MONIKA CZEREPOWICKA* | UNIWERSYTET WARMIŃSKO-MAZURSKI W OLSZTYNIE

Michał Woźniak, *Jak znaleźć igłę w stogu siana? Automatyczna ekstrakcja wielosegmentowych jednostek leksykalnych z tekstu polskiego*

PRACE INSTYTUTU JĘZYKA POLSKIEGO PAN, KRAKÓW 2017, S. 157

doi: http://dx.doi.org/10.31286/JP.99.3.14

Zagadnienia związane z wyodrębnianiem i rejestracją związków frazeologicznych stanowią ważki problem teoretyczny, ale i praktyczny, z którym mierzą się leksykografowie i anotatorzy korpusów tekstów (Bańko 2001; Rosén i in. 2016; Brook i in. 2018; Nivre 2018; Żmigrodzki i in. (red.) 2018). Monografia Michała Woźniaka stanowi interesującą pozycję zarówno z punktu widzenia badań lingwistycznych nad polską frazeologią, jak i w perspektywie szerszej – z punktu widzenia lingwistyki komputerowej. O ile kwestie automatycznej analizy morfoskładniowej leksemów można uznać za sprawę w polszczyźnie rozwiązaną (por. znakowanie NKJP oraz zawartość SGJP; Przepiórkowski i in. (red.) 2012; Kieraś, Woliński 2017), o tyle problematyka frazeologiczna stanowi nadal wyzwanie. Można wnioskować o tym pośrednio – w NKJP oraz w Składnicy, banku drzew składniowych (Woliński i in. 2018) jednostki wielowyrazowe nie są oznaczane. Poszukiwanie skutecznych sposobów na automatyczne lub półautomatyczne wydobywanie ich z tekstów oraz stworzenie ich wyczerpującego opisu wydaje się aktualnie pilącym problemem. Książka Michała Woźniaka stanowi odpowiedź na to zapotrzebowanie. Autor podejmuje się zadania kluczowego z punktu widzenia maszynowego przetwarzania polszczyzny, jakim jest stworzenie i opis automatycznej metody wydobycia (ekstrakcji) jednostek frazeologicznych z tekstów. W kontekście tezy Andrzeja Bogusławskiego (1976), wedle którego jednostki języka nie są dane, lecz należy je wydobyć, zadanie to jest kluczowe również dla językoznawstwa ogólnego i polonistycznego.

Monografia składa się z pięciu rozdziałów, bibliografii oraz trzech dodatków. Autor przedstawia w niej obszary zainteresowania i problemy lingwistyki komputerowej. Omawia znaczenie identyfikacji i opisu jednostek frazeologicznych w ramach zadań lingwistyki komputerowej. Podkreśla, że ustalenia na temat nieciągłych jednostek języka są istotne tak w pracach nad tłumaczeniem maszynowym, analizą składniową i semantyczną, jak i we wspomaganej komputerowo leksykografii i glottodydaktyce.

Badacz zwraca uwagę na panującą w językoznawstwie wielość terminologiczną dotyczącą frazeologii. Zauważa, że w miejsce „klasycznych" terminów, takich jak *frazeologizm*, *związek*

* monika.czerepowicka@uwm.edu.pl; ORCID: 0000-0002-4697-7058

frazeologiczny, coraz częściej pojawiają się bardziej techniczne, takie jak *wielosegmentowa jednostka języka*, *nieciągła jednostka języka*. Na tego typu terminy decyduje się również autor monografii.

Michał Woźniak przywołuje najważniejsze stanowiska metodologiczne dotyczące rozumienia i wyodrębniania jednostek wielowyrazowych (Wiktor Winogradow, Stanisław Skorupka, Andrzej Bogusławski, Andrzej Maria Lewicki, Wojciech Chlebda, Igor Mielczuk, Ivan Sag i współpracownicy, Rosamund Moon). Zasadniczą wartość tego rozdziału stanowi wszakże autorska definicja jednostki wielowyrazowej, u której podstaw leży przekonanie, że frazeologia jest zjawiskiem stopniowalnym. Autor dochodzi do niego po omówieniu stanu badań. Żadna z prezentowanych teorii nie daje bowiem jego zdaniem narzędzi, które pozwoliłyby na opracowanie danych empirycznych w sposób automatyczny. Twórcy klasyfikacji stosują binarny podział na jednostki i niejednostki frazeologiczne, mimo to często wskazują połączenia wątpliwe, stanowiące pogranicze składni i frazeologii. Woźniak interpretuje frazeologiczność jako swego rodzaju *continuum*, dające się umieścić na skali określonej liczbowo: jeden jej kraniec stanowią swobodne połączenia składniowe, drugi – wyrażenia ściśle frazeologiczne. Obszar pomiędzy skrajnościami wypełniają jednostki m n i e j lub b a r d z i e j frazeologiczne. Budując własną definicję jednostki wielowyrazowej, autor uwzględnia większość z kryteriów znanych z literatury przedmiotu, czyli nieregularność semantyczną, leksykalną, składniową, swoistość statystyczną, pragmatyczną, odtwarzalność, stałość leksykalną, morfosyntaktyczną, konwencjonalizację, nieprzekładalność, zamkniętość–otwartość klas substytucyjnych oraz ekspresywność i obrazowość. Każdemu z kryteriów przypisuje wagę, stosując w tym celu tzw. współczynnik istotności (*wi*). Jako najważniejsze kryterium traktuje nieregularność znaczeniową (*wi* = 4), następnie otwartość–zamkniętość klas substytucyjnych, pragmatyczność, konwencjonalizację (dla wszystkich *wi* = 3) i w końcu nieprzetłumaczalność, ekspresywność, obrazowość oraz stałość leksykalną (*wi* = 2). Kryteriom przypisuje jedną z trzech wartości: silną (=2), zauważalną (=1) lub zerową. Na przykład wartość parametru stałości leksykalnej jest znaczna (=2), jeśli żaden z członów połączenia nie podlega wymianie, np. jak w połączeniach *wolna ręka*, *martwa natura*. Jeśli możliwe są podstawienia któregoś z członów i liczba substytucji wynosi nie więcej niż 5, to wartość parametru jest zauważalna (=1), np. *siwy włos*, *duża litera*. Stopień frazeologiczności wyrażenia ustala się na podstawie iloczynu współczynnika istotności kryterium i stopnia jego spełnienia. Autor podaje definicję wielowyrazowej jednostki języka: jest nią połączenie co najmniej dwuwyrazowe, które na podstawie sumowania wagi kryteriów uzyskało przynajmniej 8 punktów. Podkreśla jednak, że zaproponowane kryteria i sposób ich oceny stanowią szkielet metodologiczny, za pomocą którego można modelować pojęcie jednostki języka. Przyjęty przez niego próg punktowy można obniżać lub podwyższać, uzyskując w ten sposób mniej lub bardziej „rygorystyczną" definicję jednostki.

Omawiając metody automatycznej ekstrakcji jednostek wielowyrazowych, Woźniak koncentruje się na jednej z nich, polegającej na stosowaniu „okna" o ustalonej szerokości, którą wyznacza liczba wyrazów. „Okno" takie jest przesuwane wzdłuż tekstu, a jego zawartość zapisywana. W ten sposób dochodzi do wyłuskania dowolnej długości *n*-gramów, czyli *n*-długich wyrażeń języka naturalnego. W monografii znajdujemy opis automatycznej ekstrakcji bigramów,

czyli połączeń dwuwyrazowych. Do ich wydobycia autor stosuje autorski algorytm, ufundowany na przekonaniu o praktycznej ewaluacji reguły, koniecznej z punktu widzenia maszynowego przetwarzania. Zastosowany algorytm ma charakter hybrydowy – łączy zarówno metody lingwistyczne, jak i statystyczne. Jego działanie przebiega w kilku krokach. Pierwszy to przygotowanie tekstu – korpusu (od segmentacji do znakowania morfoskładniowego) oraz sporządzenie listy modeli składniowych. Autor decyduje się na trzy podstawowe wzorce, typowe dla wyrażeń nominalnych: połączenia związku zgody rzeczownika z przymiotnikiem i przymiotnika z rzeczownikiem oraz połączenie związku rządu dwu rzeczowników, z których drugi występuje w dopełniaczu. Kolejny krok stanowi wyszukiwanie kandydatów jednostek leksykalnych. Na tym etapie zbierane są także dane frekwencyjne dotyczące i połączeń wyrazowych, i ich poszczególnych segmentów. Ustalana jest frekwencyjna łączliwość leksykalna wyrażeń, czyli liczba podstawień prawo- i lewostronnych wyrazów wchodzących w skład potencjalnych jednostek leksykalnych. Wygenerowana w ten sposób lista jest następnie filtrowana. Jednym z pierwszych kryteriów jest frekwencja (odrzucane są wyrażenia, które wystąpiły w korpusie treningowym mniej niż 5 razy). Dalszy etap to weryfikacja statystyczna. Dla każdego kandydata podaje się wartości miar asocjacyjnych, takich jak między innymi informacja wzajemna (PMI), statystyka *z* (*z-score*), współczynnik Dice'a, podkreślając, że dają one zróżnicowane wyniki. Czwarty etap prac nad ekstrakcją jednostek zakłada maszynowe uczenie algorytmu. Autor wykorzystuje metodę maszyn wektorów nośnych (SVM). Skuteczność klasyfikatora jest sprawdzana na zbiorze treningowym, składającym się z ponad dwóch tysięcy ręcznie oznakowanych wyrażeń, z których blisko 40 procent stanowią jednostki wielosegmentowe, reszta to standardowe konstrukcje składniowe. Skuteczność algorytmu ocenia się przez zestawienie jego wyników na zbiorze porównawczym, stanowiącym 5-milionowy podzbiór *sample* korpusu Instytutu Podstaw Informatyki Polskiej Akademii Nauk. Wyodrębnione z korpusu bigramy, realizujące jeden z zadanych wzorców składniowych i o frekwencji wynoszącej więcej niż 10, zostały ręcznie opracowane zgodnie z zaproponowaną definicją jednostki wielowyrazowej. W ten sposób powstała lista ponad 800 jednostek wielosegmentowych, jej część jest dostępna w książce w postaci *Dodatku A*. W monografii znajdujemy również listę 100 najwyżej ocenionych jednostek w nieanotowanym morfosyntaktycznie korpusie notatek prasowych Polskiej Agencji Prasowej (*Dodatek B*) oraz listę najwyżej i najniżej ocenionych wyrażeń według maszyny wektorów nośnych (również *Dodatek B*).

Finalny etap prac stanowi ocena skuteczności algorytmu. Autor, by uniknąć zawyżonych wyników, stosuje metodę ewaluacji krzyżowej. W tym celu zbiór treningowy dzieli na kilka zbiorów testowych równej wielkości. Podczas gdy jeden z nich służy za zbiór testowy, pozostałe, złączone w całość, stanowią odniesienie, czyli zbiór porównawczy. Procedura jest powtarzana tyle razy, ile wyróżniono podzbiorów treningowych. Zaletą takiego sposobu działania jest identyczna struktura obu typów zbiorów, a co za tym idzie – ich rzeczywista porównywalność.

Analizując wyniki działania algorytmu, autor realnie ocenia jego skuteczność jako przeciętną. Michał Woźniak jest świadom, że metodę ekstrakcji warto byłoby wykorzystać do wyłuskania innych jednostek, na przykład dłuższych, lub bigramów werbalnych. Sugeruje, by podczas ewaluacji zastosować inne miary asocjacji lub zmodyfikowane wersje już użytych.

Zauważa jednak, że tradycyjne metody nie mogą przekroczyć stosunkowo niskiego progu skuteczności. Postęp w automatycznej ekstrakcji wiąże z maszynowym uczeniem, najlepiej z jednoczesnym wdrożeniem metod statystycznych i lingwistycznych. Dlatego proponuje, by zastosowaną w algorytmie maszynę wektorów nośnych zastąpić sieciami neuronowymi lub klasyfikatorami Bayesowskimi, co wyznacza obiecujący kierunek rozwoju.

Monografia Woźniaka podejmuje temat mechanizmów i metod automatycznej operacji na języku naturalnym. Robi to w sposób zrozumiały i klarowny. Niekiedy w tekście zdarzają się niekonsekwencje i nieścisłości (jak np. liczba jednostek w *Dodatku A* deklarowana przez autora w rozdziale IV na s. 107 i rzeczywista liczba jednostek w dodatku). Autor ma dobre rozeznanie w koncepcjach teoretycznych, referuje je sprawnie, z lekkością. Szkoda, że w tekście zabrakło bezpośrednich odesłań bibliograficznych przy niektórych nazwiskach, na przykład Winogradowa, Skorupki, Mielczuka. W wykazie bibliografii brakuje kilku pozycji, na które autor powołuje się w monografii (np. Cowie 1998 – błędnie zapisana w tekście jako Cowie 1978, por. s. 45). Trzeba przyznać, że zawiłości natury informatycznej są przedstawiane w sposób przystępny, choć podejmowanie analogicznych treści i zagadnień w różnych miejscach książki może powodować wrażenie dezorientacji.

Bez względu na wskazane usterki monografia niewątpliwie zasługuje na uwagę. Wydaje się szczególnie istotna w świetle badań językoznawstwa komputerowego, które posługuje się obszernymi danymi językowymi. Propozycja autorskiej definicji jednostki języka jest również inspirująca z punktu widzenia teorii jednostki leksykalnej. W swojej klasyfikacji Woźniak ujmuje bodaj wszystkie kryteria wymieniane przez badaczy jako cechy definicyjne wielosegmentowych jednostek leksykalnych. Jednocześnie, wprowadzając współczynnik istotności, nadaje im pewną hierarchię, dzięki czemu potwierdza swoje osadzenie w tradycji językoznawczej. Trzeba podkreślić twórcze czerpanie autora z tradycji, jak i najnowszych badań lingwistycznych. Perspektywa skalarnego charakteru frazeologiczności pobrzmiewała już co prawda we wcześniejszych pracach (por. Chlebda 1991; Pajdzińska 1991), ale dopiero teraz doczekała się rozwinięcia i umocowania formalnego, tym bardziej istotnych, że uzasadnionych z perspektywy automatycznego przetwarzania języka.

Zarówno zawartość dodatków A i B, jak i przygotowany na potrzeby algorytmu zbiór porównawczy stanowią cenny materiał dalszych badań lingwistycznych i leksykograficznych oraz eksperymentów badawczych.

Bibliografia

Bańko M. 2001: *Z pogranicza leksykografii i językoznawstwa. Studia o słowniku jednojęzycznym*, Wydział Polonistyki Uniwersytetu Warszawskiego, Warszawa.

Bogusławski A. 1976: *Zasady rejestracji jednostek języka*, „Poradnik Językowy", z. 8(342), s. 356–364.

Brook J., Chan K., Baldwin T. 2018: *Semi-automated resolution of inconsistency for a harmonized multiword-expression and dependency-parse annotation*, [w:] S. Markantonatou, C. Ramisch, A. Savary, V. Vincze (red.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, Language Science Press, Berlin, s. 245–262.

Chlebda W. 1991: *Elementy frazematyki. Wprowadzenie do frazeologii nadawcy*, Wyższa Szkoła Pedagogiczna w Opolu, Opole.

Cowie A.P. (red.) 1998: *Phraseology: Theory, Analysis, and Applications*, Oxford University Press, Oxford.

Kieraś W., Woliński M. 2017: *Słownik gramatyczny języka polskiego – wersja internetowa*, „Język Polski" XCVII, z. 1, s. 84–93.

Nivre J. 2018: *Multiword Expressions in Dependency Parsing, invited talk at the 12th MWE Workshop, Berlin, August, 2016* (online: https://cl.lingfil.uu.se/~nivre/docs/Nivre-MWE16.pdf, dostęp: 24 maja 2019).

NKJP: Narodowy Korpus Języka Polskiego (online: http://nkjp.pl/, dostęp: 24 maja 2019).

Pajdzińska A. 1991: *Wartościowanie we frazeologii*, [w:] J. Puzynina, J. Anusiewicz (red.), *Język a kultura*, t. 3, Wrocław, s. 15–28.

Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (red.) 2012: *Narodowy Korpus Języka Polskiego: praca zbiorowa*, Wydawnictwo Naukowe PWN, Warszawa.

Rosén V., Thunes M., Haugereid P., Losnegaard G.S., Dyvik H., Meurer P., Lyse G.I., De Smedt K. 2016: *The enrichment of lexical resources through incremental parsebanking*, „Language Resources and Evaluation", Vol. 50(2), s. 291–319, doi: 10.1007/s10579-016-9356-5.

SGJP: Z. Saloni, M. Woliński, R. Wołosz, W. Gruszczyński, D. Skowrońska, *Słownik gramatyczny języka polskiego*, wyd. 3 online, Warszawa 2015 (online: http://sgjp.pl, dostęp: 24 maja 2019).

Woliński M., Hajnicz E., Bartosiak T. 2018: *A new version of Składnica treebank of Polish harmonised with the Walenty valency dictionary*, [w:] N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (red.), *Proceedings of the eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Paris, s. 1839–1844 (online: https://www.aclweb.org/anthology/L18-1, dostęp: 24 maja 2019).

Żmigrodzki P., Bańko M., Batko-Tokarz B., Bobrowski J., Czelakowska A., Grochowski M., Przybylska R., Waniakowa J., Węgrzynek K. (red.) 2018: Wielki słownik języka polskiego PAN. *Geneza, koncepcja, zasady opracowania*, Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków.