2018 PRACE JĘZYKOZNAWCZE XX/2
ISSN 1509-5304 75–97

Aleksander Kiklewicz
Uniwersytet Warmińsko-Mazurski, Olsztyn
e-mail: aleksander.kiklewicz@uwm.edu.pl

Korpus internetowy jako źródło informacji lingwistycznej: ograniczenia*

Internet corpus as a source of linguistic information: some limitations

The author shows several limitations of the corpus-based linguistic information in syntactic studies. In the case of the most frequent phenomena, corpus analysis is effective, but it does not always allow to document less typical phenomena (for example, occasional and potential combinations of tokens). One of the author's conclusions is that corpus analysis should be configured with introspection and qualitative analysis.

Słowa kluczowe: lingwistyka korpusowa, korpus internetowy, składnia, argument propozycjonalny, komplement zdaniowy, introspekcja lingwistyczna
Key words: corpus linguistics, Internet corpus, syntax, propositional argument, clausal complement, linguistic introspection

Przedmiotem niniejszego artykułu są korpusy językowe w internecie jako źródło informacji o jednostkach językowych różnego formatu (czy też różnego stopnia złożoności) ze szczególnym uwzględnieniem jednostek składniowych. W ostatnich dziesięcioleciach nastąpił prawdziwy *boom* w dziedzinie lingwistyki korpusowej, a korpusy internetowe języków narodowych zaczęto traktować jako narzędzie, które pozwala zarówno na przyśpieszenie procedur badawczych (za sprawą konkordancji), jak i na zdobycie jakościowo nowej wiedzy o systemie języka i jego realizacji w tekstach. Jednocześnie badacze zdają sobie sprawę z pewnych ograniczeń metody korpusowej: po pierwsze,

* Prezentowany artykuł jest przygotowany w ramach realizacji projektu naukowo-badawczego „Właściwości składniowe czasowników jako baza ich zintegrowanego opisu leksykograficznego (w perspektywie konfrontacji polsko-bułgarsko-rosyjskiej)" przyznanego na lata 2014–2017 przez Narodowe Centrum Nauki RP (nr grantu: 2013/11/B/HS2/03116).

każdy korpus, nawet największy[1], stanowi pewną ekscerpcję materiału językowego, więc nie obejmuje całego obszaru działalności językowej ani wszystkich możliwości systemu języka. Nawet gdybyśmy wyobrazili sobie maksymalnie pojemny korpus, który zawiera wszystkie teksty wytworzone w danym języku, musimy zdawać sobie sprawę, że w trakcie tworzenia korpusu ciągle będą pojawiać się nowe teksty, a pewne, odnotowane w korpusie elementy będą stawać się funkcjonalnie nierelewantne.

Po drugie, korpus jest fenomenem empirycznym, czyli zbiorem jednostek – nie jest nawet zbiorem danych. Jednostki te wymagają kategoryzacji, a otrzymana w ten sposób informacja wymaga interpretacji lingwistycznej, gdyż za każdym faktem mowy stoją jakieś reguły, normy, prawa lub algorytmy systemu języka (oczywiście realizujące się przy współudziale tzw. czynników zewnętrznych).

1

Językoznawstwo, w najogólniejszym wymiarze, zajmuje się opisem czterech kategorii obiektów, postulowanych w ogólnej teorii systemów, w szczególności w wersji J. A. Urmancewa (1978: 10 i n.), który rozróżnił parametry, niezbędne do opisu każdego systemu:

1) elementy podstawowe (*первичные элементы*);

2) właściwości, funkcje jako podstawy kategoryzacji i klasyfikacji elementów podstawowych (*основания*);

3) klasy ufundowane na relacjach podobieństwa elementów (*отношения единства*);

4) kompozycje tworzone na podstawie określonych reguł łączliwości jednostek podstawowych zgodnie z ich funkcją i przynależnością do klasy (*законы композиции*).

Badania empiryczne mają za zadanie dostarczenie językoznawcom wstępnej informacji o tych obiektach, a mianowicie: 1) rejestrację czy też inwentaryzację jednostek; 2) eksplikację opozycji jednostek, ufundowanych na ich określonych, relewantnych właściwościach; 3) eksplikację relacji tożsamości (w szczególności relacji zamienności) jednostek tworzących

[1] Dla przykładu Narodowy Korpus Języka Polskiego obejmuje 1,5 mld słów (podkorpus zrównoważony – 250 mln słów). Objętość niektórych innych korpusów: COSMAS (język niemiecki) – 42 mld słów; Narodowy Korpus Języka Rosyjskiego – 600 mln słów; Integrum (korpus języka rosyjskiego) – 500 mln dokumentów; British National Corpus – 100 mln słów; The Oxford English Corpus – 2,5 mld słów; Corpus of Contemporary American English – 520 mln słów; Czeski Korpus Narodowy – 4 mld słów (język pisany), 7 mln słów (język mówiony).

tę samą klasę; 4) ukazanie wszystkich możliwych kompozycji jednostek (np. konstrukcji składniowych). Korpus językowy w internecie w największym stopniu nadaje się do tego, aby służyć realizacji tych wszystkich zadań.

Badania korpusowe są szczególnie istotne w zakresie składni. O ile w przypadku leksyki istnieją uznane przez tradycję źródła lingwistyczne: słowniki (nie tylko opisowe, lecz także ortograficzne, synonimów, antonimów, wyrazów wieloznacznych, wyrazów zapożyczonych, neologizmów, archaizmów itd.), a korpusy internetowe są przydatne przy realizacji szczególnych zadań, np. w badaniach semazjologicznych, przy określeniu częstości użycia wyrazów itd., o tyle informację o zdaniach i grupach wyrazowych w wystarczającym wymiarze możemy uzyskać, czerpiąc ze źródeł tekstowych, m.in. zgromadzonych w dużych korpusach, dostępnych w internecie. Należy pamiętać, że jednostki składniowe (z wyjątkiem jednostek frazeologicznych, o utrwalonym składzie i utrwalonej strukturze) mają charakter kompozycyjny i są tworzone (nie odtwarzane, jak leksemy) w procesach działalności mownej. Wobec tego nie do wyobrażenia jest jakikolwiek rejestr „wszystkich zdań" nawet najbiedniejszego języka – w odróżnieniu od leksykografii, która dąży do coraz szerszego ujęcia słownictwa[2], które ma charakter zamknięty. Jeśli rozważymy jeden z dowolnych przykładów – zdanie pochodzące z powieści Joanny Chmielewskiej:

(1) *Rytmiczne podskoki, do których sądziłam, że już przywykłam, dały mi się we znaki dopiero wtedy, kiedy opuściwszy sterówkę, zaczęłam wykonywać normalne czynności.*

łatwo się przekonać, że niepowtarzalność, unikatowość jednostek zdaniowych polega nie tylko na specyficznym skonfigurowaniu leksemów, lecz także na konstrukcji zdaniowej jako zespole związków składniowych (międzywyrazowych i międzyzdaniowych) oraz gramatycznych form realizacji pozycji syntaktycznych. Na przykład początku zdania trudno jest przyporządkować jakiemukolwiek schematowi zdań złożonych, gdyż jest to rezultat specyficznego procesu przekształcenia struktury syntaktycznej:

(2) *Rytmiczne podskoki, do których sądziłam, że już przywykłam…* < *Rytmiczne podskoki, do których, jak sądziłam, już przywykłam…*

Można jedynie teoretycznie wyobrażać sobie możliwość algorytmicznego, maszynowego programowania wszystkich dopuszczalnych przez system języka konfiguracji związków i form gramatycznych (łącznie z modyfikacjami), jako że wykracza to poza operacjonalne możliwości współczesnego językoznawstwa.

[2] Por. projekt „300 000 polskich słów" J. Wawrzyńczyka i P. Wierzchonia, 2016.

2

Istnieją różne źródła wiedzy o językowych regułach kompozycji (zob. Kiklewicz 2009: 207 i n.). Są to źródła językowe, czyli pierwszego rzędu: teksty pisane, nagrania, teksty elektroniczne, w tym wyszukiwarki i (zawierające konkordancję dokumentów) korpusy internetowe, intuicja językowa, w pewnym stopniu także dane eksperymentalne (zwłaszcza uzyskane za pośrednictwem ankietowania) i in. Do źródeł lingwistycznych, tzn. drugiego rzędu, należy zaliczyć: opisy gramatyczne języka, słowniki syntaktyczne (np. słowniki walencyjne, słowniki rekcji czasowników, słowniki kolokacji itp.), słowniki opisowe (zawierające egzemplifikację wyrazów hasłowych w postaci zdań i grup wyrazowych). Jeśli chodzi o źródła drugiego rzędu, ich poważnym mankamentem jest niekompletność. Gramatyki bazują przede wszystkim na języku literatury artystycznej (choć są też, co prawda, nieliczne wyjątki od tej reguły, zob. Topolińska 1984: 4) i nie odzwierciedlają wielu osobliwości składni tekstów retorycznych, technicznych, specjalistycznych i in. Egzemplifikacje w słownikach opisowych są podporządkowane zasadzie uzualizmu: wyselekcjonuje się przykłady użycia jednostki hasłowej w tekstach literatury artystycznej lub dopuszczalne z punktu widzenia normy użytkowej przykłady użycia w mowie potocznej (więcej o tym zob.: Korytkowska/Kiklewicz 2016: 299 i n.). Poza tym egzemplifikacje w popularnych słownikach opisowych mają charakter pomocniczy i, w istocie rzeczy, marginalny, więc trudno oczekiwać od nich pełnej reprezentacji struktur składniowych języka.

Wydawałoby się, że problem rozwiązują źródła specjalistyczne – słowniki syntaktyczne, tym bardziej że w ostatnim czasie powstaje sporo słowników walencyjnych na platformie elektronicznej, jak np. WALENTY (http://zil.ipipan.waw.pl/Walenty) czy FrameNet (https://framenet.icsi.berkeley.edu/fndrupal). Analiza jednak wykazuje, że i te źródła odzwierciedlają pewne fragmenty systemu syntaktycznego języka, a przede wszystkim eksponują go w określonej perspektywie, zgodnie z teoretycznymi poglądami i programem badawczym określonego zespołu wykonawców.

Można to skrótowo pokazać na przykładzie niektórych słowników rosyjskich. Tak więc w pracy: Kiklewicz/Korytkowska 2013: 54 i n. już zwrócono uwagę na pewne ograniczenia *Słownika opisowo-kombinatorycznego* I. A. Mielczuka i A. K. Żołkowskiego (1984). Po pierwsze, *Słownik* nie zawiera pełnego wykazu struktur zdaniowych. Na przykład w przypadku czasownika *бороться* ‘walczyć’ w znaczeniu: ‘*X* dokłada wiele starań w celu zwalczenia sytuacji *Y*’ zostały przytoczone dwie gramatyczne formy realizacji drugiego argumentu:

с S_{Instr}
против S_{Gen}
Autorzy przytaczają ilustracje takich użyć czasownika:

(3) *бороться с сорняками*
(4) *бороться с пьянством*
(5) *бороться против опозданий*
(6) *бороться против нарушения графика*

W powyższym wykazie brakuje jednak wskazania na eksplikację drugiego argumentu w formie zdania zależnego, choć, skoro w definicji czasownika wspomniano o zwalczaniu sytuacji, możliwość werbalizacji tego elementu znaczenia w pełnej, dyskretnej formie zdaniowej jest wręcz oczekiwana, a poza tym ta forma jest odnotowana w tekstach, por.:

(7) *Как бороться с тем, что я очень впечатлительный? (интернет).*
(8) *Есть ли смысл вообще бороться с тем, от чего легко могут отказаться потенциальные нарушители при сохранении тех же рисков для других участников рынка (Александр Лобыкин).*
(9) *А как будем бороться с тем, что ремонтируют то, что не надо? (интернет).*

Po drugie, przytoczone w haśle egzemplifikacje nie odzwierciedlają wszystkich typów dystrybucyjnych. Na przykład dla czasownika *гневаться* ‘gniewać się’ są przytoczone gramatyczne formy wyrażenia drugiego i trzeciego argumentu, a także są wskazane ich możliwe i niemożliwe konfiguracje:
∅ + на S_{Acc}
на S_{Acc} + за S_{Acc}
на S_{Acc} // ∅ + из-за S_{Gen}
∅ + на то, что/C_{Con} SENT
на S_{Acc} + за то, что/C_{Con} SENT
на S_{Acc} // ∅ + из-за того, что/C_{Con} SENT
Ilustracje zdaniowe reprezentują jednak tylko cztery struktury syntaktyczne:

(10) *Истинный художник не гневается на знатока.*
(11) *Истинный художник не гневается на критику знатока [на то, что его критикуют].*
(12) *Отец гневается на сына за опоздание [за то, что он опоздал].*
(13) *Он гневается на учителей из-за сына [из-за того, что они не могут выучить его сына].*

Brakuje jednak ilustracji zdaniowych innych typów łączliwości czasownika (przykłady pochodzą z internetu):

(14) *Генерал гневался из-за отъезда в Париж дочери Анны.*
(15) *На видео видно, как этот человек гневается из-за того, что у него отняли какой то пиксель.*

Może to sugerować, że autorzy *Słownika* nie uznają realizacji drugiego argumentu w formie *из-за того, что*/C_{Con} SENT lub *из-за* S_{Gen} bez wypełnienia pozycji drugiego argumentu, jednak, jak widzimy, praktyka językowa nie potwierdza tego postulatu.

W materiale językowym (w korpusie, a także w wyszukiwarce internetowej) można odnaleźć także nieodnotowane w *Słowniku* zdania ze spójnikiem *что* oraz ze spójnikiem *потому что,* por.:

(16) *Иисус гневался также и потому, что любил родителей и понимал их стремления и тревогу.*
(17) *Он гневался еще и потому, что никто не находится вне заботы и любви Бога.*
(18) *А гневалась она потому, что ненавидела всякую домостроевщину.*
(19) *Ты гневаешься, что я тебя с женою неласково принял и мало наградил.*
(20) *И не гневайтесь, россияне, что они строят виллы и ездят на Канарские острова.*
(21) *Одни были возмущены,* […] *гневались, что их фамилий не оказалось в скандальной хронике.*

Nawet gdyby okazało się, że niektóre konstrukcje składniowe (np. ze spójnikiem *потому что*) występują przeważnie w tekstach określonego stylu, a mianowicie w tekstach o tematyce religijnej, nie daje to podstawy, żeby nie uwzględniać ich w opisie lingwistycznym, gdyż są one niesprzeczne z systemem językowym, tzn. nie stanowią dewiacji z punktu widzenia normy ogólnej.

Słownik jednoznacznie wskazuje na niemożliwość występowania formy *за* S_{Acc} bez formy *на* S_{Acc}, jednak stoi to w sprzeczności z praktyką językową. W *Słowniku języka rosyjskiego* pod redakcją A. P. Jewgieniewej w haśle *гневаться* napotykamy właśnie przykład użycia grupy wyrazowej o strukturze *гневаться* + *за* S_{Acc} bez wykładnika drugiego argumentu:

(22) *Степан Михайлович не будет гневаться за нарушение его приказания.*

Co prawda, jest to zdanie z prozy XIX w., ale i we współczesnych tekstach można odnotować konstrukcje tego typu (przykłady pochodzą z internetu):

(23) *Всуе гневаешься за всякое лихоимство.*
(24) *Гневается великий сын Латоны за то, что обесчестил царь Агамемнон жреца его Хриса.*
(25) *А если врач такую табличку не повесил, тогда за что гневаться?*
(26) *Братья гневались за свою нацию.*
(27) *Сам Иоанн Васильевич, названный мучителем, не гневался за подобные грубости.*
(28) *Гневаться за своих — свойство не преданной души, а слабой.*
(29) *Николай Чудотворец гневается за то, что его мощи Путин оскорбил.*
(30) *Похоже, боженька гневается за всё.*

Po trzecie, porządek eksplikacji form gramatycznych w schematach walencyjnych (określanych w *Słowniku* jako *government pattern*) budzi zastrzeżenia, szczególnie jeśli chodzi o drugi argument (czyli argument propozycjonalny). W pierwszej kolejności zamieszczono formy skondensowane (*на* S_{Acc}, *за* S_{Acc}, *из-за* S_{Gen}), będące rezultatem przekształcenia struktury składniowej, w której pozycja ta jest zarezerwowana dla zdania zależnego (*на то, что* SENT; *за то, что* SENT; *из-за того, что* SENT). Stoi to w sprzeczności z faktem, że formy zdaniowe – z punktu widzenia systemu języka – mają charakter podstawowy, wynikający z konotowanej przez czasownik struktury propozycjonalnej.

Po czwarte, może budzić wątpliwości ujęcie czasownika *гневаться* jako dwuargumentowego. Owszem, wiele realizacji zdaniowych zawiera obydwa człony: *na kogoś* i *z powodu czegoś*. Jednak trudno jest wytłumaczyć istnienie formy gramatycznej *на* + S_{Acc}, która wskazuje zarówno na drugi, jak i na trzeci argument, gdyż powstała w ten sposób homonimia blokuje jednoczesne użycie obydwu wykładników. Bardziej przekonującym i logicznym byłoby wytłumaczenie, że forma ta jest przyporządkowana temu samemu argumentowi ze znaczeniem przyczyny czy podstawy stanu emocjonalnego. Wobec tego czasownik *гневаться* należałoby potraktować jako predykat dwuargumentowy: $P(x, q)$. Za tym rozwiązaniem przemawia także fakt, że istnieje wiele realizacji zdaniowych, w których wypełnione są dwie pozycje: experiencera (x) i efektora/kauzatora (q), np. zdania (14) – (24). Poza tym definicja znaczenia tego czasownika w słowniku opisowym ('być rozgniewanym, poirytowanym') nie wskazuje na to, że przewidywana jest obecność trzech argumentów: x może być rozgniewany, poirytowany za sprawą q, natomiast informacja o q może być wyrażona w bardziej prosty (np. *за* S_{Acc}), lub bardziej złożony sposób (*на* S_{Acc} + *за* S_{Acc}).

Występowanie trzech zależnych od predykatu pozycji składniowych (*ktoś gniewa się na kogoś z powodu czegoś*) można wytłumaczyć jako proces rozszczepienia argumentu propozycjonalnego. Zjawisko to (znane w składni teoretycznej, zob. Korytkowska 1992: 92 i n.; Karolak 2002: 105; Apresjan 2010: 351) polega na tym, że jeden z członów argumentu propozycjonalnego q zostaje przeniesiony ze struktury zdania zależnego do struktury zdania głównego:

$$[V\ N_x\ (V_q\ N_x \ldots)] \rightarrow V\ N_x\ \mathbf{N_{xq}}\ (V_q\ \mathbf{N_x} \ldots)$$

Powyższy schemat wskazuje na fakt koreferencji syntaktemów (grup nominalnych), która zachodzi w zdaniach tego typu. Na przykład w zdaniu:

(31) *Он гневается на Ивана за то, что он не выполнил обещание.*

koreferencja dotyczy grup nominalnych *на Ивана* oraz *он*. Gdyby wymóg koreferencji nie został spełniony, zdanie należałoby uznać za niepoprawne, por.:

(32) **Он гневается на Ивана за то, что Михаил не выполнил обещание.*

3

Rola korpusów internetowych jest szczególnie ważna w przypadku opisów aksjomatycznych, tzn. takich, które są oparte na bazowej koncepcji jako modelu lingwistycznym, za pomocą którego generuje się właściwości dystrybucyjne poszczególnych jednostek. Zadanie językoznawcy polega tu na tym, aby zweryfikować możliwości walencyjne jednostki, a mianowicie takie, które wynikają ze struktury pojęciowej i w większym lub mniejszym stopniu są ogólne dla jednostek należących do tej samej klasy semantycznej (np. czasowników mentalnych lub emotywnych). Właśnie na takich założeniach teoretycznych jest ufundowany model składni eksplikacyjnej stosowany w badaniach polskiej szkoły składni semantycznej, założonej przez S. Karolaka (zob. 1984; 2002).

Zgodnie ze wspomnianą koncepcją za punkt wyjściowy opisu struktur zdaniowych przyjmuje się implikowaną przez znaczenie leksykalne czasownika strukturę propozycjonalną jako projekcję typowej sytuacji, ufundowanej na przyporządkowanej czasownikowi czynności, stanie, procesie lub relacji. System językowy dysponuje pewnym zbiorem środków realizacji, tzn. wyrażenia w formach gramatycznych elementów struktury pojęciowej czasownika. Jakkolwiek w innych modelach lingwistycznych (np. w teorii rządu) właściwości formalnogramatyczne związków syntaktycznych stanowią punkt wyjściowy, a ich interpretacja semantyczna następuje na kolejnych poziomach procedury badawczej, to model składni eksplikacyjnej opiera się na zasadzie „sens > tekst", czyli uwzględnia fakt naturalnej, zachodzącej w procesach działalności językowej pochodności struktur formalnogramatycznych od struktur semantycznych. Badanie korpusowe pozwala na to, aby potwierdzić istnienie jednych eksplikacji oraz stwierdzić ograniczenia dystrybucyjne w postaci braku występowania pewnych form gramatycznych w pozycjach otwieranych przez predykat.

Składnia eksplikacyjna postuluje ponadto hierarchiczny układ składniowych form manifestacji pozycji argumentowych, a mianowicie ze względu na to, że formy te mają charakter bardziej lub mniej dyskretny, zupełny,

izosemiczny. W przypadku realizacji argumentów propozycjonalnych, tzn. takich, które wskazują na zdarzenia lub stany rzeczy, za najbardziej zupełną formę eksplikacji uznaje się zdanie zależne, gdyż za sprawą predykacji sentencjonalny charakter argumentu oddaje się w najbardziej transparentny sposób. Wszystkie pozostałe formy gramatyczne traktuje się jako pochodne, powstałe na skutek kompresji lub, odwrotnie, rozszczepienia argumentu propozycjonalnego. Za przykład takiej, zhierarchizowanej klasy realizacji zdaniowych może posłużyć rosyjski czasownik *знать* ‘wiedzieć; znać’:

$q \rightarrow S_q$	*Эраст знает сам, что происходит в душе его.*
$q \rightarrow S_q\ S_q$	*О том, что произошло вчера, он знает, что это была провокация националистов.*
$q \rightarrow NV_{aq}\ S_q$	*О данном событии он знает, что его причины – в эскалации конфликта.*
$q \rightarrow V_q\ NV_q$	*О том, что произошло на площади, Федор знает все подробности.*
$q \rightarrow NV_q$	*Левон Оганезов знает множество историй.*
$q \rightarrow NV_{aq}\ \varnothing_q$	*Петя знает правильное решение задачи.*
$q \rightarrow NV_{aq}\ NV_{aq}\ \varnothing_q$	*О данном событии Иван знает все подробности.*
$q \rightarrow NV_{aq}\ NV_{Pq}\ \varnothing_{Vq}$	*Святитель Василий знал медицину как языческое искусство.*
$q \rightarrow N_{aq}\ \varnothing_q$	*Батюшка знает меня.*
$q \rightarrow N_{aq}\ N_{Pq}\ \varnothing_{Vq}$	*Любители поэзии знают Шульпякова как одного из последних переводчиков.*
$q \rightarrow \varnothing_q$	*[С Иваном не спорьте] – Иван знает.*

Schematy eksplikacyjne (typu $V\,N_x\,S_q$) mają charakter typów kompozycyjnych, bardziej lub mniej regularnych dla jednostek należących do określonej klasy semantycznej. Specyfika relacji między semantyką a gramatyką polega na tym, że podobieństwa w jednym obszarze niekoniecznie pociągają za sobą podobieństwa w drugim obszarze – korelacje mają raczej charakter probabilistyczny. Na przykład czasownik *знать*, o czym świadczą przytoczone ilustracje, dopuszcza jedenaście form realizacji argumentu propozycjonalnego *q,* podczas gdy czasownik *вздумать* ‘zamyślić’, o tej samej strukturze propozycjonalnej *P* (*x, q*), przewiduje tylko trzy takie formy (występuje tu m.in. bezokolicznik, nieobecny w schemacie walencyjnym czasownika *знать*):

$q \rightarrow S_q$	*Купец вздумал, что в этой палке заключена чудесная сила.*
$q \rightarrow VI_q$	*Один раз даже его лошадь вздумала бить задом.*
$q \rightarrow N_{aq}\ \varnothing_q$	*С яростью она вздумала о немцах.*

W związku z tym materiał pochodzący z korpusu powinien rozstrzygać o tym, czy określona, oczekiwana za sprawą przynależności jednostki do klasy semantycznej eksplikacja gramatyczna występuje (jako cecha walencyjna czasownika), jak również dostarczać informacji o częstości i ew. o sferach (stylach funkcjonalnych, dyskursach) jej występowania. Jest to przedsięwzięcie o tyle ważne, że słowniki popularne nie zawsze uwzględniają te, podstawowe z systemowego punktu widzenia reprezentacje gramatyczne, oddając pierwszeństwo formom syntaktycznym, powstałym na skutek kondensacji i częściej spotykanym w codziennej praktyce językowej. Brakuje wskazania na te zjawiska także w wielu słownikach syntaktycznych (walencyjnych), które ograniczają się do opisu właściwości morfosyntaktycznych, uwzględniając jedynie formy gramatyczne rzeczownika w pozycjach implikowanych przez czasownik (por. np. Araszonkawa/Lemciuhowa 1991; Łazutkina 2012).

Praktyka badawcza wykazuje, że korpusy tylko częściowo pomagają w rozwiązaniu tych problemów. Za przykład niech posłuży realizacja argumentów propozycjonalnych w formie komplementu zdaniowego, czyli w konstrukcjach zdaniowych typu *x wie, że…; x wątpi, czy…; x zastanawia się, jak…* Z systemowego punktu widzenia konstrukcje te, jak już zaznaczono w drugim punkcie, uznaje się za najbardziej reprezentatywną formę eksplikacji argumentów propozycjonalnych, dopuszczalną przez większość czasowników (predykatów wyższego rzędu)[3]. Z uzualnego punktu widzenia ich status jest bardziej zróżnicowany: w tekstach jednych stylów mowy komplement zdaniowy występuje częściej, w tekstach innych stylów rzadziej, na co wskazują np. dane, opublikowane przez M. Korytkowską i W. Małdżijewą (2002: 161)[4]. W tekstach forów internetowych na konstrukcje zdaniowe z czasownikiem mentalnym w pozycji orzeczenia, w których pozycję argumentu propozycjonalnego zajmuje komplement zdaniowy, przypada 65,9% jednostek w języku rosyjskim i 62,1% jednostek w języku polskim (Kiklewicz 2016: 106).

Jak widzimy, jest to wręcz regularna forma syntaktyczna, choć nie wszystkie czasowniki (jako predykaty drugiego rzędu) dopuszczają taki typ kolokacji. Na przykład rosyjskie czasowniki mentalne *намереваться/намериться, передумывать/передумать, подумывать/подумать, проникать/проникнуть, удумывать/удумать* nie przewidują zdania zależnego

[3] Z mojej analizy wynika, że komplement zdaniowy jest dopuszczalny w pozycji argumentu propozycjonalnego w przypadku 85,6% czasowników mentalnych i 88,6% czasowników emotywnych w języku bułgarskim, w przypadku 87,1% czasowników mentalnych i 92,3% czasowników emotywnych w języku polskim, w przypadku 88,3% czasowników mentalnych i 93,8% czasowników emotywnych w języku rosyjskim.

[4] Według tych danych najczęściej argument propozycjonalny jest wyrażany w formie zdania zależnego w tekstach artystycznych i publicystycznych, a najrzadziej – w tekstach oficjalno-urzędowych.

w pozycji drugiego argumentu. Granica między tym, co jest dopuszczalne w polu walencji predykatu czasownikowego, a co nie jest, ma jednak charakter rozmyty. Zdarza się, że intuicja językowa podpowiada nam, że komplement zdaniowy jest możliwy co najmniej w jednym ze stylów funkcjonalnych (naukowym, technicznym, prawniczo-ekonomicznym itd.), jednak badanie korpusowe nie potwierdza tej hipotezy. Na przykład polski czasownik *argumentować* ma charakter trzymiejscowy: *P* (*x, q, r*), czyli *ktoś argumentuje, tzn. uzasadnia pewien stan rzeczy poprzez odwołanie się do innego stanu rzeczy*. W polu walencji predykatu znajdują się dwa argumenty propozycjonalne, które realizują się na kilka sposobów: w formie S, NV lub ∅, por.:

NV_q, S_r *Szef argumentował konieczność wyjazdu tym, że istnieje niebezpieczeństwo aresztowania.*

S_q, NV_r *Szef argumentował to, że wyjeżdża, niebezpieczeństwem aresztowania.*

NV_q, NV_r *Szef argumentował swój wyjazd niebezpieczeństwem aresztowania.*

S_q, $Ø_r$ *Primakow [bez niedomówień] argumentował, dlaczego Rosja nie zgodzi się na rozszerzenie NATO.*

NV_q, $Ø_r$ *Komentator [długo] argumentował przyczyny agresji tego państwa.*

$Ø_q$, $Ø_r$ *Andrzej [ma cięty dowcip], [rzeczowo i błyskotliwie] argumentuje.*

Wszystkie powyższe schematy eksplikacyjne zostały odnotowane w NKJP, brakuje jednak materialnego potwierdzenia struktury zdaniowej, w której obydwa argumenty propozycjonalne byłyby realizowane w formie komplementu zdaniowego, czyli typu S_q, S_r. Niewspomnienie tej konfiguracji form miałoby znaczyć, że jest ona zablokowana przez system współczesnego języka polskiego. W tej sytuacji sięgamy do intuicji językowej, która mówi nam, że konstrukcje zdaniowe typu $V\ N_x\ S_q\ S_r$ są poprawne z punktu widzenia systemu języka, w każdym razie jest możliwe w jednym ze stylów funkcjonalnych, por. skonstruowane w taki sposób zdanie:

(33) *To, że brakuje stosownego zapisu, lekarze argumentują tym, że trudno stwierdzić jednoznacznie, która jednostka chorobowa kwalifikuje do wskazania.*

Podobnie jest w języku rosyjskim. Czasownik *интерпретировать* funkcjonuje jako trzymiejscowy predykat wyższego rzędu: *P* (*x, q, r*), czyli *ktoś interpretuje pewien stan rzeczy poprzez odwołanie się do innego stanu rzeczy* (*poprzez upodobnienie go z innym stanem rzeczy*). Zazwyczaj argumenty propozycjonalne *q, r* realizują się w formie grupy nominalnej, ufundowanej na rzeczowniku abstrakcyjnym (w szczególności dewerbatywnym), np.:

$NV_q\ NV_r$ *Он интерпретирует его замысел как попытку проникновения в историю вопроса.*

Konstrukcje tego typu w dużym asortymencie odnajdujemy w Narodowym Korpusie Języka Rosyjskiego. Nie występują tam jednak przykłady wypełnienia tych pozycji przez komplement zdaniowy. Powstaje pytanie: czy czasownik *интерпретировать* rzeczywiście blokuje możliwość występowania takich form syntaktycznych? Intuicja językowa podpowiada nam, że możliwość taka istnieje – i rzeczywiście, w innych źródłach, w szczególności w wyszukiwarce internetowej Rambler.ru, odnajdujemy przykłady użycia tego czasownika, zawierające jedno, a nawet dwa zdania zależne, por.:

$S_q\ NV_r$ *То, что происходит, можно интерпретировать как начало избирательной кампании.*

$NV_q\ S_r$ *Совокупность предлагаемых норм можно интерпретировать как то, что для расчета прибыли КИК на основании финансовой отчетности КИК должна располагаться в стране-партнере РФ.*

$S_q\ S_r$ *То, что нарисовано на картинке, можно интерпретировать как то, что «тирамису побеждает пончик с желе».*

Za tym, że powyższe formy zdaniowe odzwierciedlają rzeczywiste, zakodowane w systemie języka właściwości walencyjne czasownika, przemawia nie tylko fakt, że da się odnotować je w materiale językowym (w tekstach pisanych, mówionych, elektronicznych), lecz także ich niesprzeczność z intuicją językową, zwłaszcza z intuicją badacza. W związku z tym warto nawiązać do wypowiedzenia K. Polańskiego, który przywiązywał dużą wagę do tego źródła informacji lingwistycznej. W przedmowie do *Słownika syntaktyczno-generatywnego czasowników polskich* wybitny językoznawca pisał:

> Zgodnie z postulatem gramatyki generatywnej nie pozwalającym lingwiście ograniczać się do tzw. korpusu tekstowego trzeba w poszukiwaniu faktów językowych bardzo często sięgać do intuicji językowej własnej i innych współużytkowników języka (Polański 1980: 6).

Za przykład zintegrowanej metody ekscerpcji obiektów lingwistycznych (przy zastosowaniu korpusu) może posłużyć badanie zjawiska rozszczepienia argumentu propozycjonalnego w zdaniach z czasownikami emotywnymi. Analiza korpusowa wykazuje, że we współczesnych tekstach, choć sporadycznie, są spotykane konstrukcje, w których argument propozycjonalny jest reprezentowany przez dwa zależne od predykatu głównego człony (N_{Nom} + *как* N_{Nom}):

(34) ***Он*** *раздражает меня* ***как факт****, внезапно нарушивший наш покой своим нежданным присутствием (Нина Щербак).*
(35) *Рисовать я отчего-то не люблю, особенно меня раздражает* ***белый карандаш как невероятная глупость****, но зато обожаю играть в театр (Михаил Шишкин).*

Na to, że grupy wyrazowe [*он, как факт*] i [*белый карандаш, как глупость*] reprezentują ten sam aktant sentencjonalny, wskazuje możliwość transformacji:

(36) *Меня раздражает то, что он представляет собой факт, внезапно нарушивший наш покой…*
(37) *Особенно меня раздражает то, что белый карандаш представляет собой невероятную глупость.*

Gdy powstaje kwestia walencji innego czasownika tej grupy semantycznej – *удивлять,* o bliskim w zasadzie znaczeniu (por. też polskie odpowiedniki: *drażnić* vs *dziwić*), korpus ujawnia kilka przykładów pochodzących z tekstów XIX w.:

(38) *Но и между произведениями древних, не большая ли часть удивляют нас как памятники высокого духа, а не как образцы совершенного вкуса (Василий Жуковский).*
(39) *Как люди, отжившие свой век, они удивляли и забавляли нас своей оригинальностью и разными причудами (Федор Буслаев).*
(40) *Как иностранец, он удивляет князя Владимира и его княгиню заморскими подарками (Федор Буслаев).*
(41) *Как общественный симптом, — продолжал он, — это меня нисколько не удивляет (Петр Боборыкин).*

Powstaje dylemat: czy te nieliczne egzemplarze zdaniowe, pochodzące z tekstów sprzed 150–200 lat, dają podstawę do twierdzenia, że czasownikowi *удивлять*, podobnie jak innym czasownikom w klasie *verba sentiendi*, we współczesnym języku rosyjskim przysługuje ta cecha walencyjna? Intuicja językowa podpowiada nam, że konstrukcje takie możliwe są także dziś, ale żeby je udokumentować, musimy sięgnąć do innych źródeł. Tak więc w wyszukiwarce internetowej Rambler.ru odnajdujemy interesujące nas zdania z czasownikiem *удивлять*, i wcale nie można twierdzić, że są to zdania pojedyncze, por.:

(42) ***Он*** *меня удивляет, скорее,* ***как чудо Природы****.*
(43) ***Он*** *заслуживает уважение, как специалист, но порой удивляет* ***как человек****.*
(44) *Меня это удивляет как факт и не удивляет как результат их работы.*
(45) ***Клонирование*** *является одним из распространенных видов финансового мошенничества и уже мало кого удивляет* ***как явление****.*
(46) ***Появление*** *на футбольном поле во время игры животных уже мало кого удивляет* ***как явление****.*

(47) *Меня* ***он*** *удивляет* ***как явление****, а не в том смысле, чего от него можно ждать.*
(48) *Твоя энергетика меня всегда восхищала, а с годами, безусловно,* ***она*** *удивляет* ***как феномен****!*
(49) *Сочетание добра и милосердия в ней при высокой развитости этих чувств просто удивляет как феномен, парадокс.*
(50) *Так и подмывает спросить: если* ***«да»*** *удивляет* ***как нечто в высшей степени необычное****, то почему так поражает «нет»?*
(51) *Но* ***она*** *уже удивляет* ***как нечто из ряда вон выходящее****.*

Można więc z całą pewnością twierdzić, że rozszczepienie argumentu propozycjonalnego (przy zastosowaniu konstrukcji przyimkowo-rzeczownikowej *как* + N_{Nom}) w zdaniach z czasownikiem *удивлять/удивить* we współczesnym języku rosyjskim jest zjawiskiem udokumentowanym, a pod tym względem czasownik ten nie wyróżnia się na tle innych jednostek klasy *verba sentiendi*.

Intuicja czy też introspekcja, jak pisze W. D. Meurers (2005), stanowi nie tylko alternatywne źródło informacji lingwistycznej – w stosunku do źródeł tekstowych, zamanifestowanych, lecz także czynnik określający kierunki wyszukiwania danych w korpusach internetowych[5]. Innymi słowy, badanie korpusu (choć to twierdzenie może brzmieć banalnie) zakłada istnienie wstępnej, posiadanej przez podmiot wiedzy językowej. Wyszukiwanie w korpusie, którego celem jest m.in. zdobycie informacji na temat struktury i funkcjonowania systemu języka, jest wobec tego zapośredniczone poprzez wiedzę o systemie języka – wiedzę natywną lub lingwistycznie wyprofilowaną. Jest to szczególnie zauważalne w tych sytuacjach, gdy korpus nie dysponuje narzędziami wyszukiwania zautomatyzowanego, opartego na algorytmach. Na przykład A. Mustajoki (2006: 53 i n.) rozważa możliwości badania korpusowego rosyjskich zdań bezpodmiotowych typu

(52) *Лодку унесло ветром.*

Z uwagi na to, że korpus Integrum, z którego korzysta badacz, nie dysponuje możliwością gramatycznej selekcji materiału, należy zastosować jakieś inne narzędzia zawężenia pola wyszukiwania. Mustajoki (przywołując własną kompetencję językową) zwraca uwagę na fakt, że w zdaniach bezpodmiotowych występują określone typy czasowników oraz rzeczowników w pozycji kauzatora. Wobec tego badacz ogranicza wyszukiwanie do wybranych 180 jednostek w formie 3. osoby lp czasu przeszłego: *унесло, убило,*

[5] S. Marzo, K. Heylen i G. de Sutter G. (2012: 2 i n.) piszą, że badania korpusowe powinny opierać się na konfiguracji dwóch ujęć: kwantytatywnego i kwalitatywnego. O wsparciu analizy ilościowej przez analizę jakościową w badaniach korpusowych czytamy także w zamieszczonym w tym samym tomie artykule A. Fetzer i M. Johansson (2012: 90).

зарезало, ударило i in. Stosując tę metodę, udało się ograniczyć liczbę konstrukcji do rozsądnej skali. W przypadkach, gdy ogólna liczba wystąpień niektórych jednostek (np. *унесло, убило*) była zbyt wysoka, dokonano ich losowej ekscerpcji. W ten sposób zgromadzono podkorpus składający się z 2304 zdań, które zostały przeniesione do macierzy programu Excel i poddane dalszej analizie lingwistycznej i statystycznej.

W jeszcze większym stopniu wstępna wiedza językowa jest wymagana w badaniach pragmalingwistycznych. Niech za przykład posłuży fragment dialogu z komunikacji potocznej:

(53) – *Gdzie idziesz?*
– *Wracam za chwilę.*

Jest to dość typowy przykład tzw. pośrednich aktów mowy: z semantycznego punktu widzenia druga replika nie koresponduje z pytaniem, ale ten, kto odpowiada (*B*), nie uwzględnia dosłownego sensu pytania, lecz jego implikacje (a z psychologicznego punktu widzenia – motyw czy też bodziec wewnętrzny aktu mowy). Otóż *B* zdaje sobie sprawę, że pytanie implikuje niepokój pytającego (*A*) w związku ze stanem rzeczy, który może zostać spowodowany przez odejście *B*. Przy uwzględnieniu podtekstu powyższy dialog można przedstawić następująco:

(54) – *Gdzie idziesz? [Niepokoję się w związku z tym, że odchodzisz i będzie mi ciebie brak.]*
– *[Nie musisz się niepokoić, bo] Wracam za chwilę.*

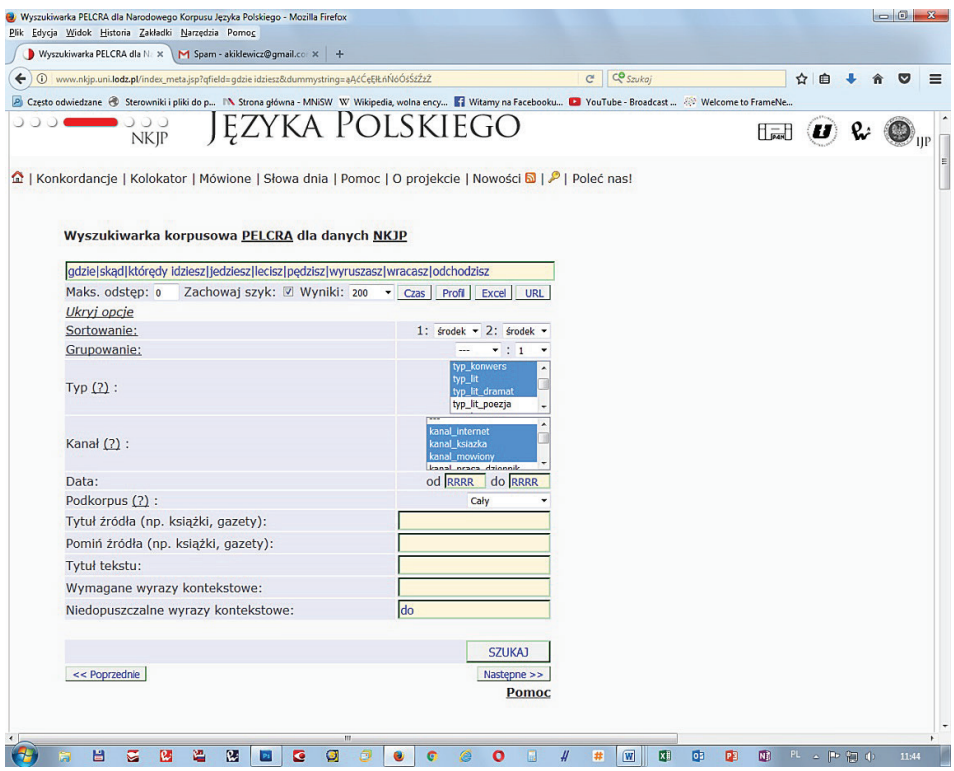
Wyobraźmy sobie sytuację badacza, który jest zainteresowany badaniem korpusowym faktów tego rodzaju. Przede wszystkim powinien zakładać, że możliwe są inne pośrednie akty mowy w sytuacjach dialogowych, w których pierwsza replika brzmi tak samo. Wobec tego najrozsądniej zacząć od sprawdzenia w korpusie fragmentów tekstowych, zawierających frazę *gdzie idziesz*. Wyszukiwarka PELCRA wyświetla 209 takich akapitów, z tym że nie wszystkie mają charakter dialogowy. Aby zawęzić obszar wyszukiwania, można skorzystać z opcji zaawansowanych. Możemy dookreślić typ i kanał tekstów, celowo unikając np. konkordancji tekstów urzędowych czy informacyjnych. Można też celowo zablokować konteksty z przyimkiem *do*. W ten sposób otrzymujemy 33 przykłady, z których, jak wykazuje analiza, siedem odpowiada profilowi wyszukiwania, por.:

(55) – *Gdzie idziesz ?…*
– *Poszukam twego pana – odparł (Bolesław Prus).*
(56) *BARTODZIEJ – Gdzie idziesz?*
ANATOL – Mam je w biurku. Ja już mam biurko, wiesz? (Sławomir Mrożek).
(57) – *Gdzie idziesz?*
– *Zaraz przyjdę – odpowiedziałem (Ireneusz Iredyński).*

(58) – *Gdzie idziesz?*
– *To wszystko zipnąć mi nie daje – powiedział z troską Szerucki (Leopold Buczkowski).*
(59) – *Gdzie idziesz?*
– *Zadzwonić. Chcę mieć to wszystko za sobą (Jacek Głębski).*

Jak widzimy na tych nielicznych przykładach, najczęściej – w przypadku pośrednich aktów mowy – odpowiedź na pytanie o kierunek zawiera informację o celu czynności. Gdybyśmy chcieli się przekonać, że jest to regularna cecha pośrednich aktów mowy w sytuacjach zapytania, należałoby rozszerzyć obszar poszukiwania. Można byłoby dodać dokumenty pochodzące z internetu, a przede wszystkim skorzystać z wiedzy, że w skład zdania pytającego wchodzi zaimek pytający oraz czasownik wskazujący na czynność fizyczną, skutkującą zmianą stanu rzeczy. Są to nie tylko czasowniki ruchu, ale także np. czasowniki kauzatywne. Wobec tego zapytanie może (przykładowo) wyglądać następująco: gdzie | skąd | którędy idziesz | jedziesz | lecisz | pędzisz | wyruszasz | wracasz | odchodzisz

W wyniku wyszukiwania otrzymujemy 168 przykładów, które można skopiować do pliku Excel i poddać dalszej analizie.



4

Badania korpusowe, jak zaznaczają specjaliści (zob. Biber/Fitzmaurice/Reppen 2002; Gries 2009; Mustajoki 2006; Płungian 2017), są szczególnie przydatne przy określeniu częstości występowania jednostek lub ich połączeń w mowie. Ujęcie statystyczne jest stosowane także przy badaniu systemu języka, np. gdy zgromadzona na podstawie danych korpusowych informacja o łączliwości syntaktycznej jednostek określonej klasy jest interpretowana ze względu na liczbę jednostek dopuszczających każdy typ łączliwości. Na przykład z moich badań wynika, że rosyjskie czasowniki mentalne w różnym stopniu dopuszczają bardziej lub mniej dyskretne realizacje argumentu propozycjonalnego: 72,2% czasowników dopuszcza w tej pozycji zdanie zależne, 87,0% czasowników – wykładnik predykatu propozycji zależnej (w formie rzeczownika abstrakcyjnego lub bezokolicznika), 66,2% – rzeczownik przedmiotowy (jako wykładnik argumentu propozycji zależnej), 48,9% – zero składniowe.

W badaniach składniowych, jak można było już obserwować, sięganie do korpusów spełnia także inny cel: rejestrację związków składniowych, weryfikację dopuszczalności pewnych form gramatycznych w pozycjach konotowanych przez czasownik. W związku z tym powstaje pytanie o interpretację tych przypadków, gdy korpus nie odnotowuje prognozowanych form (występujących np. w konstrukcjach z innymi jednostkami tej samej klasy). Pierwsze wytłumaczenie może być techniczne: korpus jest za mały albo stylistycznie wyprofilowany tak, że nie reprezentuje w wystarczającym wymiarze wszystkich stylów funkcjonalnych. Na przykład Narodowy Korpus Języka Rosyjskiego zawiera 40% dokumentów pochodzących z tekstów literatury artystycznej, a w ogóle w korpusie przeważają teksty artystyczne i publicystyczne. Właśnie dlatego korpus nie zaspokaja potrzeb użytkowników, którym zależy na wyszukaniu konstrukcji językowych, charakterystycznych np. dla stylu urzędowo-oficjalnego lub technicznego.

Braki w korpusie mogą istnieć także z powodów semantycznych czy też ontologicznych. System językowy jest ambiwalentny nie tylko w odniesieniu do sytuacji użycia (tzn. występuje w sytuacjach oficjalnych, nieoficjalnych, w kontaktach trwałych, nietrwałych itd., więcej o „społecznych rolach językowych" zob.: Grabias 2004: 270 i n.), lecz także do sytuacji referencyjnych. To znaczy, że obiektom nominacji służą zarówno fakty rzeczywistości, jak i elementy tzw. możliwych światów, np. ludzkiej wyobraźni. W dużym stopniu do powstania takich, alternatywnych w stosunku do rzeczywistości, kontrfaktycznych nominacji przyczynia się system języka – za sprawą przynależności jednostek do klas i zachodzących w ich obrębie procesów

analogii. Tak więc skoro rzeczowniki *słońce* i *pole* należą do jednej klasy paradygmatycznej, system języka przewiduje w obydwu przypadkach formy liczby mnogiej: *to pole // te pola, to słońce // te słońca.* Jest oczywiste, że rzeczownik *słońca* w lm ma charakter kontrfaktyczny i, teoretycznie rzecz biorąc, nie ma prawa bytu. Zdarza się jednak, że sporadycznie spotykamy takie użycia, jak np. formę *słońcom* w powieści Stanisława Lema „Solaris":

(60) […] *I z nie zmniejszoną chyżością pomknie ku podwójnym słońcom Solaris.*

Zgodnie z lingwistyczną tradycją jednostki takie uznaje się za potencjalne. W związku z tym M. Bańko (2001: 58 i n.) rozważa kwestię ograniczonej odmienności niektórych (tzw. trzecioosobowych) czasowników. Tak więc w przypadku *fermentować* można zastanawiać się, czy ten wyraz występuje w formach 1. i 2. osoby: *fermentuję, fermentujesz.* Bańko nie wyklucza takiej możliwości, pisząc:

> Chcąc przekonać kogoś, że *fermentować* może wystąpić w formach różnych osób, ucieklibyśmy się zapewne do argumentów składniowych. Skoro można powiedzieć: *Sok fermentuje,* to jakiś szalony (albo dowcipny) poeta, pisząc odę do soku, mógłby użyć formy drugiej osoby *fermentujesz*. Inny szalony poeta mógłby napisać wiersz z sokiem jako „ja" lirycznym, co uzasadniałoby użycie formy *fermentuję*. Z tego, że dany czasownik łączy się z podmiotem w mianowniku, wynika więc, że można odmieniać go przez osoby (2001: 58).

Jak widzimy, badacz-językoznawca, chcąc udowodnić fakty systemu języka, posługuje się własną intuicją językową, nie wykluczając, że takie, potencjalne fakty mogą być udokumentowane, ale w związku z tym pisze:

> W gruncie rzeczy jednak udokumentowane przykłady nie są potrzebne. Nikt spośród językoznawców nie neguje istnienia form potencjalnych, różnice zdań dotyczą tylko tego, w jakim stopniu ma uwzględniać fakty systemowe, a w jakim stopniu uzus (tamże; zob. też Łazutkina 2014: 51).

W dużym stopniu nawiązuje to do socjologicznej teorii É. Durkheima, a mianowicie jego ujęcia faktu społecznego:

> Zbiorowy zwyczaj […] na mocy przywileju, pozbawionego odpowiednika w biologii, zostaje raz na zawsze wyrażony w formule, która przechodzi z ust do ust, przekazywana jest przez wychowanie, utrwala się nawet na piśmie. Taki jest początek i taka jest natura przepisów moralności i prawa, aforyzmów i przysłów ludowych, kodeksów dobrego smaku, stwarzanych przez szkoły literackie, artykułów wiary, w jakich sekty religijne lub polityczne streszczają swoje wierzenia itd. Żadna z owych reguł nie wyczerpuje się w zastosowaniach, jakich dokonują poszczególni ludzie, ponieważ mogą one istnieć nawet wtedy, gdy nikt ich w danym momencie nie stosuje (Durkheim 2007: 34 i n.).

W podobny sposób w językoznawstwie istnienie faktów systemowych (należących do kodu) w pewnym stopniu jest niezależne od performancji,

która ma charakter realny/aktualny lub potencjalny. Na przykład forma gramatyczna *miesiączkował* (rm) z uzualnego punktu widzenia wydaje się niemożliwa, ale z punktu widzenia systemu języka jest tak samo (czyli za sprawą opozycji gramatycznej) prawomocna, jak forma *miesiączkowała* (rż). Forma rodzaju męskiego nie została udokumentowana w korpusie internetowym ani w innych źródłach tekstowych, ale spotykamy ją na stronie internetowej http://www.rymy.eu jako jeden z rymów do wyrazu *zredukował*. Poza tym jest ona możliwa w użyciach metaforycznych lub w zdaniach warunkowych: *Gdyby facet miesiączkował…* Skoro w obrębie danej klasy jednostek istnieją regularnie realizowane opozycje, np. <1. osoba // 2. osoba // 3. osoba>, odmienność czasowników należy uznać za fakt niezależny od performancji. W tym sensie i w odniesieniu do takich faktów należy przyznać rację L. Hjelmslevowi, który pisał, że teoria lingwistyczna jest niezależna od doświadczenia (czyli od działalności mownej) ani nie zawiera postulatu o istnieniu obiektów lingwistycznych. Najważniejsze jest, aby teoria nie zawierała twierdzeń sprzecznych ze sobą. W jego słynnych *Prolegomenach…* m.in. czytamy:

> Zakłada się, że teoria nie tylko daje nam narzędzie poznania określonego obiektu. Powinna ona zostać zbudowana w taki sposób, żeby umożliwić poznanie wszystkich wyobrażanych obiektów tego samego rodzaju. […] Teoria uzbraja nas na wypadek spotkania nie tylko z obiektami, które spotkaliśmy wcześniej, lecz także z każdym możliwym obiektem (Hjelmslev 2006: 41 i n.; tłumaczenie moje – A. K.).

W składni takie, algorytmiczne ujęcie nie jest jednak usprawiedliwione. Owszem związki składniowe mają regularny charakter[6], ale walencji czasowników nie można zaprogramować w sposób aksjomatyczny. Zgodność semantyki i gramatyki ma charakter względny, probabilistyczny, dlatego nie wszystkie jednostki, należące do tej samej klasy semantycznej, mają jednakowe właściwości dystrybucyjne. Chcąc udokumentować poszczególne występowania pewnych związków składniowych, sięgamy do źródeł empirycznych (które pełnią w tym przypadku funkcję rozstrzygającą), a korzystając z korpusów internetowych, możemy dodatkowo zbadać ich funkcjonalność, w szczególności częstość używania w mowie. W sytuacji, gdy korpus nie potwierdza pewnych form syntaktycznych, zwracamy się do intuicji językowej – własnej lub innych użytkowników języka, np. poprzez prowadzenie ankietowania. Fakty językowe, które postulujemy w taki sposób, mają specyficzny charakter: można je traktować jako potencjalizmy (więcej o tym pojęciu: Barkowicz 2015: 266).

[6] Na przykład czasownik osobowy otwiera pozycję dla rzeczownika (grupy nominalnej) w mianowniku.

Podobnie jak w fizyce kwantowej, w przypadku potencjalizmów można stosować (choć oczywiście w innym sensie i celu) pojęcie superpozycji stanów: jednostki lub kompozycje potencjalne istnieją i nie istnieją jednocześnie. Mimo że nie można odwołać się do udokumentowanych faktów ich użycia w korpusach lub innych źródłach materialnych (w sensie fizycznym nie istnieją), to jednak stanowią one elementy zintegrowane z systemem języka, gdyż domykają regularne opozycje elementów i w tym sensie są zaprogramowane w systemie. W zależności od tego, jak pisze Bańko (zob. wyżej), czy opis lingwistyczny jest ukierunkowany na system, czy na jego mowną realizację (performancję), potencjalizmy będą traktowane w taki lub inny sposób, tzn. stanowić obiekty lingwistyczne lub nie.

***

Informacja korpusowa stanowi ważny element współczesnych badań lingwistycznych zarówno w zakresie lingwistyki języka, jak i lingwistyki mowy. W ten sposób potwierdzamy udokumentowanie faktów językowych (kompatybilnych z określoną teorią lingwistyczną), jak również ułatwiamy, doskonalimy badanie funkcjonalnych właściwości jednostek językowych, w szczególności badanie frekwencji. Z praktyki badawczej jednak wynika przekonanie o pewnych ograniczeniach analizy korpusowej. Korpusy okazują się niewystarczające pod względem reprezentacyjności, w szczególności jeśli chodzi o fakty o niskiej frekwencji, a także nacechowane pod względem stylistycznym. Wyszukiwanie niektórych połączeń jednostek nie kończy się sukcesem, choć okazuje się, że są one udokumentowane w innych źródłach, np. w wyszukiwarkach internetowych. Można z tego wysnuć następujące wnioski. Po pierwsze, korpusy internetowe wymagają doskonalenia zarówno pod względem objętości, jak i pod względem reprezentacji różnych stylów funkcjonalnych języka, szczególnie wyspecjalizowanych, najbardziej oddalonych od stylu potocznego.

Po drugie, potwierdzając słuszność napisanych przez prof. Polańskiego czterdzieści lat temu słów, należy uznać konieczność dywersyfikacji źródeł oraz stosowania zasady komplementarności: jedne źródła powinny uzupełniać inne (łącznie z introspekcją). Tak jest np. przyjęte w praktyce badań etnolingwistycznych grupy prowadzonej przez prof. J. Bartmińskiego.

Odwołanie się do introspekcji (czy intuicji językowej) może wydawać się paradoksalne – na tle faktu, że korpusy internetowe zawierają setki tysięcy, nawet miliony wyrazów, a średniostatystyczny użytkownik języka zna ich zaledwie kilkadziesiąt tysięcy. Taki stan rzeczy – po trzecie – jest napomnieniem, że języka nie można utożsamić z żadną formą jego manifestacji. Podobnie jak – według Protagorasa – człowiek jest miarą wszystkich rzeczy,

jest on też miarą tożsamości systemu języka. Fakty językowe, podobnie jak każde inne, stają się faktami w kontekście ludzkich nastawień, wstępnych ustaleń, założeń sądzeniowych. Gdy angielski poseł lord McCartney w XVIII w. wręczył w prezencie bogdychanowi europejskie ryciny, przedstawiające członków angielskiej rodziny królewskiej, Chińczyków zaskoczyły światłocienie, które zinterpretowano jako ślady pobicia. Tak w tej sytuacji, jak i w językoznawstwie czy innej nauce istotna jest założeniowość empiryczna: „Założenia wprowadza się w ten sposób, aby zapewniały teoretyczne podstawy obserwowanym faktom" (Boczar 2000: 187). Nie jest to myśl nowa, jednak w sytuacji popularnego dziś panempiryzmu i instrumentalizacji badań (zob. Dębowski 2007b: 120) warto przypomnieć o stymulującej roli założeniowości i epistemologizacji w nauce[7].

Lingwistyka korpusowa wykorzystuje nowoczesne narzędzia badawcze, jednak otrzymywana w ten sposób informacja wymaga interpretacji lingwistycznej przy zastosowaniu tradycyjnych, bardziej lub mniej skonwencjonalizowanych metod naukowych. Pamiętajmy przysłowie: „Nowego kłamstwa słucha się chętniej niż starej prawdy" i starajmy się utrzymać konsensus między tradycją a nowatorstwem.

Literatura

Arashonkava H. U./Lemciuhova V. P. [= Арашонкава Г. У./Лемцюгова В. П.] (1991): *Кіраванне ў беларускай і рускай мовах. Слоўнік-даведнік*. Мінск.

Bańko M. (2001): *Co jest niewłaściwego w czasownikach niewłaściwych?* [W:] Chruszczyński W. (red.): *Nie bez znaczenia… Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15 000 dni pracy naukowej*. Białystok. 55–65.

Boczar J. (2000): *Metodologiczne implikacje założeniowości w nauce.* [W:] *Folia Philosophica.* 18, 183–189.

Barkowicz A. A. [= Баркович А. А.] (2015): *Модель развития языковых новаций в контексте компьютерно-опосредованной коммуникации*. [B:] *Вестник Тихоокеанского государственного университета.* 4/39, 263–272.

Biber D., Fitzmaurice S. M., Reppen R. (ed.) (2002): *Using Corpora to Explore Linguistic Variation*. Amsterdam.

Dębowski J. (2007a): *Bezzałożeniowy humanista. Rzeczywistość czy utopia?* [W:] Kowalewski J., Piasek W. (red.): *Zaangażowanie czy izolacja?* Olsztyn, 3140.

Dębowski J. (2007b): *Wolność nauki a zasada bezzałożeniowości. Między „epistemologicznym anarchizmem" a „epistemologicznym restrykcjonizmem"*. [W:] Szulakiewicz M., Karpus Z. (red.): *Wolność w epoce poszukiwań*. Toruń, 109–122.

Dębowski J. (2014): *Zasada bezzałożeniowości*. [W:] Plotka W. (red.): *Wprowadzenie do fenomenologii. Interpretacje, zastosowania, problemy*. II. Warszawa, 7–36.

Durkheim É. (2007): *Zasady metody socjologicznej*. Przekł. J. Szacki. Warszawa.

[7] Oczywiście nie można lekceważyć faktu, że zasada bezzałożeniowości także ma rację bytu. Argumenty na jej korzyść przywołano m.in. w publikacjach J. Dębowskiego (2007a; 2007b; 2014).

Fetzer A., Johansson M. (2012): *Cognitive verbs in context. A contrastive analysis of English and French argumentative discourse*. [In:] Sutter G. de, Heylen K., Marzo S. (eds.): *Corpus Studies in Contrastive Linguistics*. Amsterdam. 89–117.

Grabias S. (2004): *Język w zachowaniach społecznych*. Lublin.

Gries S. T. (2009): *What is Corpus Linguistics?* [In:] *Language and Linguistics Compass*. 3, 1–17.

Hjelmslev L. [= Ельмслев, Л. (2006): *Пролегомены к теории языка*. Москва.

Karolak S. (1984): *Składnia wyrażeń predykatywnych*. [W:] Topolińska Z. (red.): *Gramatyka współczesnego języka polskiego. Składnia*. Warszawa. 11–212.

Karolak S. (2002): *Podstawowe struktury składniowe języka polskiego*. Warszawa.

Kiklewicz A. (2009): *Źródła w językoznawstwie*. [In:] *Humanistyka i Przyrodoznawstwo*. XV, 205–216.

Kiklewicz A. (2016): *Синтаксические характеристики русских и польских интернет-форумов* (*на материале простых и сложных предложений с ментальными предикатами*). [In:] Tosović B./Wonisch A. (Hrsg.): *Interaktion von Internet und Stilistik, Internet und Stil*. Graz. 93–110.

Korytkowska M. (1992): *Typy pozycji predykatowo-argumentowych*. Warszawa [Gramatyka konfrontatywna bułgarsko-polska, t. 5].

Korytkowska M., Kiklewicz A. (2016): *Opis właściwości walencyjnych czasowników na podstawie teorii składni eksplikacyjnej – problemy konfrontatywne i leksykograficzne* (*na przykładzie języka bułgarskiego, polskiego i rosyjskiego*). [In:] Skwarska K., Kaczmarska E. (red.): *Výzkum slovesné valency ve slovanských zemích*. Praha, 291–304.

Korytkowska M., Małdżijewa, W. (2002): *Od zdania złożonego do zdania pojedynczego. Nominalizacja argumentu propozycjonalnego w języku polskim i bułgarskim*. Toruń.

Łazutkina E. M. [= Лазуткина Е. М.] (2012): *Словарь грамматической сочетаемости слов русского языка*. Москва.

Łazutkina E. M. [= Лазуткина Е. М.] (2014): *Вариативность в системе грамматических норм современного русского литературного языка*. [В:] *Труды Института русского языка им. В. В. Виноградова*. 1, 223–329.

Marzo S., Heylen K., Sutter G. de (2012): *Developments in Corpus-based Contrastive Linguistics*. [In:] Sutter G. de, Heylen K., Marzo S. (eds.): *Corpus Studies in Contrastive Linguistics*. Amsterdam. 1–7.

Meurers W. D. (2005): *On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of German*. [In:] *Lingua*. 115/11, 1619–1639.

Mustajoki A. (2006): *The Integrum Database as a Powerful Tool in Research on Contemporary Russian*. [In:] Никипорец-Такигава Г. (ред.): *Integrum: точные методы и гуманитарные науки*. Москва. 50–75.

Płungian W. A. [= Плунгян В. А.] (2005): *Зачем мы делаем национальный корпус русского языка?* [In:] *Отечественные записки*. 2, 296–308.

Polański K. (1980): *Słownik syntaktyczno-generatywny czasowników polskich. 1/A-M*. Wrocław etc.

Przepiórkowski A., Bańko M., Górski R. et al. (2012): *Narodowy korpus języka polskiego*. Warszawa.

Topolińska Z. (red.) (1984): *Gramatyka współczesnego języka polskiego. Składnia*. Warszawa.

Urmancew J. A. [= Урманцев Ю. А.] (1978): *Начала общей теории систем*. [In:] Горский Д. П. (ред.): *Системный анализ и научное знание*. Москва. 7–41.

Wawrzyńczyk J., Wierzchoń P. (2016): *300 tysięcy polskich słów. Indeks a fronte*. Warszawa.

Summary

The aim of present analysis is to show the use of Internet corpus in syntactic studies of Slavic languages (especially Russian). Corpus analysis is treated as a research tool, useful in describing linguistic system as well as linguistic activity. The information coming from the corpus allows to determine the frequency of occurrence of units and their combinations in texts as well as the regularity of occurrences of features/properties in the paradigmatic classes. Corpus analysis also provides the ability to verify whether a particular valence property is characteristic for a given word or not. The author shows that the use of Internet corpus in the syntactic research has its limitations. In the case of frequent phenomena, corpus analysis is effective, but does not always allow to document less typical phenomena (for example occasional and potential combinations of tokens). One of the author's conclusions is that corpus analysis should be configured with introspection and qualitative analysis.