TORUŃSKIE STUDIA BIBLIOLOGICZNE
2011, nr 1 (6)

Piotr Malak
Instytut Informacji Naukowej i Bibliologii
Uniwersytet Mikołaja Kopernika w Toruniu
e-mail: piomk@umk.pl

Metody statystyczne w komputerowym przetwarzaniu języka naturalnego

Wśród metod komputerowego przetwarzania dokumentów języka naturalnego niepoślednie miejsce zajmują metody statystyczne. Analiza statystyczna tekstów, frekwencji poszczególnych wyrazów czy zależności współwystępowania konstrukcji wielowyrazowych jest jednym z najczęściej wykorzystywanych narzędzi w wyszukiwaniu informacji (ang. *information retrieval*).

Systemy wyszukiwawcze w znacznym stopniu wykorzystują statystyczne metody komputerowego przetwarzania tekstów, zarówno z pojedynczych dokumentów, jak i – przede wszystkim – z całych ich kolekcji. Gwoli uściślenia należy dodać, że najczęściej analizowane są teksty z poszczególnych dokumentów, natomiast wnioski wyciągane są na podstawie porównania określonych właściwości dotyczących jednego dokumentu z wartościami tych samych cech stwierdzonych dla całego zbioru dokumentów. Wnioskowanie o prawidłowościach językowych jest przeprowadzane na podstawie statystycznej analizy odpowiednio dużych zbiorów tekstów.

Statystyka w przetwarzaniu języka naturalnego

Jak podaje Mieczysław Sobczyk, statystyka jest nauką dotyczącą ilościowych metod badania zjawisk (inaczej procesów) masowych[1]. Pojęcie

[1] Por. M. Sobczyk, *Statystyka*, wyd. 3 zm., Warszawa 2000; tenże, *Statystyka. Podstawy teoretyczne, przykłady – zadania*, Lublin 1998.

masowości zakłada badanie odpowiednio dużego zbioru jednostek, które cechują się podobnymi, ale nieidentycznymi właściwościami. Wynikiem badań statystycznych są reguły bądź wnioski dotyczące uśrednionych wartości cech badanych zbiorowości. Owe reguły to tzw. prawidłowości statystyczne. Badania statystyczne dotyczą tzw. zbiorowości statystycznej (populacji, masy statystycznej). Populacja oznacza zbiór elementów objętych badaniem statystycznym. Poszczególne elementy składowe populacji są nazywane jednostkami statystycznymi, przy czym w obrębie jednej zbiorowości statystycznej można wyróżnić wiele jednostek statystycznych (np. podzbiór leksemów, zdań czy też całych tekstów badanego zbioru dokumentów)[2].

Chris Manning i Hinrich Schütze – autorzy pracy *Foundations of statistical natural language processing* – w interesujący sposób streścili umiejscowienie i przynależność statystycznego nurtu przetwarzania języka naturalnego (ang. *Natural Language Processing*, dalej: NLP). Badania kwantytatywne nad językiem naturalnym zdefiniowali jako dyscyplinę łączącą wszystkie podejścia ilościowe do automatycznego przetwarzania języka, włączając w to modelowanie probabilistyczne, teorię informacji oraz algebrę liniową. Pomimo potencjalnej wieloznaczności tego pojęcia Manning i Schütze konkludują, że na przestrzeni ostatniej dekady *statystyczne NLP* było terminem używanym najpowszechniej do oznaczenia wszystkich prac nad przetwarzaniem języka naturalnego niewprowadzających symboliki ani logiki[3].

Należy zgodzić się z powyższymi wywodami, ponieważ badania statystyczne języka naturalnego rzeczywiście korzystają z osiągnięć teorii informacji, teorii prawdopodobieństwa oraz rozwiązań algebry linio-

[2] Tenże, *Statystyka…*, s. 11–13.

[3] Tłumaczenie własne autora na podstawie: Ch. D. Manning, H. Schütze, *Foundations of statistical natural language processing*, Cambridge 1999, s. XXXI–XXXII: "A final remark is in order on the title we have chosen for this book. Calling the field Statistical Natural Language Processing might seem questionable to someone who takes their definition of a statistical method from a standard introduction to statistics. Statistical NLP as we define it comprises all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra. While probability theory is the foundation for formal statistical reasoning, we take the basic meaning of the term 'statistics' as being broader, encompassing all quantitative approaches to data (a definition which one can quickly confirm in almost any dictionary). Although there is thus some potential for ambiguity, Statistical NLP has been the most widely used term to refer to non-symbolic and non-logical work on NLP over the past decade, and we have decided to keep with this term".

wej do przeprowadzenia wieloaspektowej analizy wyrażeń językowych. W takim też uniwersalnym znaczeniu będą używane w niniejszym artykule terminy *lingwistyka kwantytatywna* czy też *lingwistyka statystyczna*.

Na opracowanie kwantytatywne zbioru dokumentów składają się w dużej części operacje mechaniczne przygotowujące poszczególne dokumenty do właściwego procesu analizy. Są to operacje takie, jak np. wykluczenie z tekstu wyrazów znajdujących się na liście słów mało znaczących (ang. *stop list*) w celu obniżenia kosztów przetwarzania elementów tekstu, które nie wnoszą wartościowych informacji, zliczenie częstości wystąpień danego wyrazu (ang. *term frequency*) czy porównanie częstości występowania poszczególnych wyrazów w różnych dokumentach badanego zbioru.

Operacje tego typu, ważne dla dokonania poprawnej analizy dokumentu, nie wymagają udziału człowieka, mogą z powodzeniem zostać przeprowadzone przez specjalistyczne oprogramowanie. Zastosowanie komputerów do badań nad tekstami języka naturalnego pozwala na obniżenie kosztów operacji mechanicznych oraz zwielokrotnienie liczby tych operacji wykonanych w określonym czasie w porównaniu do analizy przeprowadzanej przez człowieka. W związku z tym oczywisty jest fakt scedowania na komputery jak największej części prac związanych z opracowaniem zbioru dokumentów i pozostawienia człowiekowi kontroli nad zautomatyzowanym procesem.

W niniejszym artykule zostaną zaprezentowane podstawy kwantytatywnej analizy tekstów języka naturalnego oraz wybrane metody komputerowego przetwarzania języka naturalnego. Zostanie również przeprowadzona dyskusja przyjętych w badaniach NLP terminów.

Analiza kwantytatywna tekstów

Analiza kwantytatywna języka naturalnego wykorzystuje bardzo duże zbiory danych do generowania wniosków o tekstach bądź języku. Metody statystyczne stosowane w badaniach NLP w określonym zakresie pozwalają uzyskać wiarygodne i wartościowe wyniki analiz przy niskich kosztach operacyjnych. Jak podaje Agnieszka Mykowiecka, analiza frekwencyjna znajduje zastosowanie w indeksowaniu lub klasyfikacji dokumentów, wskazywaniu kategorii tematycznej treści dokumentów lub określaniu języka tekstu. Oprócz pojedynczych elementów języka anali-

zie mogą podlegać złączenia, czyli tzw. współwystępowanie składników. Określenie częstości występowania poszczególnych złączeń wyrazów może być wykorzystane np. przy wskazywaniu znaczenia wyrazów wieloznacznych (w zależności od częstości poszczególnych złączeń)[4].

Fundamentalne znaczenie dla przybliżenia statystyki oraz możliwości jej zastosowań w badaniach nad językiem w Polsce mają prace Jadwigi Sambor. W tym miejscu można wymienić m.in. jej publikację *Językoznawstwo statystyczne dla pracowników informacji naukowej* (Warszawa 1978) czy dzieła zbiorowe powstałe we współpracy z Rolfem Hammerlem *Statystyka dla językoznawców* (Warszawa 1990) oraz *O statystycznych prawach językowych* (Warszawa 1993). W swych pracach autorka daje wyczerpujący wstęp do rachunku prawdopodobieństwa i statystycznych metod analizy tekstu oraz wymienia przykłady użycia poszczególnych opisywanych przez siebie metod[5].

Lingwistyka kwantytatywna

Analizą statystyczną prawidłowości ilościowych w tekstach i w języku zajmuje się lingwistyka kwantytatywna. Według definicji słownikowej prawidłowości owe dotyczą m.in. frekwencji (częstości) występowania wyrażeń i struktur językowych wszystkich poziomów języka. Ponadto w zakres lingwistycznych badań kwantytatywnych wchodzi także prawdopodobieństwo występowania wyrażeń i struktur w różnych kontekstach, rodzajach tekstów czy stylach wypowiedzi. Analizowane są również zależności pomiędzy częstością występowania wyrażeń i struktur a innymi cechami tych wyrażeń i struktur lub ich wartością informacyjną. Jak podaje Bożenna Bojar, autorka *Słownika encyklopedycznego informacji, języków i systemów informacyjno-wyszukiwawczych*, wyniki badań języko-

[4] A. Mykowiecka, *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*, Warszawa 2007, s. 188–191.

[5] Bardzo dobry i wyczerpujący wstęp do rachunku prawdopodobieństwa, wnioskowania statystycznego oraz wykorzystania metod statystycznych w badaniach języka zawiera pozycja J. Sambor, *Językoznawstwo statystyczne dla pracowników informacji naukowej*, Warszawa 1978. Praca R. Hammerl, J. Sambor, *Statystyka dla językoznawców*, Warszawa 1990, jest z kolei bardzo szczegółowym wprowadzeniem zarówno do metod statystycznych, teorii informacji, jak i praw oraz prawidłowości językowych uzyskanych na podstawie analiz statystycznych języka. Natomiast pozycja R. Hammerl, J. Sambor, *O statystycznych prawach językowych*, Warszawa 1993, prezentuje dość szczegółowo prawa oraz prawidłowości statystyczne dotyczące języka naturalnego wraz z dyskusją nad terminem *prawo językowe*.

znawstwa statystycznego dowodzą, że częstość występowania wyrażeń jest ich cechą systemową i jako taka powinna być uwzględniana w opisach systemów językowych, formalizacjach transformacji językowych, nauczaniu języków oraz innych pracach związanych z przetwarzaniem języka. *Lingwistyka kwantytatywna* jest traktowana w przywołanym *Słowniku encyklopedycznym informacji...* jako synonim terminu *lingwistyka statystyczna*[6].

Interesującą analizę oraz wprowadzenie do dyscypliny prezentuje również Adam Pawłowski w swojej pracy *Metody kwantytatywne w sekwencyjnej analizie tekstu.* W publikacji tej znajdziemy dyskusję zarówno na temat przedmiotu, jak i celu lingwistyki kwantytatywnej oraz zwięzły, systematyczny opis poszczególnych praw i prawidłowości statystycznych dotyczących tekstów języka naturalnego. W omawianej pracy autor prezentuje także metody sekwencyjnego modelowania struktur tekstu oraz szczegółową dyskusję analizy sekwencyjnej tekstów[7].

Należy również wspomnieć przywołaną już pracę A. Mykowieckiej, której jeden rozdział jest poświęcony statystycznym modelom języka. Autorka zaprezentowała w nim wprowadzenie do metod statystycznych w badaniach języka naturalnego, jak i możliwości praktycznego zastosowania poszczególnych metod i technologii w opracowywaniu modeli języka naturalnego[8].

Badania statystyczne tekstów języka naturalnego mogą dotyczyć elementów różnych poziomów języka. Poniżej przedstawiono jednostki badań oraz definicje wybranych terminów, stosowanych w badaniach kwantytatywnych języka naturalnego.

Jednostki badania kwantytatywnego tekstów

W lingwistyce kwantytatywnej jednostkami badania są podstawowe elementy różnych poziomów języka. Mogą to być np. elementy graficzne (grafemy, symbole i znaki), fonologiczne (fonemy, sylaby), morfologiczne (morfemy gramatyczne, części mowy) czy składniowe (typy zdań,

[6] *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*, oprac. B. Bojar, Warszawa 2002, s. 149.

[7] A. Pawłowski, *Metody kwantytatywne w sekwencyjnej analizie tekstu*, Warszawa 2001, s. 6–74.

[8] A. Mykowiecka, dz. cyt., s. 187–230.

części zdania). Jak podają J. Sambor i R. Hammerl, inwentarze tych jednostek w systemie językowym w aspekcie ściśle synchronicznym można traktować jako skończone i małe[9].

Natomiast w przypadku analizy jednostek leksykalnych można przyjąć, że mamy do czynienia z populacjami nieskończonymi. Adam Pawłowski, językoznawca specjalizujący się w lingwistyce korpusowej, w artykule *Uwagi na temat korpusu języka polskiego (reprezentatywność, aktualność, nazwa)* przy okazji dyskusji wokół metody reprezentacyjnej w badaniach języka analizuje pojęcia *skończoności* oraz *otwartości* poszczególnych podsystemów systemu języka. Autor artykułu wyróżnia podsystemy zamknięte, o niewielkiej i łatwej do określenia liczbie jednostek (np. system fonologiczny), systemy półotwarte, cechujące się przewagą liczbową jednostek potencjalnych nad jednostkami faktycznie obserwowanymi, przy czym dzięki zastosowaniu kombinatoryki można obliczyć liczbę jednostek potencjalnych (np. repertuar morfemów). Ostatni typ podzbiorów, wyróżniony przez tego badacza, stanowią systemy otwarte, czyli takie, w których liczba elementów jest teoretycznie skończona, lecz w praktyce nieprzeliczalna. Przykładem podsystemów otwartych jest system leksykalny danego języka[10].

Autorzy podręcznika *Statystyka dla językoznawców* wyróżniają, za Zygmuntem Salonim[11], następujące jednostki leksykalne:

– słowo,
– słowoforma (forma wyrazowa),
– leksem,
– wyraz,
– hasło[12].

Definicje najważniejszych pojęć

Definicje przyjęte w pracy *Statystyka dla językoznawców* lub *O statystycznych prawach językowych* są podane w postaci skróconej, treścio-

[9] R. Hammerl, J. Sambor, *Statystyka…*, s. 16–17; tychże, *O statystycznych…*, s. 21–22.

[10] A. Pawłowski, *Uwagi na temat korpusu języka polskiego (reprezentatywność, aktualność, nazwa)*, [w:] *Językoznawstwo w Polsce: stan i perspektywy*, pod red. S. Gajdy, Opole 2003, s. 165–166.

[11] Z. Saloni, *Kategoria rodzaju we współczesnym języku polskim*, [w:] *Kategorie gramatyczne grup imiennych w języku polskim*, pod red. R. Laskowskiego, Wrocław–Warszawa 1976, s. 43–78. Cyt. za: R. Hammerl, J. Sambor, *Statystyka…*, s. 17.

[12] R. Hammerl, J. Sambor, *O statystycznych…*, s. 17–19.

wo dostosowanej do potrzeb komputerowego przetwarzania tekstów języka naturalnego. Ciekawą i obszerną dyskusję tych pojęć przeprowadził Janusz S. Bień, który w swoich pracach analizuje szczególnie znaczenie i definicję terminów *wyraz*, *słowo* oraz *leksem*, a także wprowadza własną (dosyć powszechnie obecnie przyjętą) jednostkę – *fleksem*[13].

Należy w tym miejscu nadmienić również, że wielu polskich badaczy analizujących komputerowo język naturalny sięga do opracowań Jana Tokarskiego, który w swych publikacjach rozważał możliwości zautomatyzowania niektórych etapów prac nad słownikami oraz wskazywał pomysły zrealizowania wybranych operacji automatycznie, za pomocą komputerów. Interesujące rozważania nad znaczeniem terminów *wyraz* oraz *forma* można znaleźć w pozycji J. Tokarskiego *Fleksja polska*[14].

Definicje bardziej szczegółowe niż w pracach J. Sambor i R. Hammerla, a jednocześnie bliższe zastosowaniom w informacji naukowej, znajdziemy w przywoływanym już *Słowniku encyklopedycznym informacji*... Omawiane tu terminy można za tym wydawnictwem zdefiniować następująco:

> *wyraz* – traktowany jako synonim terminu *słowo*, jest wyrażeniem elementarnym. W językach naturalnych wyrazy składają się z morfemów leksykalnych lub z morfemów leksykalnych i gramatycznych. Termin *wyraz* może być interpretowany jako *leksem* (wyraz systemowy, czyli wyrażenie poziomu leksykalnego) albo jako *słowoforma*, czyli wyrażenie tekstu (wyraz tekstowy). W celu ułatwienia jednoznacznego wskazania wyrazów w tekstach można dodatkowo zdefiniować je jako ciągi liter pomiędzy znakami delimitacji tekstu (spacje, znaki przestankowe). Ponadto pojedyncze wyrazy można określić jako ciąg morfemów, pomiędzy którymi nie może wystąpić żaden inny morfem[15].

[13] J. S. Bień, *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji* [on-line]. Biblioteka Cyfrowa Katedry Lingwistyki Formalnej Uniwersytetu Warszawskiego [dostęp 15 grudnia 2010]. Dostępny w World Wide Web: http://bc.klf.uw.edu.pl/12/2/emph.pdf; tenże, *O pojęciu wyrazu morfologicznego* [on-line]. Biblioteka Cyfrowa Katedry Lingwistyki Formalnej Uniwersytetu Warszawskiego [dostęp 15 grudnia 2010]. Dostępny w World Wide Web: http://bc.klf.uw.edu.pl/62/1/jsb-zsE.pdf; tenże, *Aparat pojęciowy wybranych systemów przetwarzania tekstów polskich*. Biuletyn Polskiego Towarzystwa Językoznawczego [on-line] 2006, z. 62 [dostęp 15 grudnia 2010]. Dostępny w World Wide Web: http://www.ptj.civ.pl/component/option,com_docman/task,doc_download/gid,20/Itemid,8/. Rozważania te są rozwinięciem ustaleń poczynionych przez J. Tokarskiego.

[14] J. Tokarski, *Fleksja polska*, Warszawa 1978, s. 20–24.

[15] Por. *Słownik encyklopedyczny informacji*..., s. 301.

W *Encyklopedii językoznawstwa ogólnego* termin *wyraz* jest definiowany (w rozumieniu potocznym) jako najmniejsza znacząca jednostka językowa, cechująca się względną samodzielnością składniową[16]. Według kolejnych definicji:

– *słowoforma* – jest wyrażeniem będącym elementem tekstu. Stanowi realizację leksemu poprzez nadanie mu odpowiedniej formy językowej oraz połączenie z odpowiednim morfemem[17];
– *morfem* – jest to najmniejsze wyrażenie przekazujące znaczenie. Można wyróżnić morfemy gramatyczne (fleksyjne oraz słowotwórcze) oraz morfemy leksykalne (rdzenie)[18];
– *termin* – jest wyrażeniem o ściśle ustalonym znaczeniu w danej dziedzinie nauki lub techniki[19].

Termin *leksem* nie został zdefiniowany w *Słowniku encyklopedycznym informacji...* bezpośrednio. Z definicji terminu *wyraz* można wywnioskować, że *leksem* jest to wyrażenie poziomu leksykalnego, czyli wyraz systemowy[20].

Ponadto *Słownik encyklopedyczny informacji...* podaje trzy inne definicje, przydatne do statystycznego przetwarzania języka naturalnego. Są to pojęcia: *słowa klucze*, *słowo kluczowe* oraz *temat*. Ich definicje przybierają następującą postać:

– *słowa klucze* – są to wyrazy cechujące się w danym tekście lub korpusie tekstów frekwencją znacząco większą niż w danym języku naturalnym. Stanowią one wykładniki głównych tematów tekstu, są również charakterystyczne dla danego autora[21];
– *słowo kluczowe* – jest to wyrażenie z tekstu dokumentu lub zapytania informacyjnego charakteryzujące jego treść. W przypadku dokumentów słowa kluczowe pochodzą często z tytułu lub tytułów rozdziałów[22];
– *temat* – definiowany również jako przedmiot dokumentu, to, czego dotyczą zawarte w dokumencie informacje. W informacji naukowej utożsamiany niekiedy z głównym przedmiotem dokumentu, zna-

[16] *Encyklopedia językoznawstwa ogólnego*, wyd. 2 popr. i uzup., pod red. K. Polańskiego, Wrocław 1999, s. 595.
[17] Tamże, s. 246.
[18] Tamże, s. 163.
[19] Tamże, s. 277.
[20] *Słownik encyklopedyczny informacji...*, s. 301.
[21] Tamże, s. 242.
[22] Tamże, s. 246.

czeniowo najważniejszym, dla omówienia którego powstał dokument[23].

Natomiast termin *hasło* został w *Słowniku encyklopedycznym informacji…* zdefiniowany wyłącznie w kontekście zastosowania w systemach informacyjno-wyszukiwawczych jako wyrażenie o funkcji porządkującej lub wyszukiwawczej w danym zbiorze informacyjnym (słownik, indeks, tekst, zbiór charakterystyk wyszukiwawczych dokumentów)[24].

Nieco odmiennie definiują owe pojęcia autorzy prowadzący badania w zakresie komputerowego przetwarzania języka naturalnego. Największe różnice dotyczą terminów *słowo*, *wyraz* oraz *hasło*. Definicja słownikowa utożsamia ze sobą dwa terminy: *słowo* oraz *wyraz*. Natomiast w pracach lingwistycznych spotykamy wyraźne zróżnicowanie znaczeń przypisywanych obu pojęciom. Jadwiga Sambor definiuje *słowo* jako jednostkę tekstu (lub języka) wyodrębnianą w procedurze segmentacyjnej, odpowiadającą w większości przypadków ciągowi liter pomiędzy odstępami. Z kolei termin *wyraz* wspomniana badaczka traktuje jako pojęcie nadrzędne do terminów *słowo*, *słowoforma* i *leksem*. W pracach J. Sambor termin *wyraz* jest używany zamiast wskazanych trzech terminów w kontekście wskazującym jednoznacznie rodzaj zastępowanej jednostki[25].

Termin *hasło* w pracach dotyczących przetwarzania języka naturalnego jest definiowany jako zwyczajowo przyjęta w leksykografii danego języka forma gramatyczna leksemu (np. bezokolicznik dla czasowników w j. polskim). Pojęcie *hasło* można zdefiniować również jako zbiór słowoform reprezentowany przez określoną postać danej słowoformy[26].

Przytoczone powyżej terminy są podstawowymi pojęciami stosowanymi w przetwarzaniu tekstów. Określają one m.in. jednostki badania statystycznego wyrażeń języka naturalnego. Jednostki te odznaczają się konkretnymi cechami statystycznymi, które zostaną zaprezentowane poniżej.

Należy przy okazji odnotować pewne różnice terminologiczne pomiędzy językiem polskim a angielskim, które wynikają z różnic typów obu języków[27]. W angielskiej literaturze przedmiotu termin *token* bardzo

[23] Tamże, s. 272.

[24] Tamże, s. 76.

[25] R. Hammerl, J. Sambor, *Statystyka…*, s. 17–19; tychże, *O statystycznych…*, s. 21–22. Definicję techniczną terminu *słowo*, jako ciągu znaków pomiędzy dwiema spacjami, przyjmuje również A. Mykowiecka, dz. cyt., s. 67.

[26] R. Hammerl, J. Sambor, *Statystyka…*, s. 18; tychże, *O statystycznych…*, s. 21.

[27] Odnotowanie różnic terminologicznych jest o tyle sensowne i usprawiedliwione, że same badania przetwarzania języka naturalnego zostały rozpoczęte w krajach anglo-

często jest wykorzystywany do różnych graficznie postaci tego samego wyrazu. Termin *word token* z kolei jest stosowany na określenie każdego wystąpienia wyrazu w tekście (z uwzględnieniem powyższej uwagi). Natomiast w celu oznaczenia różnych znaczeniowo słów jest stosowane pojęcie *word type*. W terminologii polskiej angielskiemu pojęciu *token* odpowiadają terminy *słowo*/*wyraz*, natomiast terminowi *word type* odpowiada *hasło* (*wyraz słownikowy*). Pewne wątpliwości znaczeniowe mogą pojawić się również dla pojęcia *term* (*termin*). Powszechnie przyjętą definicją tego pojęcia w języku polskim jest wyrażenie o ściśle ustalonym znaczeniu w danej dziedzinie. Natomiast w tekstach anglojęzycznych poświęconych NLP określenie *term* wydaje się stosowane zamiennie z określeniem *word type* dla oznaczenia każdego odmiennego znaczeniowo wystąpienia danego słowa. Bardzo często we wzorach związanych z przetwarzaniem tekstów języka naturalnego można spotkać oznaczenie *t* (jako skrót od *term*) pokrywające się znaczeniowo z pojęciem *word type*[28].

Cechy statystyczne jednostek leksykalnych

Jednostki tekstu lub języka w danej zbiorowości statystycznej mogą być badane kwantytatywnie ze względu na określoną cechę statystyczną X. Różne realizacje liczbowe x_i tej cechy w przypadku poszczególnych badanych jednostek odwzorowują ich zróżnicowanie pod kątem danej cechy X. Owe cechy statystyczne, ze względu na sposób ich zróżnicowania, można podzielić na:

– cechy ilościowe, które z kolei można podzielić na ciągłe (mierzalne – w danym przedziale wartości zmienne mogą przyjmować dowolne wartości liczbowe) lub skokowe (przeliczalne – w danym przedziale wartości zmienne mogą przyjmować tylko określone wartości liczbowe, np. liczby naturalne) – w badaniach lingwistycznych

saskich (głównie USA), a poziom zaawansowania tych badań dla języka angielskiego jest najwyższy. Z powodu prymatu krajów anglojęzycznych w owych badaniach stosowana w nich terminologia jest oryginalnie pochodzenia angielskiego.

[28] Por. m.in. Ch. D. Manning, H. Schütze, dz. cyt.; D. Jurafsky, J. H. Martin, *Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, New Yersey 1999; P. Jackson, I. Moulinier, *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, Amsterdam–Philadelphia 2002; Ch. D. Manning, P. Raghavan, H. Schütze, *An introduction to Information Retrieval*, Cambridge 2009.

częściej analizuje się cechy przeliczalne;
– cechy jakościowe, których wartości zmiennych nie wyraża się liczbami, np. stosunek liczebności leksemów rodzimych i obcych w słownictwie, zdania złożone – współrzędnie i podrzędnie[29].

Podstawową kategorią stosowaną w ilościowych obliczeniach statystycznych jest częstość absolutna (frekwencja) **F**. Jest to wskaźnik liczbowy otrzymany drogą sumowania jednostek wchodzących w skład danej próby. Podstawą sumowania mogą być wystąpienia poszczególnych jednostek bądź też wartości konkretnej cechy określającej dane jednostki. Częstość występowania słów w tekście jest cechą ilościową przeliczalną, o wartościach wyrażanych za pomocą liczb naturalnych.

Częstość absolutną **F** można przedstawić za pomocą następującego wzoru:

$$F = \sum_{i=1}^{n} f_i$$

Wzór 1: *Wzór na częstość absolutną wystąpień danego słowa*,
gdzie:
F – częstość absolutna,
n – liczebność zbioru analizowanych dokumentów,
f_i – częstość wstąpienia danego słowa w kolejnym dokumencie[30].

Ze względu na to, że najczęściej analiza statystyczna obejmuje wiele tekstów z danej dziedziny, można wprowadzić dodatkową kategorię, jaką jest częstość średnia , określana wzorem:

$$\overline{f} = \frac{F}{n} = \frac{\sum_{i=1}^{n} f_i}{n}$$

Wzór 2: *Wzór na częstość średnią*,
gdzie:
F – częstość absolutna[31].

[29] R. Hammerl, J. Sambor, *Statystyka...*, s. 19; M. Sobczyk, *Statystyka...*, s. 12–13, 92–113.
[30] *Słownik frekwencyjny polszczyzny współczesnej*, oprac. I. Kurcz i in., pod red. Z. Saloniego, Kraków 1990, s. l.
[31] Tamże.

Dla korpusów zróżnicowanych wewnętrznie podaje się wskaźniki określające zróżnicowanie frekwencji danej jednostki w poszczególnych częściach korpusu. Podstawowym wskaźnikiem równomierności rozkładu jest dyspersja. Dyspersja (rozrzut) danej cechy mierzalnej opisuje zróżnicowanie jednostek badanego zbioru ze względu na tę cechę. Podstawowe miary dyspersji odzwierciedlają rozrzut wartości danej cechy wokół średniej arytmetycznej w badanym zbiorze. Jedną ze stosowanych w statystyce miar zmienności jest odchylenie standardowe **s**, które określa przeciętne odchylenie częstości danej jednostki od częstości średniej dla całego zbioru. Odchylenie standardowe jest określane wzorem[32]:

$$s = \frac{\sum_{i=1}^{n} f_i - \overline{f}^2}{n-1}$$

Wzór 3: *Wzór na odchylenie standardowe*.

Kolejną miarą jest współczynnik zmienności **v** określający relatywne odchylenie frekwencji danego elementu od częstości średniej[33]:

$$v = \frac{s}{\overline{f}}$$

Wzór 4: *Wzór na współczynnik zmienności*.

Jednakże, jak podają redaktorzy *Słownika frekwencyjnego...*, miary owe są w zbyt dużym stopniu zależne od wartości średniej, w związku z czym na potrzeby własnych badań wprowadzili wskaźnik dyspersji złożonej. Dyspersja złożona **D**, dostosowana do korpusu tekstów, jest wyrażana wzorem[34]:

$$D = 100(1 - \frac{v}{\sqrt{n}}) = 100(-\frac{\sqrt{\sum_{i=1}^{n} f_i - \overline{f}^2}}{\sqrt{n(n-1)\,\overline{f}}}$$

Wzór 5: *Wzór na dyspersję złożoną danego słowa*.

[32] Tamże.

[33] Tamże.

[34] O dyspersji słownictwa por. tamże, s. li; R. Hammerl, J. Sambor, *Statystyka...*, s. 50 i nast.; J. Sambor, *Językoznawstwo statystyczne dla pracowników informacji naukowej*, Warszawa 1978, s. 51–53. O miarach zmienności w statystyce por. też M. Sobczyk, *Statystyka...*, s. 45–53.

Analiza statystyczna elementów z określoną cechą statystyczną w systemie pozwala ustalić tzw. udział elementów w systemie, zwany również częstością relacyjną **U**. Częstość relacyjna jest wyrażona następującym wzorem[35]:

$$U = \frac{F \cdot D}{100}$$

Wzór 6: *Wzór na częstość relacyjną danego słowa*.

Autorzy *Statystyki dla językoznawców* jako przykłady udziałów podają m.in. udział słownictwa częstego lub rzadkiego w tekście czy też udział słownictwa rodzimego i obcego w określonym słowniku danego języka[36].

Podsumowanie

Metody statystyczne ze względu na relatywnie niskie koszty oraz możliwość całkowitej automatyzacji ich wykorzystania są powszechnie i z dobrymi rezultatami stosowane w badaniach nad komputerowym przetwarzaniem języka naturalnego. W niniejszym artykule podjęto próbę przybliżenia czytelnikowi podstaw analizy statystycznej oraz ujednolicenia terminologii stosowanej przy badaniach frekwencyjnych nad tekstami języka naturalnego. Możliwość potraktowania poszczególnych słów jako elementów powiązanych łatwymi do komputerowego przetwarzania relacjami statystycznymi pozwala na znaczne uproszczenie procesu analizy treści dokumentów, zapewniając jednocześnie wyniki mieszczące się w przedziale tolerancji wartości.

A statistical approach to the natural language processing

Abstract

The article is an introduction to a statistical approach to natural language processing. The quantitative linguistics as a research discipline as well as text units ap-

[35] *Słownik frekwencyjny…*, s. li.

[36] O określaniu wartości średnich oraz ich odchyleń w badaniach językoznawczych por. R. Hammerl, J. Sambor, *Statystyka…*, s. 44–72.

plicable to statistical research have been presented. Definitions of the particular text units have been discussed in terms of their applicability to statistical natural language processing, with special attention to differences in Polish and English terminology. Statistical attributes of lexical units have also been presented as well as categories and measures used in quantitative lexical units research.