwistyce korpusowej uległa adaptacji fonetycznej, ortograficznej i morfologicznej.

2.1.1. Zapożyczenia z języka angielskiego o różnym o stopniu adaptacji

Wśród terminów o strukturze zapożyczeń można wyróżnić leksemy niezaadaptowane do języków lub zaadaptowane całkowicie lub

częściowo pod względem ortograficznym, fonetycznym, morfologicznym. Do terminów niezaadoptowanych, będących właściwie cytatami, należą rodzaje podejścia korpusowe, np. *corpus-illustrated*, *corpus-based*, *corpus-driven*. Wiąże się to z tym, że metodologia lingwistyki korpusowej wciąż jest obca w krajach zachodniosłowiańskich. Trudnością jest również znalezienie zgrabnego ekwiwalentu. Próba stworzenia czeskiego odpowiednika połączeń angielskich *corpus-based*, *corpus-driven* nie powiodła się (patrz dalej). We wszystkich trzech językach te połączenia językowe pozostają w jednej formie fleksyjnej. Jedyne, co je różni od angielskiego oryginału, to wymiana leksemu *approach*, będącego integralną częścią terminu na słowiański odpowiednik lub opis, np. cz. *corpus-based přístup*, *corpus-driven přístup*, pl. *podejście/sposób analizy corpus-based*, słc. *corpus-based prístup*, *corpus-driven prístup*.

Jedním z termínů, jehož vymezení je důležité jak pro korpusový výzkum obecně, tak pro kontrastivní lingvistiku a translatologii, je anglické spojení corpus-based (approach, discipline atd.) (Chlumská 2014, s. 222).

Problematyczną kwestią dotyczącą metodologicznych aspektów językoznawstwa korpusowego jest różnica pomiędzy podejściem *corpus-based* a *corpus-driven* w badaniu danych językowych (Zasina 2018, s. 170).

Z innych terminów funkcjonujących w językach zachodniosłowiańskich jako cytaty, należy wymienić leksemy: *lockword/lockwords*, *keyword/keywords*. Wymienione terminy są częścią metodologii badawczej stosowanej w kwantytatywnej w zachodniej lingwistyce korpusowej. Metoda zastosowania tzw. *lockwords* jest dość kontrowersyjna ze względu na zastosowanie w niej frekwencji całkowitej. W krajach zachodniosłowiańskich jest mało znana, sporadycznie stosowana. Termin ten jest używany raczej przy referowaniu badań zachodnich badaczy. Równie rzadko stosowany jest angielski leksem *keywords*, używany między innymi w metodzie słów kluczowych (patrz dalej). Ze względu na wieloznaczność odpowiedników słowiańskich: *słowo kluczowe*, *klíčové slovo*, *kľúčové slovo*, wymienione leksemy podaje się często w nawiasie jako uściślenie, w jakim znaczeniu tu występują.

KWords slouží k identifikaci klíčových slov (tzv. keywords): to jsou slova (resp. slovní tvary), která jsou úzce spojena s hlavními tématy textu a s jeho žánrem (https://kwords.korpus.cz).

Z leksemów, które zostały zapożyczone w takiej samej postaci graficznej, ale uległy adaptacji fleksyjnej, można wymienić: *tag* (cz., pl., słc.), *tagset* (cz., pl., słc.), *token* (cz., pl., słc.). Odmieniają się we wszystkich trzech językach według tych samych paradygmatów, do których należą wyrazy rodzime. Leksemy te uległy również częściowej adaptacji fonetycznej. Angielska głoska [ae] obecna w słowach *tag* i *tagset* jest adaptowana jako [a] w języku polskim lub [e] w języku czeskim i słowackim.

Następne terminy zaadaptowane zarówno morfologicznie, fonetycznie i ortograficznie, to leksemy, które do rdzenia obcego mają dołączone morfemy języka docelowego. Do takich wyrazów należą we wszystkich językach odpowiedniki angielskich terminów: *lemmatisation* ‘lematyzacja’, *anotation* ‘anotacja’, *tagging* ‘tagowanie’, por.:

– ang. *annotation*, cz. *anotace*, pl. *anotacja*, słc. *anotácia*;
– ang. *lemmatisation*, cz. *lemmatizace*, pl. *lematyzacja*, słc. *lematizácia*;
– ang. *tagging*, cz. *tagování*, pl. *tagowanie*, słc. *tagovanie*.

Na oddzielną wzmiankę zasługuje adaptacja ekwiwalentów angielskiego terminu *lemma*. Jest to wyraz pochodzenia greckiego zaadaptowany z łaciny do języka angielskiego. W językach zachodniosłowiańskich zaadaptowały się całkowicie analogicznie do leksemu *temat*, czyli w języku polskim wraz z przyrostkiem tematycznym *lemat*[4], natomiast w języku czeskim i słowackim bez niego: *lema*, *lemma*.

Z pożyczek pozostaje jeszcze wspomnieć o połączeniach językowych, które są dosłownymi tłumaczeniami angielskich wyrażeń i zwrotów, np. pl. *preferencja semantyczna*, *prozodia semantyczna*, cz.

[4] W niektórych źródłach internetowych (https://pl.wikipedia.org/wiki/Forma_słownikowa) oraz w niektórych publikacjach (Tkaczewski 2013) zdarzają się wystąpienia formy *lemma*. Poświadczenia te mają źródło w języku czeskim.

sémantická preference, *sémantická prozodie*. Są one niezrozumiałe dla osób spoza kręgu lingwistów korpusowych.

2.1.2. Odpowiedniki rodzime zapożyczonych leksemów

W każdym z badanych języków nastąpiła próba zastąpienia niektórych pożyczek leksemami rodzimymi (patrz tab. 1).

Tabela 1. Wybrane ekwiwalenty angielskich terminów stosowane w języku czeskim, polskim i słowackim.

Termin w języku angielskim	**Termin w języku czeskim**	**Termin w języku polskim**	**Termin w języku słowackim**
annotation/ tagging	značkování/anotace/tagování	znakowanie/anotacja /tagowanie	značkovanie/anotácia/tagovanie
corpus-based approach	korpusem ověřovaný výzkum, na korpusu založený výzkum	corpus-based	Brak danych
corpus-driven approach	**korpusem inspirovaný výzkum, korpusem řízený výzkum**	corpus-driven	brak danych
lemma	lemma	hasło/lemat	Lema
lemmatisation	lematizace	hasłowanie/lematyzacja	lematizácia
Tag	značka/tag	znacznik/tag	značka/tag
Token	token	okaz/token/segment	Token

We wszystkich językach zachodniosłowiańskich istnieją formy obce i rodzime dla angielskich terminów: *annotation*, *tag*. Ponadto w języku polskim wykorzystano leksem rodzimy *hasłowanie* w znaczeniu ‘lematyzacja’ oraz *okaz* – w znaczeniu ‘token’. Natomiast w języku czeskim nastąpiła próba utworzenia rodzimych odpowiedników angielskich wyrażeń *corpus-based* i *corpus-driven*. Są one pod względem ekonomii języka niezbyt wydajne. Pełnią raczej funkcję tłumaczenia terminu angielskiego niż samodzielnej jednostki terminologicznej, np.

Na jedné straně stojí přístup, v němž se postupuje od introspektivně vybudované hypotézy směrem k jejímu ověřování na rozsáhlých datech, tzv. přístup corpus-based, tedy „na korpusu založený. Do protikladu k němu bývá dáván tzv. přístup corpus-dri-

ven, „korpusem řízený“, který označuje postup, v němž sice badatel vychází od určité své hypotézy či představy (což je ostatně nevyhnutelné u všech typů výzkumu), je ovšem připraven ji na základě dat zcela přeformulovat tak, aby odpovídala reálné situaci; data zde tedy hrají skutečně klíčovou roli (Chlumská 2014, s. 222).

Warto podkreślić, że również inne leksemy często występują wymiennie, tłumacząc się wzajemnie poprzez dodanie odpowiednik rodzimego lub obcego, np.

Kolejná různicą między tagsetami jest lematyzacja (hasłowanie) – w NKJP formą hasłową skrótu jest jego rozwinięcie (Przepiórkowski, Bańko, Górski, Lewandowska-Tomaszczyk 2012, s. 68).

Morfologické značky (tagy) jsou součástí výsledku (výstupem) morfologické analýzy, která pracuje s izolovanými slovními tvary, tedy bez ohledu na jejich kontext (Chlumská 2017).

2.1.2.1. Token, segment czy okaz?

Ciekawy problem dotyczący ekwiwalentu ang. *token* w przeciwieństwie do języka czeskiego i słowackiego pojawił się w języku polskim. Możemy tu mówić o konkurencji obcego leksemu *token* i rodzimego wyrazu *okaz*. Mam tu na myśli termin odpowiadający angielskiej definicji:

Token: A single linguistic unit, most often a word, although depending on the encoding system being used, a single word can be split into more than one token, for example he’s (He + ’s).’ (Hardie, MacEnery, Baker 2006, sincerely. 159).

W ten sposób jest również definiowany termin *token*, a co za tym idzie i hasło *tokenizacja* w języku czeskim i słowackim. Stosowany jest w nich tylko jeden leksem, właśnie *token*.

Token – Nejmenší jednotka textu, většinou grafické slovo, resp. jedna jeho realizace (type-token). V korpusové lingvistice je v některých případech jedno grafické slovo rozděleno na dvě slova (např. *mohu -li*), často je také z praktických důvodů (pro snadné vyhledávání) oddělována interpunkce od předcházejícího či následujícího slova (3 tokeny: *řekl , že*). O jednotlivých **t.** v korpusu se také mluví jako o pozicích. – Velikost korpusu se udává v t.n. také v textových slovech. Rozčlenění textu na **t.** je výsledkem procesu tokenizace (Cvrček 2017).

Ani w języku czeskim, ani w słowackim nie używa się innego leksemu niż *token*.

W języku polskim mamy inną sytuację. W użyciu są trzy leksemy: *token*, *segment* oraz *okaz*. Twórcy Narodowego Korpusu Języka Polskiego nazywają *segmentacją* na poziomie słów proces tokenizacji, a jego wynik *segmentem*. Część badaczy stosuje leksem *token* lub *okaz*, lub obydwa.

Wygenerowane listy frekwencyjne mogą ponadto przedstawiać procentowy udział poszczególnych typów w stosunku do objętości całego korpusu, czyli wszystkich wyrazów w tekście (tzw. okazów, ang. tokens) (Łukasik 2008, s. 34).

Przykładowo zasób CommonCrawl6, zawierający dane całej domeny .ru tylko za lata 2012 oraz 2013, liczy (hipotetycznie) około 500 mld tokenów (okazów) języka rosyjskiego (Fedorushkov 2018, s. 56).

A zatem jedynym językiem, w którym powstał odpowiednik o strukturze rodzimej, jest język polski. Termin *okaz* ma długą tradycję w polskiej lingwistyce. Jego początki są związane z teorią znaku Peirce'a, w której jest on odpowiednikiem angielskiej opozycji *type-token*, tłumaczonej na język polski *typ-okaz* (por. Saloni 1993). Część językoznawców uznało zatem, że opozycja *type-token* pojawiająca się w lingwistyce korpusowej powinna mieć taki sam ekwiwalent, czyli *typ-okaz*.

Takiej sytuacji nie ma w języku czeskim i słowackim. Owszem w historii czeszczyzny wraz z wprowadzeniem teorii Peirce'a nastąpiły próby zastąpienia angielskiego wyrazu *token* przez leksem *exemplář* (Palek 1969). Nie zakorzeniły się jednak w językoznawstwie czeskim. Nie zostały zatem również zaproponowane jako ekwiwalenty rodzime w lingwistyce korpusowej.

Próba zastąpienia leksemu *token* przez leksem *okaz* wydaje się jednak dość kontrowersyjna. Brakuje tu leksemu, który byłby derywatem pochodzącym od leksemu *okaz* i oznaczałby proces tokenizacji, co ma miejsce w innych przypadkach *token* jako wynik tokenizacji, *segment* jako rezultat segmentacji.

2.2. Definiowanie terminów – wybrane przykłady

Terminologia lingwistyki korpusowej stosowana w językach zachodniosłowiańskich nie zawsze odzwierciedla oryginalne znaczenia terminów. Przyczyn należy szukać częściowo w chaosie terminologicznym, który cechuje również oryginalną literaturę przedmiotu. Przykładami mogą być wspomniane już terminy *corpus-based* i *corpus-driven*. Dosłowne (każde badanie wykorzystujące dane korpusowe), a nie terminologiczne (jako metody badawczej) traktowanie połączenia językowego *corpus-based* znajdziemy zarówno w angielskiej literaturze przedmiotu, jak i niektórych publikacjach wydanych w krajach zachodniosłowiańskich. Rozbieżności w stosowaniu terminu mogą mieć również źródło w zderzeniu tradycji definicyjnych (np. *token*), niepoprawnym rozumieniu terminu (*korpus referencyjny*, *KWIC*), polisemii (*konkordancja*, *słowo kluczowe*), wielodefinicyjności (*prozodia semantyczna*), jak również po prostu niewiedzy.

2.2.1. Dosłowne traktowanie terminu jako źródło rozbieżności terminologicznych

Przykładami definicji w pełni nieodzwierciadlających oryginalnych eksplikacji są czeskie terminy *KWIC* i *referenční korpus* 'korpus referencyjny'. Przyjrzyjmy się pierwszemu z wymienionych terminów. Definicja znajdująca się w angielskich źródłach (CASS 2013) odnosi *KWIC* do konkordancji, rozumianej jako lista wystąpień jednostki wyszukiwanej występującej w kontekście.

A way of displaying a node word or search term in relation to its context within a text. This usually means the node is displayed centrally in a table with co-text displayed in columns to its left and right. Here, 'key word' means 'search term' and is distinguished from keyword (CASS 2013, s. 6).

W takim znaczeniu jest również podawany termin *KWIC* w języku polskim i słowackim:

Najpopularniejsza forma konkordancji, w której wyszukiwany wyraz jest przedstawiony w otaczającym kontekście z lewej i prawej strony na środku generowanej listy (Lewandowska-Tomaszczyk 2005, s. 297).

Zobrazenie výsledkov hľadania, pri ktorých je hľadané kľúčové slovo al. reťazec znakov zobrazené zvýraznené uprostred kontextu (Šimková 2006, s. 3).

Od tych definicji odstaje definicja czeska umieszczona zarówno w słowniku terminów lingwistyki korpusowej znajdującym na stronach Czeskiego Korpusu Narodowego, jak i w publikacji *Nový encyklopedický slovník češtiny*. Definicja V. Cvrčka odnosi nas do historii i etymologii skrótowca, a nie terminu używanego współcześnie. Mamy tu do czynienia ze zjawiskiem deleksykalizacji. *KWIC* jest tu zatem rozumiany dosłownie jako rozwinięcie skrótowca *KWIC*, czyli *keyword in context* ‘słowo wyszukiwane w kontekście’.

Zkratka z angl. ***key word in context*** (***klíčové slovo v kontextu***), kterou se označuje hledaný výraz (n. kombinace výrazů) v různě velkém kontextu. Je to základní součást každé korpusové konkordance, většinou graficky odlišená od okolních (kontextových) slov. ***KWIC*** může v závislosti na typu dotazu reprezentovat jedno slovo n. jejich kombinaci (n-gram) (Cvrček 2017)[5].

Co warte podkreślenia definicja ma się nijak do praktycznego zastosowania tego terminu w Narodowym Korpusie Języka Czeskiego. W aplikacji *Kontext* w zakładce *Zobrazení* ‘sposób wyświetlenia’ mamy dwie możliwości: *KWIC* i *SENTENCE*. Przy wyborze *KWIC* otrzymujemy konkordancję, w której jednostka wyszukiwana znajduje się na środku listy, natomiast przy wyborze *Sentence* otrzymujemy zbiór jednostki wyszukiwanej w zdaniu (por. rys. 1–2).

Ponadto w tej samej aplikacji jednostka wyszukiwana jest nazwana *Node*, a nie *KWIC* zgodnie z najnowszymi trendami obecnymi w lingwistyce korpusowej (por. rys. 3).

Innym terminem, który jest różnie rozumiany przez niektórych przedstawicieli czeskiej lingwistyki korpusowej, jest również termin wielowyrazowy *referenční korpus* ‘korpus referencyjny’. Należy tu jednak sprecyzować, że na przestrzeni lat doszło tu do zmiany definicji. Dotyczy to wyłącznie środowiska Czeskiego Korpusu Narodowego (Uniwersytet Karola). W słowniku *A Glossary of Corpus Linguistics* znajdujemy dwa znaczenia tego terminu. Pierwsza dotyczy korpusu, do którego porównujemy inny, mniejszy korpus przy zastosowaniu metod badawczych opartych na frekwencji (metoda słów kluczowych), np. w sytuacji, kiedy chcemy ustalić, czy jednostka wyszukiwana jest typowa dla poszczególnego tekstu/zbioru tekstów. Podstawą porównania jest większy zbiór tekstów zawierający szerszą gamę rodzajów tekstów. Ten zbiór jest często nazywany korpusem referencyjny. Ta definicja jest zgodna z definicją czeską umieszczona na stronach www.korpus.cz.

Drugie znaczenie korpusu referencyjnego (Hardie, MacEnery, Baker 2006) dotyczy rodzaju korpusu, który nie jest przykładem konkretnego wariantu języka czy typu tekstu. Jest natomiast próbą reprezentacji języka ogólnego. I tak jest definiowany w polskiej literaturze przedmiotu[6].

Od wielu lat *korpus referencyjny* jest definiowany w innych publikacjach czeskich (por. Čermák 2017[7]). W związku z tym tylko część korpusów tworzonych w Instytucie Czeskiego Korpusu Narodowego była określana jako referencyjne. Były to zatem korpusy zamknięte, złożone z różnych gatunków testów, proporcjonalnie dobrane na podstawie przeprowadzonych badań. Wraz z odejściem prof. Čermáka nastąpiła zmiana definicyjna. Od tej pory wszystkie korpusy, które zostały stworzone w Instytucie uzyskały miano referencyjnych. Argumentem było tu dosłowne potraktowanie leksemu *referenční,* czyli taki, który ma *reference* ‘odniesienie do konkretnych źródeł’, jest zamknięty, a materiał w nim zgromadzony jest zawsze ten sam i dostępny przy każdym wyszukiwaniu (dosł. *entita, která je zpětně dostupná*

[5] W definicji *konkordancji* Cvrček 2017 wspomina, co prawda, że w ramach *konkordancji* wyróżnia się *KWIC*, ale definiuje tu *KWIC* jako wyraz wyszukiwany występujący z prawym i lewym kontekstem, co nie do końca odpowiada definicji oryginalnej.

[6] Przedstawiciele słowackiej lingwistyki korpusowej pracujący w Słowackim Korpusie Narodowym nie używają terminu *korpus referencyjny*.

[7] „[...] který se zvláště pro svou reprezentativnost a respektovanou povahu užívá jako standard k poměřování jiných korpusů, avšak bez důrazu na časový aspekt. Je pochopitelné, že zvláštní důležitost mají u některých jazyků *korpusy nářeční*, v č. zatím v zásadě však neexistující” (Čermák 2017).

www.korpus.cz). Na stronie Czeskiego Korpusu Narodowego jego przedstawiciele zaznaczają, że każdy z korpusów jest opublikowany przez CzNK jest korpusem referencyjnym (w przeciwieństwie do pierwszego znaczenia tego terminu), nawet jeśli jest specjalistyczny i zawiera tylko jeden typ tekstu.

2.2.2. Polisemia jako źródło rozbieżności terminologicznych

Wieloznaczność leksemu lub połączenia wielowyrazowego stanowiącego termin może przysparzać kłopotów z identyfikacją terminu. Takimi przykładami mogą być *konkordancja* i *słowo kluczowe*.

Termin *konkordancja* ma długą tradycję zarówno w języku angielskim, jak i w badanych przeze mnie językach zachodniosłowiańskich. Jego początek wiąże się z konkordancją biblijną. Można zatem znaleźć definicje przedkorpusowe i korpusowe.

Definicje korpusowe nie różnią się właściwie od siebie. Wszystkie tłumaczą konkordancję jako listę wystąpień wyszukiwanego leksemu por.:

> A concordance is a list of all of the occurrences of a particular search term in a corpus, presented within the context in which they occur – usually a few words to the left and right of the search term. A search term is often a single word although many concordance, programs allow users to search on multiword phrases, words containing wildcards, tags or combinations of words and tags.' (Hardie, McEnery, Baker 2006, s. 42).

> Všechny doklady (výskyty) hledaného jevu v korpusu spolu s okolním kontextem; někdy se však uživatel spokojí jen s výběrovou **k.**, je-li v ní hledaný jev dostatečně ukázán (Cvrček 2017);

> Zestaw wszystkich wystąpień danego wyrazu w analizowanym tekście wraz z ich kontekstami (Lewandowska-Tomaszczyk 2005, s. 296);

> Výpis kontextov kľúčového, hľadaného slova al. reťazca znakov s jeho zvyčajným zobrazením spolu s určitým kontextom (Šimková 2006, s. 3).

Większość zastosowań leksemu *konkordancja* jest zgodna ze znaczeniem podawanym w słownikach angielskich i badanych tu języków słowiańskich. Niemniej jednak zdarzają się poświadczenia,

z których wynika, że termin *konkordancja* jest użyty w znaczeniu pojedynczego wystąpienia jednostki wyszukiwanej. Take przykłady znalazłam zarówno w polskiej, jak i czeskiej literaturze przedmiotu.

> Aby uzyskać konkordancje dokładnych dopasowań pojedynczych słowoform lub fraz składających się z dwóch lub więcej słowoform, należy je wpisać w polu wyszukiwania i wcisnąć ikonę lupy. Wielkość liter nie ma znaczenia w zapytaniach (http://monco.frazeo.pl/help).

> Konkordancji na danej stronie może być więcej niż mogłaby to sugerować wybrana przez użytkownika wartość opcji limitu, jeżeli co najmniej jedno z pobranych z indeksu zdań zawiera więcej niż jeden kontekst pasujący do zapytania (http://monco.frazeo.pl/help).

W interfejsie programu *Frazeo* widnieje zakładka *konkordancje*, a nie *konkordancja*. Wydaje się, że źródło pluralizacji leksemu *konkordancja* tkwi w użyciu go w znaczeniu niekorpusowym. Połączenia językowe *konkordancje biblijne* są często używane właśnie w liczbie mnogiej w przeciwieństwie do wyrażenia *konkordancja* (korpusowa). Tym samym dochodzi do mylenia poświadczenia z konkordancją. Wydaje się, że może mieć to źródło właśnie w polisemii badanego terminu.

Innym przykładem, który należy do najbardziej polisemicznych terminów zarówno w języku angielskim, jak i w badanych tu językach zachodniosłowiańskich, jest odpowiednio: ang. *keyword*, cz. *klíčové slovo*, pl. *słowo kluczowe*, słc. *kľúčové slovo*. Są to terminy równie definiowane w zależności od dziedziny, w której są używane. Oprócz lingwistyki korpusowej, są stosowane m.in. w informacji naukowej, informatyce, analizie dyskursu. W angielskiej literaturze przedmiotu odnoszącej się do lingwistyki korpusowej *keyword* oznacza specjalistyczny termin, stosowany w metodzie kwantytatywnej słów kluczowych. *Keyword* jest zatem według definicji jednostką, która jest wynikiem zastosowanych logarytmów, najczęściej log-likelihood lub Chi. Jej frekwencja dla danego tekstu lub zbioru tekstów jest wyższa, niż by się tego można było spodziewać.

Zdarza się również, choć miało to miejsce, raczej w starszej literaturze przedmiotu, że leksem *keyword* jest stosowany w znaczeniu

jednostki wyszukiwnej. Teraz na to miejsce używa się terminów *node* czy *search term*.

Ta wieloznaczność jest również obecna w językach zachodniosłowiańskich. W języku czeskim można odnaleźć oba znaczenia, w języku słowackim właściwie tylko w znaczeniu terminu wyszukiwanego, w języku polskim oba znaczenia, np.

Słowo, które pojawia się w tekście lub w korpusie statystycznie częściej, niż byłoby oczekiwane, kiedy porównamy z korpusem o większym lub takim samym rozmiarze. (Lewandowska-Tomaszczyk 2005, s. 300).

Metoda słów kluczowych (keyword analysis) – metoda badawcza polegająca na porównaniu za pomocą testów statystycznych listy frekwencyjnej wyrazów danego tekstu lub grupy tekstów z listą frekwencyjną wyrazów korpusu frekwencyjnego. Pozwala ustalić, które słowa występują częściej w danym tekście lub grupie tekstów niż w języku stanowiącym normę. Są to tak zwane słowa kluczowe (Lewandowska-Tomaszczyk 2005, s. 295);

Aplikace **KWords** poskytuje základní východisko pro empiricky podloženou interpretaci textů tím, že analyzuje slova v zadaném textu a porovnává jejich frekvenci s referenčním korpusem. Výsledkem takové analýzy je identifikace klíčových slov (tzv. keywords), tj. jednotek vyskytujících se signifikantně častěji v analyzovaném textu než v korpusu, který představuje neutrální jazykový úzus.' (Fidler, Cvrček 2015).

Słowo kluczowe jako jednotka wyszukiwana:

– cz. Vyhledávání je možné zpřesňovat také zadáním části jména korpusu nebo jeho popisu do vyhledávacího řádku, výsledný seznam korpusů se přitom podle takto zadaných klíčových slov nebo jejich částí interaktivně filtruje (Křen 2020).
– słc. *Kľúčové slovo* (*keyword*) hľadané slovo al. reťazec znakov zapisovaný do vyhľadávacieho okienka a po vyhľadaní zobrazovaný zvyčajne uprostred kontextu – konkordančného výpisu (Šimková 2006, s. 3).

3. Podsumowanie

Źródłem dla terminologii lingwistyki korpusowej w językach zachodniosłowiańskich jest bez wątpienia angielszczyzna. W każdym

z badanych tu języków mamy do czynienia zarówno z procesem przyswajania pożyczek oznaczających terminy występujące w lingwistyce korpusowej, jak i z próbami tworzenia terminologii na bazie rodzimego słownictwa. Znaczenie terminów jest bądź zgodne ze znaczeniem oryginalnym, bądź zmodyfikowane. W niektórych przypadkach dochodzi do rozbieżności znaczenia nie tylko na płaszczyźnie międzyjęzykowej, ale również w obrębie jednego języka. Wytłumaczenie tego zjawiska nie jest związane z charakterem lingwistyki korpusowej, ale ze specyfiką terminologii:

[…] różnicom terminologicznym sprzyjają rozbieżne rozumienia tego samego zjawiska językowego, stosowanie przez różnych badaczy różnych badaczy różnych terminów do nazwania tych samych zjawisk, błędy w tłumaczeniu terminów na inne języki, pary dubletowe, czyli istnienie w jednym języku zarówno terminu obcego, jak i rodzimego, różne rozumienie jednego terminu przez różnych badaczy, a także semantyczne zazębianie się terminów (Witalisz 2012, s. 11).

Źródła

www.korpus.cz
www.korpus.pl
www.korpus.sk

Literatura

C v r č e k Václav, 2017, *Konkordance*, [w:] Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. Online: https://www.czechency.org/slovnik/KONKORDANCE [dostęp: 31.12.2020].

C v r č e k Václav, 2017, *KWIC*, [w:] Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. Online: https://www.czechency.org/slovnik/KWIC [dostęp: 8. 12. 2020].

C v r č e k Václav, 2017, *Token*, [w:] Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny. Online: https://www.czechency.org/slovnik/TOKEN [dostęp: 31. 12. 2020].

C h l u m s k á Lucie, 2014, *Není korpus jako korpus: Korpusy v kontrastivní lingvistice a translatologii*, „Časopis pro moderní filologii" 96, cz. 2, s. 221–232.

C h l u m s k á Lucie, 2017. Online: https://wiki.korpus.cz/doku.php/seznamy:tagy [dostęp: 9.12.2020].

Č e r m á k František, 2017, *Korpusová lingvistika*, [w:] Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny.*

Online: https://www.czechency.org/slovnik/KORPUSOVÁ LINGVISTIKA [dostęp: 12. 12. 2020].

Čermák František, 2017, *Typy korpusů*, [w:] Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. Online: https://www.czechency.org/slovnik/TYPY KORPUSŮ [dostęp: 20.12.2020].

CASS, 2013, Corpus Linguistics: Some Key Terms.

Fedorushkov Yuri, 2018, *Prolegomena do tagowania frazemów w równoległym korpusie rosyjsko-polskim (literatura piękna) w aspekcie przekładoznawczym*, „Acta Polono-Ruthenica XXIII/2", Olsztyn.

Fidler Mosako, Cvrček Václav, 2015. Online: https://kwords.korpus.cz [dostęp: 8.12.2020].

http://monco.frazeo.pl.

Grabowski Łukasz, Hebal-Jezierska Milena, 2016, *O różnych korpusowych metodach badawczych – próba krytycznej refleksji*, „Komunikacja Specjalistyczna 11", s. 65–83.

Hardie Andrew, McEnery Tony, Baker Paul, 2006, *A Glossary of Corpus Linguistics*, Edinburgh.

Hebal-Jezierska Milena, 2014, *Praktyczny przewodnik po korpusach języków słowiańskich*, Warszawa.

Křen Michal, 2020. Online: https://wiki.korpus.cz/doku.php/manualy:kontext:novy_dotaz [dostęp: 20.12.2020].

Lewandowska-Tomaszczyk Barbara (red.), 2005, *Podstawy językoznawstwa korpusowego*, Łódź.

Łukasik Marek, 2008, *Narzędzia lingwistyki korpusowej w warsztacie terminologa, terminografa i tłumacza tekstów specjalistycznych (cz. 1)*, [w:] *Debiuty Naukowe 1. Wiedza – korpus – słownik*, Warszawa, s. 23–47.

Mańczak-Wohlfeld Elżbieta, 2008, *Morfologia zapożyczeń angielskich w językach europejskich*, Kraków.

McEnery Tony, Hardie Andrew, 2012, *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press.

Pęzik Piotr, 2020 *Budowa i zastosowania korpusu monitorującego MoncoPL,* „Forum Lingwistyczne" 7, s. 133–150.

Palek Bohumil, 1969, *Type-token a lingvistika*, „Slovo a slovesnost" 30, číslo 3, s. 263–268.

Przepiórkowski Adam, Bańko Mirosław, Górski Rafał, Lewandowska-Tomaszczyk Barbara (red.) *Narodowy Korpus Języka Polskiego.* Wydawnictwo Naukowe PWN, Warszawa.

Saloni Zygmunt, 1993, *Okaz,* [w:] Kazimierz Polański (red.) *Encyklopedia językoznawstwa ogólnego*, Wrocław, s. 402.

Stewart Dominic, 2013, *Semantic prosody: a critical evaluation*, New York, London.

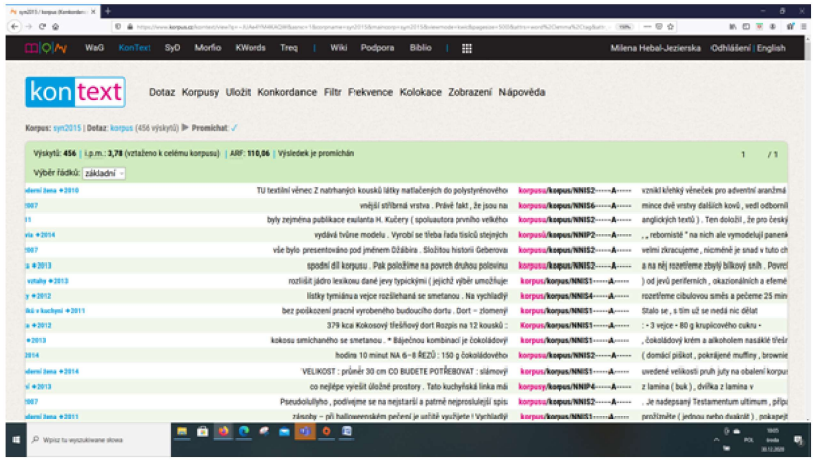
Šimková Maria, 2006, *Výberový slovník termínov z počítačovej a korpusovej lingvistiky*. Online: https://korpus.sk/attachments/what/2006-simkova-vyberovy_slovnik_terminov.pdf [dostęp: 20.12.2020].

Tkaczewski Dariusz, 2013, *Ottův slovník naučný na tle czeskiej tradycji leksykograficznej: encyklopedia – twórcy – język*, Katowice.

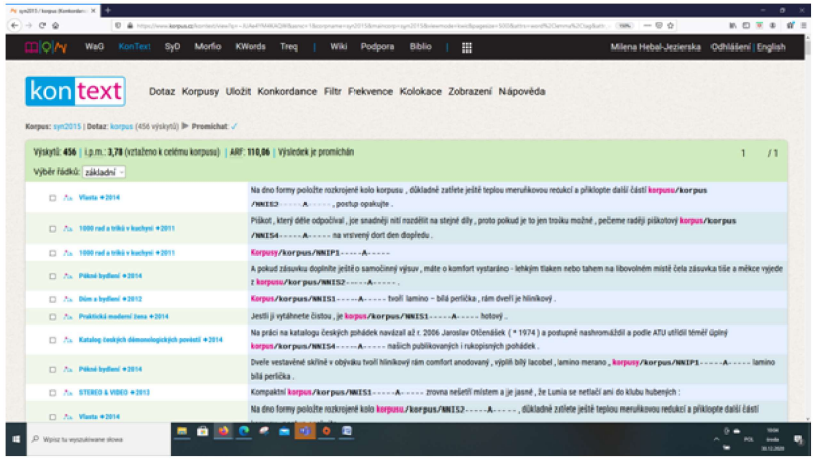
Witalisz Alicja, 2012, *O rozbieżności terminologicznej w teorii zapożyczeń językowych* [w:] Dorota Brzozowska, Władysław Chłopicki (red.), *Termin w językoznawstwie*, Kraków 2012, 107–114.

Zabawa Marcin, 2017, *Neosemantyzmy i zapożyczenia semantyczne jako odzwierciedlenie postępu technologicznego i zmian kulturowo-obyczajowych*, „Język Polski" XCVII, nr 2, s. 94–104.

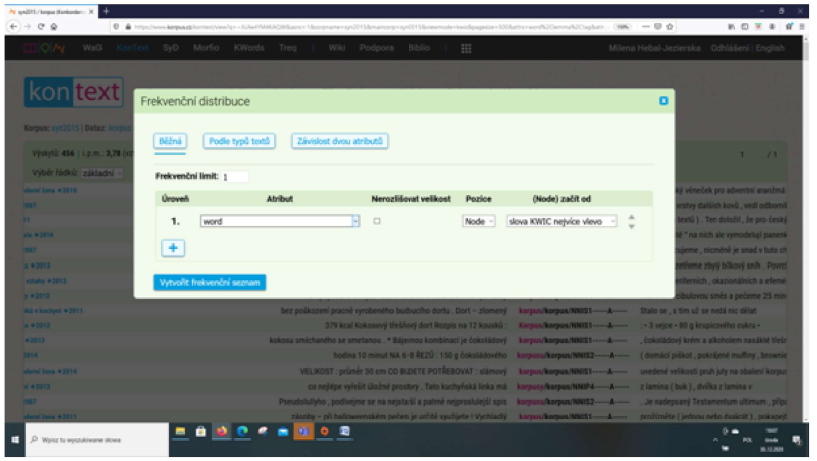
Zasina Adrian, 2018, *Językoznawstwo korpusowe*, [w:] Igor Borkowski (red.), *Empiryczne podejście w badaniach humanistycznych, dziennikarstwo i media, Metodologie i praktyki badawcze*, Wrocław, s. 169–178.



Rys. 1. Przykład sposobu wyświetlenia konkordancji typu KWIC



Rys. 2. Przykład sposobu wyświetlenia konkordancji typu Sentence



Rys. 3. Jednostka wyszukiwana nazwana Node