TRAITEMENT AUTOMATIQUE DES LANGUES

ROCZNIKI HUMANISTYCZNE
Tom LXI, zeszyt 8 – 2013

GRAŻYNA VETULANI

PROBLEMY I KORZYŚCI WYNIKAJĄCE Z AUTOMATYCZNEGO PRZETWARZANIA KORPUSÓW – NA PRZYKŁADZIE BADAŃ Z ZAKRESU PREDYKACJI RZECZOWNIKOWEJ W JĘZYKU POLSKIM

DRAWBACKS AND ADVANTAGES OF THE COMPUTER CORPORA PROCESSING. CASE STUDY OF NOMINAL PREDICATION IN POLISH

A b s t r a c t

This paper reports on our work related to nominal predication in Polish and exploring electronic corpora with help of text processing tools. Various aspects and challenges related with the applied methodology are presented. Despite encountered problems, nowadays, it is practically impossible to imagine solutions ignoring advantages of corpus linguistics. In fact this methodology appeared very efficient. In a relatively short time we developed an application-oriented dictionary of Polish predicative nouns and now we continue to extend it within the same paradigm.

Key words: corpus linguistics, text processing, nominal predication.

0. WSTĘP

Pozyskiwanie jednostek do badań lingwistycznych z korpusów językowych niesie za sobą cały szereg zalet i wad bez względu na fakt, czy odbywa się metodą tradycyjną (poprzez analizę naoczną tekstu pisanego lub treści przyhasłowych zawartych w słownikach), czy za pomocą wyspecjalizowa-

Prof. dr hab. GRAŻYNA VETULANI – Instytut Filologii Romańskiej Uniwersytetu im. Adama Mickiewicza w Poznaniu; adres do korespondencji – e-mail: gravet@amu.edu.pl

nych programów komputerowych, wspomagających lingwistę w przeszukiwaniu specjalnie przygotowanych do tego celu korpusów elektronicznych. W obydwu wypadkach badanie języka odbywa się poprzez analizę materiału dostępnego empirycznie, co stawia lingwistykę pośród innych nauk odwołujących się do obserwacji faktów i zjawisk. Dla lingwisty rzeczywistością obserwowaną jest zgromadzona w korpusach „materia", czyli konkretne realizacje językowe (wykonania na poziomie *parole*), takie jak: słowa, krótsze i dłuższe fragmenty wypowiedzi, zdania, teksty pisane, ale także nagrania mowy pochodzące z różnych rejestrów i typów dyskursu.

W niniejszej pracy pragniemy zwrócić uwagę na problemy i korzyści napotkane podczas badań prowadzonych od szeregu już lat w zakresie predykacji rzeczownikowej języka polskiego przy wykorzystaniu zarówno korpusu tradycyjnego (na pierwszym etapie badań, we wczesnych latach 90. ubiegłego wieku, w wyniku których powstało pierwsze opracowanie z tego zakresu pt. *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych*, G. Vetulani, 2000), jak i korpusów elektronicznych oraz odpowiednio przygotowanych do ich przetwarzania narzędzi informatycznych (od końca lat 90. do chwili obecnej, tj. na drugim i trzecim (obecnym) etapie[1]).

1. DYNAMICZNY ROZWÓJ LINGWISTYKI KORPUSOWEJ

Z uwagi na fakt, że typ, jakość, a także wielkość materiału obserwacyjnego mają zasadniczy wpływ na uzyskiwane wyniki (a co za tym idzie – na wnioski ogólne w odniesieniu do danego systemu językowego), jednym

[1] Główne prace na etapie drugim, podczas których nastąpiło wykorzystanie zinformatyzowanych korpusów języka polskiego oraz programów komputerowych do ich obróbki, były prowadzone w ramach projektu Ministerstwa Nauki i Szkolnictwa Wyższego Nr R00 02802: POLSKA PLATFORMA BEZPIECZEŃSTWA PUBLICZNEGO pt. „Technologie przetwarzania tekstu polskiego zorientowane na potrzeby bezpieczeństwa publicznego; komunikacja człowieka z systemem informatycznym w warunkach kryzysowych przy użyciu języka naturalnego". Projekt był realizowany w UAM w Poznaniu okresie od 15.12.2006 do 28.02.2010 pod kierownictwem Zygmunta Vetulaniego. Prace bieżące (etap trzeci) odbywają się w ramach projektu pt. „Rozbudowa zasobów cyfrowych języka polskiego w zakresie słowników walencyjnych w kierunku leksykonu-gramatyki zorientowana na potrzeby zastosowań informatycznych w humanistyce", który jest finansowany ze środków Narodowego Programu Rozwoju Humanistyki (MNiSW Nr 0022/FNiTP/H11/80/2011); czas trwania projektu: od 1.02.2012 do 31.01.2015; kierownik projektu: Grażyna Vetulani.

z pierwszych problemów, jakie należy rozwiązać w momencie przystępowania do badań, są kwestie związane z wyborem odpowiedniego materiału źródłowego, czyli *korpusu*. Stosowane metody, w tym wykorzystywane korpusy, idą zawsze w parze z określonymi w danym czasie (epoce) możliwościami technicznymi[2].

Rzeczą oczywistą jest, że obecnie obserwacja zjawisk językowych różni się w swoim charakterze od tej, która miała miejsce dawniej. Bez trudu da się zauważyć – przynajmniej jeśli chodzi o języki, dla których stworzono już nowoczesne i reprezentatywne korpusy – że dostępność do faktów językowych jest dzisiaj dużo większa niż dawniej. Zamiana korpusów tradycyjnych na nowe, nieporównanie większe (aktualnie wręcz gigantyczne) i odpowiednio przygotowane do automatycznego przeszukiwania, stworzyła dogodne możliwości obserwowania zjawisk językowych i tworzenia na tej podstawie teorii (np. modeli odnośnie do systemu), powodując zarazem przyspieszenie badań i – przy okazji – rozwój nowych kierunków w lingwistyce bez wyhamowywania starszych (por. uwagi w G. Vetulani, 2012: 48-50).

Badania korpusowe są dziś niezwykle zaawansowane. Dzięki nim wnosi się do nauki wiedzę na podstawie zarejestrowanych zaświadczeń (uzusu), a nie – jak to bywało niekiedy w przeszłości – poprzez preparowane przykłady językowe. Szybki rozwój badań opartych na autentycznych realizacjach językowych, wsparty rozwojem nowych technologii[3], pociągnął za sobą wykształcenie się odrębnej gałęzi, jaką jest *lingwistyka korpusowa*. Gromadzenie i tworzenie zasobów językowych zaowocowało zatem powstaniem niezależnego nurtu badań, który – jeśli sądzić po realizowanych obecnie pracach z zakresu szeroko rozumianej technologii języka – będzie nadal rozwijał się w sposób dynamiczny. Zauważmy, że rozwój ten idzie w parze z powszechną cyfryzacją, którą objęte są różnego rodzaju zbiory, w tym stare i nowe zasoby biblioteczne dostarczające tekstów do badań. Odbywająca się na naszych oczach wszechobecna cyfryzacja jest, z jednej strony, wynikiem postępu technologicznego, z drugiej zaś polityki Komisji Europejskiej. Z inicjatywy tej ostatniej powstaje Europejska Biblioteka Cyfrowa w celu zachowania dorobku kulturowego Unii[4].

---

[2] Por. ponadto uwagi na temat stosunku językoznawców do tekstów źródłowych w aspekcie historycznym w pracy pt. *Z pogranicza leksykografii i językoznawstwa* (M. Bańko, 2001: 26-28).

[3] W Europie rozwój ten następuje w dużej mierze dzięki nakładom Komisji Europejskiej, która od wielu juz lat finansuje wiele projektów z zakresu inżynierii języka.

[4] Chodzi w szczególności o prace nad Europejską Biblioteką Cyfrową EUROPEANA, umożliwiającą zintegrowany dostęp do cyfrowych zbiorów europejskich bibliotek, archiwów i

1.1. KORPUSY CYFROWE NIEODZOWNYM ELEMENTEM WARSZTATU BADAWCZEGO

Aby prowadzić badania językowe szybko i skutecznie, należy dysponować cyfrowym zapisem tekstów (pisanych lub mówionych). I choć obecnie nie ma już trudności z uzyskaniem postaci cyfrowej wypowiedzi, gdyż w każdym momencie redagowania tekstu używa się do tego celu komputera, to samo zgromadzenie danych w pamięci maszyny nie stwarza jeszcze wystarczających warunków do prowadzenia badań. Muszą być one w sposób szczególny przygotowane, tj. „znakowane lingwistycznie" (A. Przepiórkowski, 2004: 5). Dlatego też elementy korpusu elektronicznego podlegają anotowaniu (indeksowaniu, tagowaniu) za pomocą właściwych znaczników, bez których ich dalsze przetwarzanie byłoby mniej efektywne. Typowe znaczniki dotyczą morfo-składni, ale mogą też odnosić się do semantyki, pragmatyki, prozodii itp.

Od pewnego już czasu, z różnym skutkiem dla różnych języków, buduje się odpowiednio zorganizowane bazy językowe z przeznaczeniem do badań naukowych. W zasadzie można uznać, że tworzenie i wykorzystywanie autentycznych korpusów stało się dziś obowiązujące i że są one nieodzownym elementem warsztatu językoznawczego. Prace prowadzi się zarówno na korpusach jednojęzycznych, jak i dwu- i wielojęzycznych, a ostatnio wiele się również czyni w zakresie budowania i wykorzystywania do badań porównawczych oraz traduktologicznych tzw. korpusów równoległych (cf. D. Bralewski, 2012; D. Vitas, 2008). Istnieją ponadto rozmaite inne, specjalistyczne bazy danych tekstowych: dziedzinowe, terminologiczne, tezaurusy itd.

Jeśli chodzi o badania nad predykacją rzeczownikową języka polskiego, podkreślmy, że po pierwszym etapie badań, kiedy to za źródło służył nam słownik tradycyjny (*Słownik Języka Polskiego*, M. Szymczak, red., 1983), niemal natychmiast, tj. z chwilą nastania w Polsce odpowiednich warunków technicznych, zmieniliśmy metody badawcze, aby kontynuować prace na korpusie zinformatyzowanym (wówczas nie mówiło się jeszcze o korpusie reprezentatywnym polszczyzny, czyli *korpusie narodowym*), wykorzystując odpowiednio przygotowane do jego przetwarzania programy komputerowe, uwzględniające specyfikę języka polskiego. Wszystkie narzędzia wykorzy-

muzeów. EUROPEANA została uruchomiona w 2008 r. Inicjatywa ta wynikała z potrzeby przyspieszenia przyjęcia wspólnych norm koniecznych do funkcjonowania wielojęzycznych bibliotek i archiwów *online* oraz potrzeby wsparcia badań naukowych w tej dziedzinie.

stywane podczas tych prac zostały wytworzone przez zespół informatyków pracujących pod kierownictwem Z. Vetulaniego na UAM w Poznaniu w ramach takich projektów, jak: projekt KBN, 1994-1996: POLEX – POLSKA LEKSYKALNA BAZA DANYCH oraz projekty Komisji Europejskiej: CEGLEX (CPERNICUS 1032, 1995-1996) oraz GRAMLEX (COPERNICUS 621, 1996-1998). Stworzenie programów komputerowych zdolnych do obróbki automatycznej tekstów języka polskiego, tj., między innymi, słowników elektronicznych języka polskiego (morfologicznych), programów indeksujących wyrazy w korpusie lub generujących pożądane konkordancje, było warunkiem koniecznym, gdyż takie narzędzia dla języka polskiego wówczas po prostu nie istniały (lub były dopiero w budowie), a ponadto udostępniona nam wersja korpusu była nieotagowana (G. Vetulani et al., 2006, 2007).

1.2. PIERWSZE KORPUSY A KORPUS JĘZYKA POLSKIEGO

Prace nad korpusami tekstowymi rozpoczęli w latach sześćdziesiątych H. Kucera i W.N. Francis, którzy stworzyli tzw. *Brown Corpus* (ok. 1 000 000 słów), dając początek lingwistyce komputerowej (M. Bauer & B. Aarts, 2000). Wśród największych istniejących obecnie baz językowych należy wymienić korpusy dla języków: angielskiego – *American National Corpus* (22 miliony słów), *British National Corpus* (100 milionów), the *Brown Corpus* "family" (oryginalny BC – ponad milion słów), *Oxford English Corpus* (2 miliardy), *PennTreebank* i wiele innych; niemieckiego – *German Reference Corpus* (4 miliardy); czeskiego – *Czech National Corpus* (1300 milionów); rosyjskiego – *Russian National Corpus* (350 milionów); hiszpańskiego – *Spanish Text Corpus* (660 milionów). Szereg korpusów zostało utworzonych w celach komercyjnych przez firmy oferujące usługi z zakresu inżynierii języka. Posród takich firm na uwagę zasługuje działająca od 2003 r. firma Lexical Computing Ltd. (A. Kilgariff), która dysponuje pakietem wielomilionowych (często powyżej 100 milionów słów) korpusów dla 52 języków. Lingwistyka korpusowa znajduje się w ciągłym rozwoju i nic nie wskazuje na to, by ta sytuacja miała ulec zmianie. Z jednej strony obserwuje się nieustanne dążenia do powiększania oraz ulepszenia istniejących już korpusów, z drugiej zaś do budowania korpusów nowych dla języków, które nie mają jeszcze (lub mają, lecz w ograniczonej formie) swojej narodowej reprezentacji tekstowej w postaci cyfrowej.

Podczas prac nad predykacją rzeczownikową języka polskiego, w okresie przechodzenia z metody tradycyjnej na komputerową (koniec lat 90.), największym – i jedynym wówczas dostępnym do badań naukowych – polskim korpusem był *Korpus IPI PAN*, liczący ok. 200 milionów słów. Od tamtego momentu korpus IPI PAN jest stale rozwijany, zmierzając do zapewnienia reprezentatywności zjawisk języka polskiego (nadmieńmy, że jego autorzy już teraz przypisują mu cechy korpusu narodowego). W analizie wykorzystano jednak jego wersję okrojoną, jedynie dostępną dla naszych prac, składającą się z ok. 80 milionów słów zawartych w tekstach beletrystycznych, naukowych, prasowych oraz licznych stenogramach sejmowych i senackich, będących zapisem dyskursu mówionego. Uwagi, które zamieszczamy poniżej, odnoszą się do wersji nam udostępnionej.

2. ZAUTOMATYZOWANA EKSPLORACJA KORPUSU

2.1. ROZWIJANIE SŁOWNIKA RZECZOWNIKÓW PREDYKATYWNYCH JĘZYKA POLSKIEGO

Zastosowanie metody wspomagania komputerowego miało na celu przyspieszenie badań w zwiazku z opisem predykatów nominalnych języka polskiego oraz stworzenie (dla tej klasy jednostek) słownika lepszej jakości w stosunku do pierwszego opracowania, które powstało po pierwszym etapie prac (G. Vetulani, 2000). Chodziło także o bezpośrednią i szybką konfrontację z autentycznymi faktami językowymi, tj. sprawdzenie występowania tych jednostek we współczesnej polszczyźnie, co zapewniał obrany korpus. Były powody, by sądzić, że technologia pozwoli na odkrycie w łatwy sposób (dzięki wspomaganiu komputerowemu) nowych znaczeń dla form, które uprzednio zostały przeanalizowane metodą „ręczną" na podstawie lektury słownika tradycyjnego i dla których został zaproponowany konkretny format opisu (do wykorzystania informatycznego). Były to jednostki Klasy I (*ibidem*), w której znalazły się nazwy różnych czynności, zachowań, operacji, technik itd. Podkreślmy jednak, że zmiana metody (w tym korpusu) nie miała na celu weryfikacji przyjętych na wstępie założeń metodologicznych (*lexique-grammaire*, cf. M. Gross, 1975), zgodnie z którymi powstał pierwszy opis tych jednostek, ani zmiany zaproponowanego formatu opisu semantycznego (w formie kodu odzwierciedlającego użycie gramatyczne jednostki i przeznaczonego do zastosowań informatycznych).

2.1.1. *Główne problemy wynikające z automatycznego przetwarzania korpusu*

Jak powszechnie wiadomo, podstawowe trudności związane z analizą automatyczną zinformatyzowanego korpusu wynikają ze złożoności języka naturalnego i niedoskonałości (niedopasowania) do jego specyfiki programów informatycznych. Podczas przeszukiwania tekstu, w celu wygenerowania kontekstów interesujących lingwistę, programy muszą radzić sobie z problemami związanymi z wielkością materiału, wieloznacznością językową, segmentacją zdań, a także jakością samego korpusu (np. zapisem ortograficznym itd.).

Dokładny opis napotkanych trudności, które wystąpiły w trakcie badań, można znaleźć w G. Vetulani 2012: 69-81. W tym miejscu ograniczymy się do wymienienia najczęstszych spośród nich. A zatem dochodziło do:

– nierozpoznania homonimii wyrazowej (np. system wygenerował zestawienie wyrazowe *mieć poza sobą*, „uznając", że chodzi o rzeczownik predykatywny *poza*, gdy w rzeczywistości był to przyimek (dodajmy na marginesie, że fakt ten można łatwo zrozumieć, ponieważ w języku polskim możliwa jest struktura *mieć jakąś pozę* w sensie 'przybrać jakąś pozę');
– błędnego wytypowania formy czasownikowej na czasownik podporowy dla danego predykatu, gdy tymczasem był on użyty w swoim pełnym znaczeniu (por. *nie ma innej metody* jako konstrukcję bezosobową i konstrukcję *ktoś nie ma jakiejś metody na coś...*);
– wygenerowania z korpusu rzeczownika z cechą [+konkr] lub [+osob] jako predykatu, choć wiadomo, że funkcję tę mogą pełnić jedynie formy w użyciu abstrakcyjnym (np. system rozpoznał predykat *gierka* w znaczeniu 'gra' z kontekstu, w którym chodziło o *Edwarda Gierka*);
– nierozpoznania dwóch elementarnych zdań, tj. struktur predykatywno-argumentowych (na poziomie struktury głębokiej) w pojedynczym (powierzchniowo) zdaniu prostym (np. zdanie: *... ale też by dawało gwarancję bezpieczeństwa przede wszystkim naszym pacjentom...* zawiera predykat *gwarancja*, który wraz z czasownikiem *dawać* (*dawać gwarancję*) oznacza 'gwarantować' oraz predykat *bezpieczeństwo*, który występuje tutaj pod postacią grupy imiennej jako zredukowane zdanie proste: 'pacjenci są bezpieczni');
– generowania bardzo długich list (liczonych w tysiącach linii) określonych zestawień wyrazowych wynikających z niewyważenia korpusu (z jego niezrównoważenia, gdy idzie o reprezentatywność w stosunku do całego

systemu języka polskiego); np. predykat *dyskusja* pojawił się 5788 razy w zestawieniu z czasownikiem *otwierać*, ponieważ korpus zawierał bardzo dużą liczbę stenogramów z sesji posiedzeń Sejmu i Senatu, podczas których marszałek prowadzący obrady otwierał lub zamykał dyskusję, wypowiadając formuły: *otwieram dyskusję*, *zamykam dyskusję*);
– generowania wielu kontekstów trudnych do zaakceptowania – mimo ich autentyzmu (zaświadczonego przez korpus) – z powodu niepoprawności użycia, niegramatyczności lub zbytniej oryginalności, choć niektóre z nich mogły uchodzić za innowacje językowe, a na pewno były zrozumiałe w konkretnej sytuacji komunikacyjnej (np. *?dokonujemy niezwykłego fikołka*); tego typu problemy nie pojawiają się w badaniach opartych na dziełach normatywnych i uznanych tekstach literackich.

Wyżej wymienione i wiele innych, podobnych, przypadków wymagały wnikliwej i niekiedy żmudnej lektury na etapie sprawdzającym, kiedy to leksykografowie metodą „ręczną" zatwierdzali lub odrzucali przykłady użyć.

2.1.2. *Korzyści wynikające z zastosowanej metody*

Mimo opisanych wyżej trudności obrana metoda okazała się owocna. Na sukces składa się wiele przyczyn:
– skuteczny okazał się kod opracowany w pierwszej fazie badań nad predykacją rzeczownikową języka polskiego (G. Vetulani, 2000); kod ten wykorzystano przy budowie programu informatycznego zastosowanego w drugiej fazie, tj. podczas analizy wspomaganej komputerowo;
– skrócony został czas potrzebny do przebadania obranego korpusu (bardzo dużego materiału językowego);
– otrzymano duży, liczący ponad 14600 jednostek, przydatny i do badań podstawowych, i do aplikacji w informatyce, tłumaczeniu lub dydaktyce, zbiór charakterystycznych związków wyrazowych języka polskiego (konstrukcji analitycznych, tzw. kolokacji werbo-nominalnych o strukturze: *czasownik + rzeczowik predykatywny*, uwidaczniający wewnętrzne bogactwo i zróżnicowanie tych zwrotów);
– udało się podnieść jakość i wielkość słownika uzyskanego po pierwszym etapie badań; *de facto* powstał nowy słownik, w nieco zmienionym formacie, pt. *Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I*, który został przygotowany w wersji elektronicznej i dołączony do monografii (G. Vetulani, 2012).

2.2. PRACE W TOKU A KORPUS

Obecnie, w ramach większego projektu (zob. informacje podane w przypisie 2), toczą się dalsze prace nad rozbudową słownika. Rozszerzenie ma głównie polegać na opracowaniu według tej samej metodologii nowych kategorii, tj. predykatów będących nazwami cech, które w monografii z 2000 r. (G. Vetulani) zostały zakwalifikowane do Klasy II[5].

W fazie wstępnej projektu dokonano analizy istniejącego formatu (opracowanego dla rzeczowników predykatywnych Klasy I) pod kątem jego wykorzystania. Jak się szybko okazało, trzeba było poddać ten format kilku modyfikacjom, ponieważ nazwy cech – inaczej niż jednostki Klasy I – występują w języku polskim także w mianowniku (cf. *ciekawość go bierze*, *rzecznika cechuje obiektywizm*, *wesołość ogarnęła zgromadzenie*, *zazdrość nim owładnęła*, *profesjonalizm charakteryzuje ludzi powołanych do...* itp.). Nie oznacza to jednak, że trzeba było całkowicie zrezygnować z przyjętego modelu, tj. *N0 Vsup (MOD) Npred N1 N2...*[6], ponieważ oddaje on również funkcjonowanie gramatyczne nazw cech (por. wyżej wymienione jednostki: *ktoś ma naturalną ciekawość do...*, *ktoś wykazał życzliwą ciekawość*, *ktoś wykazał obiektywizm*, *ktoś wykazał się obiektywizmem*, *ktoś wpadł w wesołość*, *ktoś okazuje zazdrość wobec kogoś*, *ktoś poczuł zazdrość*, *ktoś cechuje się profesjonalizmem*, *ktoś nabrał profesjonalizmu*). W związku z powyższym istniejący format opisu już teraz (w trakcie aktualnych prac – cf. G. Vetulani, 2013: 295) jest poszerzany w celu oddania specyfiki analizowanych jednostek i zapewne będzie jeszcze dopracowywany.

Analiza kontekstów zawierających nazwy cech odbywa się na podobnych zasadach co poprzednio, tj. przy użyciu zinformatyzowanego korpusu oraz programów wyszukujących zadane konkordancje, wygenerowane automatycznie, ze słowem kluczowym będącym nazwą cechy. Tak przygotowane dane poddaje się analizie przez leksykografów[7] w celu wyznaczenia dla każdej nazwy modelu strukturalnego, będącego jednocześnie jej opisem semantycznym.

---

[5] W monografii z 2000 r. charakter jednostek należących do Klasy II został jedynie naświetlony. Obecnie dąży się do uzyskania opisu porównywalnego z tym, jakiego dokonano dla Klasy I.

[6] *N0* odnosi się do argumentu-podmiotu, *Vsup* to czasownik podporowy dla danego predykatu (tutaj: nazwy cechy), *(MOD)* odsyła do obowiązkowego, dodatkowego elementu (przeważnie przymiotnika) występującego w strukturze, *Npred* jest symbolem rzeczownika predykatywnego (nazwy cechy), a *N1, N2* to kolejne argumenty.

[7] W przetwarzaniu automatycznym tekstu oraz sprawdzaniu kontekstów metodą tradycyjną udział biorą na bieżąco: A. Kaliska, B. Kochanowski, M. Nkollo, T. Obrębski, G. Vetulani i Z. Vetulani

Z chwilą przystąpienia do prac za materiał badawczy posłużyła nam ta sama co poprzednio wersja korpusu IPI PAN (z 2004 r., a nie wersja uaktualniona, głównie z powodu ograniczeń wynikających z praw autorskich). I choć w dość szybkim tempie uzyskuje się obecnie konteksty użyć, na których podstawie przeprowadza się analizy, to w bardzo wielu przypadkach wyszukiwanie okazuje się niezadawalające. Dotychczas nie udało się uzyskać zaświadczeń (zgodnie z przyjętymi zasadami, tj. ze zdefiniowanymi modelowo strukturami, w których – jak się wydaje – nazwy cech powinny wystąpić) dla wielu jednostek, m.in. takich, jak: *ciemność*, *ciężkość*, *ewentualność*, *niekaralność*, *niesprawność*, *nieuchronność*, *niewinność*, *proporcjonalność*, *przekupstwo*, *rozpiętość*, *spójność*, *szczelność*, *terminowość*. Nie twierdzimy tutaj, że wyżej wymienione nazwy nie pojawiły się w korpusie w ogóle jako nazwy cech, tylko że nie były tam zaświadczone w znaczeniu ‘kogoś (coś) cechuje coś’ (np. *tę wypowiedź cechuje spójność* albo *ta wypowiedź odznacza się spójnością*, *ktoś odznacza się dużym stopniem niesprawności*, *wywód charakteryzuje się spójnością* albo *wywód wykazuje spójność* itp.). Chodziło bowiem o takie przykłady z języka polskiego, z których by wynikało, że predykaty rzeczownikowe (będące nazwami cech) są odpowiednikami realizacji przymiotnikowych (np. *ktoś odznacza się dużym stopniem niesprawności = ktoś jest niesprawny*, *wywód charakteryzuje się spójnością = wywód jest spójny*, *bryła wykazuje proporcjonalność = bryła jest proporcjonalna* itd., itp.).

Z uwagi na liczne przypadki tego typu, ale też przypuszczenie, że zaobserwowane braki wynikają z nieodpowiedniości obranego korpusu, niereprezentatywnego dla tego typu użyć, lub z niedoskonałości programów komputerowych, za pomocą których odbywało się jego przetwarzanie, podjęto czynności sprawdzające, aby ustalić, czy przyczyny leżą po stronie danych tekstowych czy narzędzi. Przeprowadzono doświadczenie, polegające na wygenerowaniu konkordancji służących za podstawę prac leksykografów przy wykorzystaniu różnych pakietów narzędzi. Okazało się, że na tym samym materiale tekstowym uzyskano zgodne wyniki, co uwiarygodniło narzędzia. Jednocześnie zmiana źródła danych, polegająca na przejściu do pozyskiwania konkordancji z internetu, istotnie zwiększyła liczbę zaświadczeń interesujących nas zjawisk. Doświadczenie to zwraca uwagę na fakt, że w lingwistyce, podobnie jak w tradycyjnych naukach empirycznych, o jakości pozyskiwanej wiedzy decyduje jakość przeprowadzonych obserwacji.

3. PODSUMOWANIE

Wydaje się, że obecnie nie ma odwrotu od badań korpusowych. Nawet fragmentaryczne badania prowadzone na autentycznym materiale językowym przygotowanym do analizy – jak te prowadzone w zakresie predykacji rzeczownikowej języka polskiego – mogą przyczynić się do ustanowienia standardów dla korpusów ogólnego przeznaczenia.

BIBLIOGRAFIA

Bańko Mirosław, 2001, *Z pogranicza leksykografii i językoznawstwa. Studia o słowniku jednojęzycznym*, Wydział Polonistyki Uniwersytetu Warszawkiego, Warszawa.

Bauer M. & Aarts B., 2000, « Corpus construction: a principle for qualitative data collection » [in:] *Qualitative researching with text, image and sound: a practical handbook*, [éds.] Bauer M., Gaskell G., London, Sage, 19-37

Bralewski Dariusz, 2012, *Od przekładu do słownika. Korpus równoległy w redakcji słowników tłumaczeniowych*, Oficyna Wydawnicza LEKSEM, Łask.

Gross Maurice, 1975, *Méthodes en syntaxe*, Paris.

Habert Benoît & Nazarenko Adeline & Salem André, 1997, *Les linguistiques de corpus*, Armand Colin, Paris.

Piotrowski Tadeusz, 1994, *Z zagadnień leksykografii*, PWN, Warszawa.

Przepiórkowski Adam, 2004, *Korpus IPI PAN. Wersja wstępna*, Instytut Podstaw Informatyki, Warszawa.

Vetulani Grażyna, 2013, „Budowa syntaktycznego słownika rzeczowników predykatywnych języka polskiego na potrzeby zastosowań informatycznych w dobie aktualnych wyzwań dla językoznawstwa" [in:] *Scripta manent – res novae*, [éds.] Puppel S., Tomaszkiewicz T., Wydawnictwo Naukowe UAM, Poznań, 485-498.

Vetulani Grażyna, 2012, *Kolokacje werbo-nominalnejako samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I*., Wydawnictwo Naukowe UAM, Poznań.

Vetulani Grażyna, 2010, « Élaboration d'un dictionnaire des noms prédicatifs en polonais » [in:] *Supports et prédicats non verbaux dans les langues du monde*, [éd.] Ibrahim A.H., Paris: Cellule de Recherche en Linguistique, 166–181.

Vetulani Grażyna, 2000, *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych*, Wydawnictwo Naukowe UAM, Poznań.

Vetulani Grażyna & Obrębski Tomasz & Vetulani Zygmunt, 2007, "Towards a Lexicon-Grammar of Polish: Extraxion of Verbo-Nominal Collocations from Corpora" [in:] *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*, [éds.] Wilson D.C., Sutcliffe G.C.J., Menlo Park. California, 267–268.

Vetulani Grażyna & Vetulani Zygmunt & Obrębski Tomasz, 2006, "Syntactic Lexicon of Polish Predicative Nouns" [in:] *Fifth International Conference on Language Resources and Evaluation. 24–26.05.2006*, [éd.] Calzolari N., Genoa–Paris, 1734–1737.

Vetulani Zygmunt & Obrębski Tomasz & Vetulani Grazyna & Dąbrowski Adam & Kubis Marek & Osiński Jędrzej & Walkowska Justyna & Kubacki Piotr & Witalewski Krzysztof, 2010,

Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego, Wydawnictwo Naukowe UAM, Poznań.

Vitas Duško & Krstev Cvetana, 2008, "O paralelnim korpusima, a posebno o beogradskim paralelnim korpusima i načinu njihove eksploatacije" [in:] *Die Unterschide zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, [éd.] B. Tošović, LITVerlag, Münster.

DÉSAVANTAGES ET PROFITS DU TRAITEMENT AUTOMATIQUE DES CORPUS À L'EXEMPLE DES RECHERCHES SUR LA PRÉDICATION NOMINALE EN POLONAIS

R é s u m é

Cet article rend compte des travaux menés depuis un certain temps dans le domaine de la prédication nominale en polonais dans lesquels on exploite des corpus électroniques en utilisant des outils d'analyse automatique du texte. On y présente certaines difficultés qui ont apparu en liaison avec la méthode appliquée, mais on souligne aussi qu'aujourd'hui il est pratiquement impossible de mener des recherches linguistiques autrement et que, finalement, cette méthode s'est avérée très efficace. Dans un laps de temps assez court, elle a permis de construire un dictionnaire des noms prédicatifs du polonais destiné aux applications informatiques et elle contribue à l'heure actuelle au développement du dictionnaire existant.

Mots-clés: linguistique de corpus, traitement automatique du texte, prédication nominale.