INVESTIGATIONES LINGUISTICAE VOL. XXIII, 2011

AL. NIEPODLEGŁOŚCI 4, 60-874, POZNAŃ – POLAND

Klasyfikacja operatorów metatekstowych i częstość ich występowania w krótkich tekstach naukowych w języku polskim

Agnieszka Czoska

INSTYTUT PSYCHOLOGII UNIWERSYTETU IM. A. MICKIEWICZA W POZNANIU, UL. SZAMARZEWSKIEGO 89

aczoska@amu.edu.pl

Streszczenie

The article presents an analysis of the usage frequency of different types of metatext markers in short scientific texts written in Polish. A well-known classification by Hyland (1998, 2005) was used with additional binary classifications by Bunton (1999) and Dahl (2004). Data mining was performed on the data using rule-generating algorithm *OneRule*, decision tree J48, bayesian Naive Bayes Classifier and k-Neares Neighbour classifier, in order to analyse relations between the classes of metatext markers found in the texts. The outcomes of the analysis may be used to simplify classification of metatext markers. Information on metatext markers classes frequency may also be used for preparing or adapting texts in research on the influence of metatext markers on reading and, eventually, for automatic text structure analysis and abstract generation.

1 Metatekst i operatory metatekstowe

W literaturze dotyczącej tych aspektów tekstu pisanego, które mają przede wszystkim funkcje metatekstową (dotyczą tekstu, w którym są zawarte, stanowią odnośniki do literatury) stosuje się różne terminy na określenie tak wyróżnionych elementów. Badacze piszą o metatekście (Lemarié, Robert F. Lorch, Eyrolle i Virbel, 2008) i metadyskursie (Hyland, 1998; Bunton, 1999) oraz obiektach tekstowych pełniących funkcję metatekstową (Lemarié et al., 2008, s. 29), markerach dyskursowych (Fraser, 1999; M. M. Louwerse i Mitchell, 2003, *discourse markers*), wskazówkach (meta)tekstowych (*text signalling devices*) (Lemarié et al., 2008), markerach koherencji (T. Sanders, Land i Mulder, 2007a, *coherence markers*).

Podawane w literaturze definicje metatekstu (metadyskursu) zgadzają się, że pełni on funkcję metatekstową, jednak różnią nieco w szczegółach dookreślających zakres tej funkcji. Większość tekstów mówi o sygnalizowaniu lub wskazywaniu wprost organizacji (struktury) tekstu (Fraser, 1999; Goldman i John A. Rakestraw, 2000; Lemarié et al., 2008; Hyland, 1998) rozumianej także jako jego koherencja (McNamara, Kintsch, Butler Songer i Kintsch, 1996; T. Sanders et al., 2007a) oraz relacji pomiędzy wskazanymi fragmentami tekstu (Fraser, 1999; M. Louwerse, 2001; M. M. Louwerse i Mitchell, 2003; Hyland, 1998; Lemarié et al., 2008) (czasem określanych jako retoryczne – nawiązujących do teorii struktury retorycznej tekstu RST) (Knott i Dale, 1993, 1996; Marcu, 1997)). Część definicji podaje także jako wyznacznik metatekstowości wskazywanie na autora tekstu lub jego relacje z czytelnikiem (Hyland, 1998; Mur-Dueñas, 2011) lub wyrażanie aktów tekstowych (analogiczne do aktów dialogowych czy dyskursowych (Bunton, 1999, s. 46 - 47)) i intencji autora (Lemarié et al., 2008, s. 31) wraz z jego opiniami (Hyland, 1998; Mur-Dueñas, 2011). W literaturze pojawia się także twierdzenie, że elementy metatekstowe, skoro opisują już istniejącą strukturę tekstu, nie dodają niczego do jego treści – są elementami dodatkowymi – *nonessential* (Lemarié et al., 2008, s. 29) – i mogą być usunięte z tekstu bez szkody dla zawartej w nim informacji (Goldman i John A. Rakestraw, 2000; Abdi, Rizi i Tavakoli, 2010).

Różnice w definiowaniu metatekstu wiążą się także z tym, że badacze zaliczają do metatekstu obiekty znacznie różniące się formą i wielkością. Lemarié i in. (Lemarié et al., 2008) traktują jako obiekty metatekstowe zarówno spójniki jak i większe fragmenty tekstu (np. abstrakt) oraz elementy graficzne składające się na ostateczną postać tekstu. Z kolei Fraser (1999) zalicza tu jedynie kilkuwyrazowe wyrażenia należące do klas gramatycznych spójników, przysłówków i fraz przyimkowych. Podobnie Marcu (1997) definiuje elementu metatekstowe (wyrażające relacje retoryczne RST) jako wyrażenia

regularne, najczęściej jednowyrazowe, łączące frazy. Badacze zajmujący się częstością występowania metatekstu w wybranych tekstach (Hyland, 1998; Bunton, 1999; Mur-Dueñas, 2011) traktują jako metatekstowe nie tylko wyrażenia uwzględniane przez Frasera i Marcu, ale także wskazujące wprost na autora tekstu, jak zaimki pierwszoosobowe.

1.1 Definicja operatora metatekstowego na potrzeby badania tekstów w języku polskim

Analizowane tu operatory metatekstowe można zdefiniować jako wyrażenia o funkcji metatekstowej, umieszczone w ciągu tekstu, będące wskazówkami struktury tekstu lub roli danego fragmentu w większej całości. Stanowią one podzbiór wyrażeń metatekstowych definiowanych przez wymienionych wyżej badaczy, zbliżający się najbardziej do zakresu wyznaczanego przez definicję Frasera (1999) lub zbioru realizowanych tekstowo wskazówek metatekstowych Lemarié (2008). Zaliczenie do operatorów jedynie metatekstu umieszczonego w ciągu tekstu oznacza, że jako operatory nie będą klasyfikowane tytuły rozdziałów i inne wskazówki podziału tekstu na części wyróżnione graficznie (np. spis treści), zaliczające się jednak do metatekstu (Lemarié et al., 2008).

Pojęcie operatora metatekstowego pojawiło się już w polskojęzycznej literaturze (Winiarska, 2001). Winiarska pisze (2001, s. 21): „operatory metatekstowe są instrukcjami wskazującymi odbiorcy, w jaki sposób powinien łączyć kolejny element z tymi, które go poprzedzają, uściślają one i precyzują rodzaj relacji semantycznej, jaka ma łączyć poszczególne elementy" – co stanowi definicję spójną z podanymi wyżej. Jest też bardzo zbliżona do podawanej przez Frasera (1999), obejmującej spójniki, przysłówki i frazy przyimkowe występujące w funkcji metatekstowej. Może zostać także rozszerzona o czasowniki, gdy dotyczą one aktów tekstowych (Bunton, 1999), natomiast nie będzie obejmować zaimków pierwszoosobowych (Hyland, 1998; Mur-Dueñas, 2011). Niektórzy cytowani poprzednio badacze określają metatekst także jako instrukcje dla czytelnika (Goldman i John A. Rakestraw, 2000; McNamara et al., 1996; Lemarié et al., 2008; Fraser, 1999).

1.2 Cel badań nad metatekstem

Motywacja leżąca u podstaw cytowanych prac nad metatekstem związana jest z badaniami językoznawczymi i tekstologicznymi (Fraser, 1999; Lemarié et al., 2008), automatyczną analiza i generowaniem tekstów (Knott i Dale, 1993, 1996; Marcu, 1997) oraz badaniami psycholingwistycznymi zadającymi pytanie o wpływ obecności metatekstu (oraz spójności tekstu) na czytelnika

(McNamara et al., 1996; Goldman i John A. Rakestraw, 2000; T. J. M. Sanders i Noordman, 2000; Degand i Sanders, 2002; T. Sanders et al., 2007a). Pojawia się także coraz więcej badań porównawczych, analizujących konwencje stosowania metatekstu w wybranych socjolektach i żargonach naukowych oraz intencje i cele autora tekstu wyrażane w metatekście (Hyland, 1998; Bunton, 1999; Mur-Dueñas, 2011; Saz Rubio, 2011; Aijmer i Simon-Vandenbergen, 2004; Abdi et al., 2010).

Podczas gdy w badaniach tekstologicznych badacze (Fraser, 1999; Lemarié et al., 2008; Knott i Dale, 1993) skupiają się na podaniu definicji elementów metatekstowych oraz skategoryzowaniu ich i zbadaniu relacji (i różnic) pomiędzy znajdowanymi w tekstach markerami, w pracach socjo- i psycholingwistycznych z góry zakładana jest określona kategoryzacja wspomagająca przeszukiwanie tekstów pod kontem metatekstu. Zwykle kategoryzacje takie mają u podstaw wiele założeń dotyczących funkcji metatekstu oraz form, jakie może on przybierać (Hyland, 1998; Mur-Dueñas, 2011). Kategoryzacje te zostaną szczegółowo opisane w następnym rozdziale.

Badania nad wpływem metatekstu na przetwarzanie i zrozumiałość tekstu opierają się z kolei na manipulacji zawartością elementów metatekstowych w tekście. Kategoryzacja zaproponowana przez Lemarié i in. (2008) ma na celu między innymi zaproponowanie charakterystyki wyrażeń metatekstowych (obejmującej także ich potencjalny wpływ na czytelnika) pozwalającej dokładniej określić zakres manipulacji eksperymentalnej w badaniach tego typu, ułatwienie ich replikacji i porównywania wyników.

2 Proponowane w literaturze klasyfikacje operatorów metatekstowych

Wiele opracowań cytowanych w poprzednim rozdziale zawiera propozycje klasyfikacji operatorów lub elementów metatekstowych/metadyskursowych przyjęte na potrzeby analizy tekstu lub badania metatekstu jako takiego. Poniżej zostały przedstawione wybrane kategoryzacje, przy czym uwzględniono jedynie te ich aspekty, które odnoszą się do operatorów rozumianych jako wyrażenia umieszczane w ciągu tekstu, informujące o jego organizacji i relacjach pomiędzy jego częściami.

Przedstawiono tu klasyfikacje najpełniej charakteryzujące elementy metatekstowe, pozwalające przy tym na odróżnienie cech danego elementu jako takiego lub wynikających z jego roli w tekście. Są to typologie dobrze znane i często stosowane w badaniach tekstologicznych nad metatekstem (Abdi et al., 2010). Umożliwiają także interpretację różnych typów metatekstu w ka-

tegoriach intencji autora tekstu i instrukcji dla czytelnika. Z tego powodu stanowią dobry punkt wyjścia do badań nad zróżnicowaniem i częstością występowania metatekstu, jak i jego wpływu na przetwarzanie tekstu i umożliwiają porównanie badań prowadzonych w tradycji tekstologicznej i psycholingwistycznej.

2.1 Operatory tekstowe i interpersonalne

Hyland (1998) i Mur-Dueñas (2001) zastosowali w swoich analizach porównawczych zawartości metadyskursu w tekstach naukowych podział na metatekst o funkcji interpersonalnej albo tekstowej (Hyland, 1998) oraz interakcyjnej albo interaktywnej (Mur-Dueñas, 2011; Hyland, 2005; Hyland i Tse, 2004). Klasyfikacje te różnią nieznacznie zestawem klas, jednak definicje interpersonalności/interakcyjności i tekstowości/interaktywności opierają się w obu przypadkach na tym samym rozróżnieniu: do metadyskursu interpersonalnego zalicza się wyrażenia mające na celu konstruowanie relacji autor-czytelnik (lub autor-społeczność odbiorców), zaś tekstowy niesie informacje o organizacji tekstu. Hyland i Tse zdecydowali się zastosować terminologię metadyskurs interaktywny i interakcyjny, gdyż bardziej odpowiada ona rozwijanej przez nich teorii o pragmatycznej roli metadyskursu jako wskazówki odnoszenia się autora do czytelnika lub całej społeczności, w obrębie której ma funkcjonować tekst. Także wskazówki struktury tekstu są wg. nich interpersonalne w tym sensie, że wskazują czytelnikowi kontekst niezbędny do interpretacji danej informacji (Hyland i Tse, 2004; Hyland, 2005). Klasyfikacja ta wykorzystywana jest na potrzeby badań porównawczych nad stosowaniem metatekstu w różnych społecznościach badawczych, ma stanowić podstawę analizy przyjmowanych w nich stylów argumentacyjnych oraz kreowanych obrazów autora i relacji autor-czytelnik/społeczność.

Zgodnie z zastosowaną na potrzeby tych badań definicją operatora metatekstowego, wyrażenia tego typu należą do metadyskursu interaktywnego, dlatego szczegółowo zostanie tu opisany jedynie podział na klasy ze względu na informacje o strukturze tekstu i relacjach pomiędzy jego fragmentami. W badaniu tym kładzie się nacisk na funkcję metatekstu (metadyskursu) jako wskazówki struktury tekstu, dlatego będzie tu stosowany termin metatekst tekstowy – zgodnie z rozróżnieniem Buntona: *It would seem preferable to reserve the term 'metatext' for this textual function and use 'metadiscourse' for broader definitions (...) which encompass the interpersonal function as well as the textual.* (Bunton, 1999, s. 44).

Klasyfikacja Hylanda (1998, s. 442) zalicza do metatekstu tekstowego (*textual metatext*) następujące wyrażenia:

- spójniki logiczne (w późniejszych artykułach zastosowano termin *transitions*; klasa obejmuje markery relacji retorycznych – addytywnej, kontrastu, wynikania),
- markery aktów dyskursowych – działań autora obejmujących treść i strukturę tekstu (*frame markers*),
- markery endoforyczne (anafory i katafory odnoszące się do innych obiektów tekstowych, wskazujące na zawarte w nich informacje),
- markery synonimiczności lub uszczegółowienia (*code glosses*; wyrażenia jak „na przykład", „mianowicie", „ponadto"),
- przypisy.

Z wyjątkiem przypisów, wszystkie kategorie mieszczą się w przyjętej tu definicji operatora metatekstowego, pozwalając na wyodrębnienie czterech typów operatorów.

Mur-Dueñas (2011, s. 3) dodaje do typów wymienianych u Hylanda:

- wyliczenia (*sequences*, zalicza się tu także podział na rozdziały itp.),
- markery topiku wprowadzające nowy temat, sygnalizujące jego podsumowanie lub zmianę (*topicalisers*).

Do klas wprowadzonych przez Mur-Dueñas mogą należeć obiekty metatekstowe wyróżniające się wizualnie z całości tekstu, nie stanowi to jednak problemu dla opisywanych tu badań, gdyż nie wprowadza ona żadnej klasy do której nie mogą należeć operatory metatekstowe. Propozycja Mur-Dueñas (6 klas wyrażeń metatekstowych) może być traktowana jako podział logiczny, gdyż zakłada, że dany obiekt należy tylko do jednej klasy, zaś zestaw klas jest wyczerpujący w ramach metatekstu tekstowego.

2.2 Klasyfikacje binarne

Interesujące z punktu widzenia automatycznej analizy danych (*data mining*) są proste, binarne kategoryzacje zaproponowane przez Buntona (1999) oraz Dahl (2004). Umożliwiają one skupienie się jedynie na jednym wymiarze różnicującym elementy metatekstowe, ale także wyróżnienie go i porównanie z innymi, jeśli zostaną zestawione z inną klasyfikacją.

2.2.1 Metatekst lokalny i globalny

Podział obiektów metatekstowych na lokalne i globalne stanowi u Buntona (1999) ostatni etap kategoryzacji metatekstu znalezionego w tekstach prac doktorskich, wcześniej podzielonego na typy według podziału podobnego

do stosowanych przez Hylanda i Mur-Dueñas (Bunton, 1999, s. 48, 45). Można go stosować w celu doprecyzowania pojęcia zasięgu w analizie Lemarié i in.

Bunton określa jako globalny każdy element metatekstowy, którego zasięg (rozumiany jako wielkość obiektu odniesienia) lub odległość od obiektu odniesienia są na poziomie rozdziału lub całości tekstu[1] (Bunton, 1999, s. 50). Podział metatekstu na globalny/lokalny jest w tym rozumieniu bardziej jednoznaczny, niż ogólne stwierdzenia o bliskości czy wielkości obiektów (Lemarié et al., 2008, s. 34). Klasyfikacja ta jest o tyle interesująca, że nie opisuje cech operatora jako takiego, ale wymaga oceny wielkości obiektu (obiektów) odniesienia operatora i jego odległości od wskazywanego fragmentu (jedynie w przypadku wskazówek endoforycznych, odwołujących się wprost do całości tekstu lub rozdziału, informacja o zasięgu jest dostępna natychmiast, w samym operatorze).

2.2.2 Metatekst retoryczny i lokalizujący

Dahl (2004) stosuje w swoich badaniach podział na metatekst lokalizujący i retoryczny, nadbudowany nad klasyfikacją Hylanda. Operator lokalizujący musi zawierać odniesienie do konkretnego fragmentu tekstu (z użyciem jego etykiety – „w rozdziale pierwszym", lub z mniejszym stopniem precyzji – „jak wspomniano wyżej"). z kolei wyrażenie zaliczane do metatekstu retorycznego wyraża wprost informację o akcie retorycznym dokonanym przez autora („podsumowując", „stwierdzamy"). Dla tej klasyfikacji decydująca jest realizacja słowna metatekstu, zliczana do formy operatora (Lemarié et al., 2008).

O ile operatory lokalizujące są łatwe do wyodrębnienia z tekstu i nie powinno być wątpliwości co zaliczenia wyrażenia do tej kategorii, operatory retoryczne mogą sprawiać wrażenia gorzej określonego typu. Jednak jeśli wziąć pod uwagę to, że wszystkie cytowane powyżej teorie metatekstu nadają mu funkcje wskazywania relacji pomiędzy elementami tekstu (lub jego struktury), można stwierdzić, że kategoryzacja proponowana przez Dahl spełnia wymogi podziału logicznego – ma dobrze zdefiniowaną klasę „lokalizujące" oraz klasę „inne". Operatory retoryczne w tym rozumieniu to wszystkie, które nie zawierają dodatkowej, szczegółowej informacji o zasięgu lub lokalizacji obiektu odniesienia.

W przeciwieństwie do klasyfikacji Buntona, podział na metatekst retoryczny i lokalizujący można przeprowadzić na operatorach wyekstrahowanych z tekstu, przy czym wciąż zawiera on informację o sposobie odnoszenia się

[1]Parafraza *either the scope or the distance is at chapter or thesis level* (Bunton, 1999, s. 50).

do obiektów odniesienia metatekstu. W dalszym badaniu zostanie przedstawione zestawienie tych dwóch klas, co stanowi część odpowiedzi na pytanie o relacje pomiędzy formą operatora i sposobem jego stosowania.

3 Badanie częstości występowania operatorów metatekstowych różnych typów w krótkich tekstach pokonferencyjnych

Przebadano 62 krótkie artykuły, z których jeden nie wszedł do przedstawionej tu analizy (nie zawierał żadnych operatorów metetekstowych). Teksty pochodziły z dwóch publikacji pokonferencyjnych Poznańskiego Forum Kognitywistycznego – studencko-doktoranckiej konferencji naukowej. Pochodzą one z lat 2009 i 2010 (4. i 5. edycja konferencji)[2] . Przeanalizowane teksty są krótkie (mają do 10 stron standardowego maszynopisu), dotyczą różnych zagadnień związanych z kognitywistyką – zaliczają się do tematyki psychologicznej, językoznawczej, informatycznej, logicznej i innych. Część artykułów ma charakter raportu z badania, część – analizy teoretycznej wybranego problemu.

Operatory wyodrębniono automatycznie, za pomocą napisanego na potrzeby badania programu wyszukującego słowa kluczowe w tekście[3], następnie sprawdzono, ile ze znalezionych wyrażeń rzeczywiście pełni funkcję metatekstową i odnosi się do tekstu, w którym są zawarte.

3.1 Znalezione operatory i ich klasyfikacja

W badaniu potraktowano operatory w sposób nieco niestandardowy, traktując jako osobne przypadki wystąpienia takich wyrażeń, jak „powyżej", „wcześniej" i „wspomniany" czy „rozdział", które w tekście często występują razem. Zdecydowano się na taki zabieg ze względu na większą przejrzystość klasyfikacji i danych oraz możliwość analizy wystąpień poszczególnych klas operatorów, a nie ich współwystąpień. W samych tekstach pojawiły się bardzo różnorodne zestawienia operatorów, często wyznaczone w powyższy sposób operatory występowały samodzielnie.

W tekstach znaleziono 48 rodzajów wyrażeń mogących pełnić rolę operatorów metatekstowych. Większość z nich występowała w różnych formach

[2]Pliki PDF z publikacjami dostępne są pod adresem **http://pfk.wikidot.com/nasze-wydawnictwa**. Obydwie publikacje datowane są na 2010 rok, konferencje odbyły się jednak w 2009 i 2010 roku, zaś artykuły zostały napisane przed konferencjami.

[3]Autorem programu jest Adam Kupś, doktorant w Instytucie Psychologii UAM

gramatycznych – do jednego rodzaju zaliczano różne formy rzeczownika czy przymiotnika, oraz czasownika – osobno formy osobowe czynne i bierne oraz formy bezosobowe. Operatory wraz z częstościami występowania w tekstach spisane zostały w Tabeli 1. Następnie poklasyfikowano znalezione operatory zgodnie z klasycznymi propozycjami przytoczonymi wyżej. Klasyfikacje operatorów przedstawia Tabela 2.

3.2 Analiza statystyczna

3.2.1 Częstość występowania operatorów

Analizy statystyczne dotyczą 61 tekstów (jedynie tych, w których odnaleziono operatory metatekstowe).Przebadane teksty miały od 1532 do 5672 słów (M=2835,39; SD=831,991)[4], od 1 do 10 sekcji (rozdziałów lub podrozdziałów; M=4,80; SD=1,711), zawierały 1 – 52 operatorów metatekstowych (M=14,39; SD=10,341; średnio 5,16 operatora na 1000 słów).

Katafory wystąpiły w 25 tekstach (M=1,20; SD=2,235; Max=12), średnia proporcja liczby katafor do sumy operatorów w tekście wyniosła 0,096.

Rozkład poszczególnych klas operatorów według podziału na metatekst tekstowy i interpersonalny przedstawia się następująco: najliczniejszą klasą były markery endoforyczne (393 operatory, od 0 do 33 w tekście; M=6,44; SD=5,584), najmniej liczną (39 wystąpień) – markery aktów dyskursowych (0 – 4 w tekście; M=0,64; SD=0,876). Znaleziono 145 markerów wyliczenia (0 – 10; M=2,38; SD=2,222), 123 markery topiku (0 – 17; M=2,02; SD=2,668), 78 spójników logicznych (0 – 9; M=1,28; SD=1,762) oraz 69 wskazówek synonimiczności (0 – 9; M=1.13; SD=2,021).

We wszystkich tekstach znaleziono 392 operatory globalne oraz 485 operatorów lokalnych (według Buntona (1999)), teksty zawierały od 1 do 29 operatorów globalnych (M=6,88; SD=5,418; średnia proporcja liczby operatorów globalnych do wszystkich wyniosła 0,465) oraz 1 – 35 lokalnych (M=8,22; SD=6,701; średnia proporcja =0,577). Średnia proporcja liczby operatorów lokalnych do globalnych wyniosła 2,1191 (SD=2,3456).

Operatorów lokalizujących i retorycznych (według Dahl (2004)) znaleziono odpowiednio 403 i 475, lokalizujących 0 – 33 w tekście (M=6,61; SD=5,572), retorycznych – od 1 do 40 (M=7,79; SD=6,778). Średnia proporcja liczby operatorów retorycznych do lokalizujących wyniosła 1,612 (SD=1,864).

[4]We wszystkich przypadkach M oznacza średnią, zaś SD – odchylenie statystyczne.

Tablica 1: *Znalezione operatory z liczbą wystąpień w przebadanych tekstach*

typ operatora	liczba wystąpień we wszystkich tekstach	liczba wystąpień w pojedynczym tekście (max)	tekstów zawierających (%)
artykuł	71	9	26,23
części	64	7	21,31
dokonany	3	1	4,92
drugi	55	4	22,95
jak widać	5	3	4,92
jak wspomniano	4	3	1,64
jak wynika	2	1	3,28
mianowicie	11	2	14,75
między innymi	27	5	26,23
na podstawie	21	5	19,67
następnie	12	2	14,75
opisano	5	2	6,56
opisany	3	1	4,92
oznacza to	12	3	11,48
paragraf	10	9	1,64
pierwszy	31	4	34,43
po drugie	20	2	26,23
po pierwsze	23	3	26,23
po trzecie	4	1	6,56
pod uwagę	7	2	9,84
podrozdział	4	2	6,56
podsumowane	1	1	1,64
podsumowując	20	3	22,95
poniżej	11	1	18,03
przykład	58	8	19,67
przytoczony	12	2	16,39
porównaj	5	2	6,56
poruszony	4	1	6,56
poświęcony	1	1	1,64
powiedziane	1	1	1,64
powyżej	32	6	11,48
praca	43	7	16,39
problem	22	7	8,20
przedstawiony	34	7	14,75
z (...) strony	34	6	11,48
vide	1	1	1,64
wcześniej	24	3	22,95
wniosek	26	2	27,87
wnioskować	6	2	8,20
wspomniany	12	6	18,03
wspomnieć	19	2	18,03
wykonany	1	1	1,64
wynika to	4	1	6,56
wyżej	40	3	40,98
zagadnienie	23	3	26,23
zarysowany	2	2	1,64
zgodnie z	41	7	14,75
znaczy to	1	1	1,64

Tablica 2: *Klasyfikacja operatorów metatekstowych.*

operator	liczba wystąpień z zasięgiem globalny	liczba wystąpień z zasięgiem lokalny	liczba wystąpień z funkcją katafory	operator retoryczny lub lokalizacyjny	funkcje tekstowe
artykuł	71	0	20	lokalizacyjny	endofora
części	63	1	14	lokalizacyjny	endofora
dokonany	3	0	0	lokalizacyjny	endofora
drugi	8	47	1	retoryczny	wyliczenie
jak widać	1	4	0	retoryczny	spójnik logiczny
jak wspomniano	3	1	0	lokalizacyjny	wyliczenie
jak wynika	0	2	0	retoryczny	spójnik logiczny
mianowicie	0	11	0	retoryczny	synonimiczny
między innymi	0	27	0	retoryczny	wyliczenie
na podstawie	11	10	0	retoryczny	topik
następnie	10	2	5	lokalizacyjny	wyliczenie
opisano	5	0	1	lokalizacyjny	endofora
opisany	1	2	0	lokalizacyjny	endofora
oznacza to	0	12	0	retoryczny	spójnik logiczny
paragraf	9	1	0	lokalizacyjny	endofora
pierwszy	5	26	1	retoryczny	wyliczenie
po drugie	3	17	0	retoryczny	wyliczenie
po pierwsze	4	19	0	retoryczny	wyliczenie
po trzecie	0	4	0	retoryczny	wyliczenie
pod uwagę	2	5	0	retoryczny	topik
podrozdział	4	0	2	lokalizacyjny	endofora
podsumowane	1	0	0	lokalizacyjny	endofora
podsumowując	10	10	0	retoryczny	akt dyskursowy
poniżej	1	10	1	lokalizacyjny	endofora
porównaj	1	4	0	retoryczny	topik
poruszony	3	1	0	lokalizacyjny	endofora
poświęcony	1	0	0	lokalizacyjny	endofora
powiedziane	1	0	0	lokalizacyjny	endofora
powyżej	18	14	0	lokalizacyjny	endofora
praca	43	0	16	lokalizacyjny	endofora
problem	9	13	1	retoryczny	topik
przedstawiony	21	13	5	lokalizacyjny	endofora
przykład	7	51	2	retoryczny	synonimiczny
przytoczony	7	5	2	lokalizacyjny	endofora
strony	0	34	0	retoryczny	spójnik logiczny
vide	1	0	0	retoryczny	topik
wcześniej	16	8	0	lokalizacyjny	endofora
wniosek	9	17	1	retoryczny	spójnik logiczny
wnioskować	1	5	0	retoryczny	akt dyskursowy
wspomniany	5	7	0	lokalizacyjny	endofora
wspomnieć	2	17	1	retoryczny	akt dyskursowy
wykonany	0	1	0	lokalizacyjny	endofora
wynika to	0	4	0	retoryczny	spójnik logiczny
wyżej	13	27	0	lokalizacyjny	endofora
zagadnienie	9	14	0	retoryczny	topik
zarysowany	2	0	0	lokalizacyjny	endofora
zgodnie z	13	28	0	retoryczny	topik
znaczy to	0	1	0	retoryczny	spójnik logiczny

Tablica 3: *Różnice grupowe dla zmiennej zależnej liczba katafor*

zmienne niezależne liczba	zmienne zależne grupujące liczba katafor 1	2	p**
słów	M=2924; SD=879,226	M=2508,23; SD=534,745	0,384
sekcji	M=4,75; SD=1,631	M=5; SD=2,041;	0,911
operatorów	M=12,58; SD=8,97*	M=21,08; SD=12,573*	0,011
liczba operatorów			
kataforycznych	-	-	-
lokalnych	M=7,74; SD=6,781*	M=10.08; SD=6,302*	0,008
globalnych	M=5,43; SD=4,06*	M=11,77; SD=6,66*	0,001
lokalizacyjnych	M=5,33; SD=3,755*	M=11,31; SD=8,341*	0,041
retorycznych	M=7,25; SD=6,871*	M=9,77; SD=6,274*	0,019
liczność klas operatorów tekstowych			
spójniki logiczne	M=1,35; SD=1,839	M=1; SD=1,472	0,272
endofory	M=5,21; SD=3,73*	M=11; SD=8,534*	0,017
markery synonimiczności	M=o,6; SD=0,893	M=3,08; SD=3,475	0,356
markery aktów dysk.	M=0,65; SD=0,876	M=0,62; SD=0,768	0,984
markery topiku	M-2,04; SD=2,729	M=1,92; SD=2,532	0,33
markery wyliczenia	M=2.21; SD=2,269	M=3; SD=2	0,228

* różnice istotne statystycznie

** poziom istotności dla różnicy median

3.2.2 Różnice grupowe

Wykonano testy różnic rozkładów oraz median zmiennych niezależnych w grupach czterech zmiennych zależnych: metatekstowych w tekście, oraz liczby operatorów lokalnych, globalnych oraz katafor. Wybrano te zmienne, gdyż opisują one tekst jako taki, lub rolę danych operatorów w tekście, a nie ich charakter w oderwaniu od tekstu. Podziału na grupy dokonano na podstawie średniej ilości operatorów w tekstach: do pierwszej w każdym przypadku zaliczono teksty zawierające poniżej średniej ilości operatorów, do drugiej – pozostałe. W przypadku liczby katafor grupa pierwsza obejmuje teksty, w których nie znaleziono żadnej katafory, lub tylko jedną, co może ograniczać interpretację danych. Zastosowano testy nieparametryczne ze względu na odmienność rozkładów wartości zmiennych od rozkładu normalnego. Podsumowanie wyników znajduje się w Tabeli 3 – 6[5]. Dla wszystkich zmiennych grupujących mediany zmiennych zależnych są wyższe w grupie powyżej średniej niż w grupie poniżej średniej.

[5]W tabelach 3 – 6 M oznacza medianę, zaś SD – odchylenie standardowe

Tablica 4: *Różnice grupowe dla zmiennej zależnej suma operatorów*

zmienne niezależne liczba	zmienne zależne grupujące suma operatorów 1	2	p**
słów	M=2720,86;SD=757,717	M=2989,58;SD=918,455	0,53
sekcji	M=4,94;SD=1,83	M=4,62;SD=1,551	0,852
operatorów	-	-	-
liczba operatorów			
kataforycznych	M=0,49;SD=0,919	M=2,15;SD=3,029	0,78
lokalnych	M=4,15;SD=2,526*	M=13,38;SD=6,812*	0
globalnych	M=3,97; SD=2,846*	M=10,35; SD=5,748*	0
lokalizacyjnych	M=3,17;SD=2,607*	M=10,5;SD=6,147*	0
retorycznych	M=3,74;SD=2,343*	M=13.23;SD=7,005*	0
liczność klas operatorów tekstowych			
spójniki logiczne	M=0,57;SD=0,884*	M=2,23;SD=2,178*	0,002
endofory	M=3,60;SD=2,546*	M=10,27;SD=6,284*	0
wskazówki synonimiczności	M=0,40;SD=0,775*	M=2,12;SD=2,688*	0,002
wskazówki aktów dysk.	M=0,31;SD=0,471*	M=1,08;SD=1,093*	0,008
wskazówki topiku	M=1,14;SD=1,089*	M=3,19;SD=3,6*	0,01
wskazówki wyliczenia	M=1,23;SD=1,19*	M=3,92;SD=2,365*	0

* różnice istotne statystycznie

** poziom istotności dla różnicy median

Tablica 5: *Różnice grupowe dla zmiennej zależnej liczba oeratorów globalnych*

zmienne niezależne liczba	zmienne zależne grupujące liczba operatorów globalnych 1	2	p**
słów	M=2662,72;SD=730,692	M=3067,81;SD=915,023	0,096
sekcji	M=4,53;SD=1,754	M=5,12;SD=1,633	0,451
operatorów	M=9,17;SD=5,742*	M=21,42;SD=11,057*	0
liczba operatorów			
kataforycznych	M=0,49;SD=0,887*	M=2,15;SD=3,042*	0,037
lokalnych	-	-	-
globalnych	-	-	-
lokalizacyjnych	M=3,63;SD=2,745*	M=10,62;SD=5,927*	0
retorycznych	M=5.54;SD=4,111*	M=10,81;SD=8,410*	0,002
liczność klas operatorów tekstowych			
spójniki logiczne	M=1,03;SD=1,74	M=1,62;SD=1,768	0,382
endofory	M=3,15;SD=2,683*	M=10,38;SD=6,073*	0
wskazówki synonimiczności	M=0,54;SD=0,886	M=1,92;2,756	0,127
wskazówki aktów dysk.	M=0,46;SD=0,701	M=0,88;SD=1,033	0,111
wskazówki topiku	M=1,57;SD=1,42	M=2,62;SD=3,669	0,996
wskazówki wyliczenia	M=1,77;SD=1,629	M=3,19;SD=2,654	0,046

* różnice istotne statystycznie

** poziom istotności dla różnicy median

Tablica 6: *Różnice grupowe dla zmiennej zależnej liczba operatorów lokalnych*

zmienne niezależne liczba	zmienne zależne grupujące liczba operatorów lokalnych 1	2	p**
słów	M=2752,97;SD=805,875	M=2981,5;SD=876,123	0,529
sekcji	M=5,05;SD=1,835	M=4,36;SD=1,399	0,145
operatorów	M=9,03;SD=5,188*	M=23,91;SD=10,415*	0
liczba operatorów			
kataforycznych	M=0,85;SD=1,565	M=1,82;SD=3,034	0,594
lokalnych	-	-	-
globalnych	-	-	-
lokalizacyjnych	M=4,97;SD=4,139*	M=9,5;SD=6,631*	0
retorycznych	M=4,05;SD=2,416*	M=14,41;SD=6,987*	0
liczność klas operatorów tekstowych			
spójniki logiczne	M=0,62;SD=0,935*	M=2,45;SD=2,241*	0,001
endofory	M=4,87;SD=4,137*	M=9,23;SD=6,74*	0,015
wskazówki synonimiczności	M=0,28;SD=0,605*	M=2,64;SD=2,7*	0
wskazówki aktów dysk.	M=0,44;SD=0,552	M=1;SD=1,195	0,309
wskazówki topiku	M=1,13;SD=1,005*	M=3,59;SD=3,8*	0,002
wskazówki wyliczenia	M=1,44;SD=1,334*	M=4,05;SD=2,516*	0,001

* różnice istotne statystycznie

** poziom istotności dla różnicy median

3.3 Regresja liniowa

Przeprowadzono także analizę regresji dla wyżej wymienionych zmiennych zależnych. Wyniki analizy regresji znajdują się w tabeli 7.

Zmienna suma operatorów korelowała najsilniej ze zmienną niezależną liczba operatorów retorycznych i liczba endofor (r^2=0,629) – najliczniejszymi klasami operatorów. Dla zmiennej liczba katafor testy wskazały największy współczynnik zależności liniowej ze zmienną liczba operatorów lokalizujących, ale bardzo zbliżone współczynniki korelacji zostały znalezione pomiędzy tą zmienną, a liczbą operatorów globalnych i endofor. Można potraktować to jako wskazówkę, które z klas operatorów najczęściej stosowane były jako kataforyczne. Najsilniejsze korelacje w przypadku zmiennej zależnej liczba operatorów globalnych zostały wskazane dla zmiennych liczba operatorów lokalizacyjnych i endofor (będących podzbiorem klasy operatorów lokalizujących), co może sugerować, że te klasy operatorów najczęściej miały globalny zasięg w przypadku przebadanych tekstów. z kolei zmienna zależna liczba operatorów lokalnych korelowała najsilniej ze zmienną liczba operatorów retorycznych (r^2=0,873) i sumą operatorów (r^2=0,771). Może to świadczyć o tym, że operatory retoryczne najczęściej przyjmowały zasięg lokalny, oraz, że częstość stosowania operatorów lokalnych jest silnie związana ze skłonnością autora do stosowania metatekstu w ogóle.

Tablica 7: *Regresja liniowa dla zmiennych opisujących liczbę operatorów w tekstach*

zmienna zależna	zmienna niezalezna	r^2	błąd standardowy
liczba operatorów globalnych	katafory	0.436*	4,106
	suma	0.635*	3,304
	sekcje	0.019	5,416
	słowa	0.076*	5,255
	endofory	0.802*	2,434
	topik	0.061	5,3
	frame	0.094	5,204
	synonimy	0.201	4,887
	wyliczenie	0.115	5,142
	sp.log	0.034	5,374
liczba operatorów lokalnych	suma	0.771	3,233
	słowa	0.081	6,481
	sekcji	0.006	6,738
	katafory	0.051	6,584
	sp.log	0.342	5,481
	endofory	0.239	5,896
	synonimy	0.177	6,132
	frame	0.298	5,662
	topik	0.441	5,053
	wyliczenie	0.619	4,172
	lokalizacyjne	0.250	5,854
suma operatorów	retoryczne	0.755*	5,158
	słowa	0.108*	9,851
	sekcje	0	10,428
	katafory	0.246*	9,057
liczba katafor	globalne	0.436*	1,74
	słowa	0.23	2,228
	sekcje	0.013	2,239
	suma	0.246	1,957
	lokalizujące	0.442*	1,683
	retoryczne	0.044	2,203
	endofory	0.435*	1,694

*istotność statystyczna

3.4 Klasyfikacja tekstów i operatorów metatekstowych z zastosowaniem wybranych algorytmów eksploracji danych

Wykorzystane algorytmy klasyfikujące przypisują klasę nowym przypadkom na podstawie wzorów wygenerowanych podczas nauki na przypadkach o z góry określonej klasie. Zastosowano tu algorytmy trzech różnych typów – oparty na równaniu regresji *Simple Linear Regression*, algorytmy regułowe *Zero Rule* i *One Rule*, drzewo decyzyjne *J48*, oparty na prawdopodobieństwie Naiwny Klasyfikator Bayesowski (*Naive Bayes*) i porównujący przypadek klasyfikowany z najbardziej podobnymi *k-Nearest Neighbour*. Klasyfikację przeprowadzono przy pomocy programu do eksploracji danych (*data mining*) Weka autorstwa Marka Halla, Eibe Frank, Geoffrey'a Holmesa, Bernharda Pfahringera, Petera Reutemanna i Iana H. Wittena[6].

Algorytmy klasyfikujące opisuje się nie w terminach zmiennych zależnych i niezależnych, ale atrybutów. Atrybut decyzyjny zawiera informacje o klasie przypadków, resztę atrybutów opisujących przypadki można nazywać cechami, stanowią one zestaw danych charakteryzujących instancje (Witten i Frank, 2005). Aby trzymać się przyjętej w dziedzinie eksploracji danych w rozdziale tym nie będą stosowane terminy zmienna zależna i niezależna w odniesieniu do wyników klasyfikacji.

Wszystkie algorytmy konstruujące model – reprezentacje regularności znalezionych w danych – przechodzą dwie fazy działania:

1. uczenia na zbiorze danych uczących (przypadków już poklasyfikowanych), w trakcie którego tworzony jest model;
2. testowania na zbiorze danych testujących, również już poklasyfikowanych (algorytm „nie widzi" informacji o ich klasie), które pozwala ocenić, na ile trafny jest wygenerowany model.

Trafność modelu (jego poprawność w klasyfikowaniu nowych przypadków) oszacować można na podstawie wielu wskaźników, jednak najczęściej stosowanymi i najłatwiejszymi do interpretacji są:

- współczynnik poprawności (*accuracy*, procent poprawnych poklasyfikowań) (Witten i Frank, 2005);
- pole pod krzywą ROC (*roc area*, AUC – stosunek pomiędzy prawdopodobieństwem fałszywego alarmu i poprawnej klasyfikacji), które musi

[6]Program i dokumentacja dostępne są na stronie Uniwersytetu Waikato http://www.cs.waikato.ac.nz/ml/weka/, podczas badania korzystano z wersji 3.6.4 oprogramowania

być większe niż 0,5 (poziom klasyfikacji losowej) (Witten i Frank, 2005). Wskaźnik ten stosowany jest jako miara dodatkowa.

Model o wysokich współczynnikach poprawności i AUC może zostać wykorzystany do klasyfikacji nowych przypadków lub zinterpretowany w celu uzyskania wiedzy o zależnościach między atrybutami. Z drugiej strony skomplikowane modele o bardzo wysokiej poprawności dla zbioru uczącego i niskiej dla testującego świadczą o przeuczeniu – zbytnim dopasowaniu modelu do danych (Witten i Frank, 2005).

Algorytmy regułowe Algorytmy regułowe dzielą przypadki na podzbiory według wartości atrybutów niedecyzyjnych. Następnie łączą je w zestawy współwystępujących wartości i zestawiają je z klasami decyzyjnymi. Model generowany przez algorytmy tego typu stanowi zestaw reguł warunkowych („*jeśli* $atrybut_k = x_k$ i $atrybut_m = x_m$ *i ... i* $atrybut_n = x_n$, *to* $atrybut_{decyzyjny} = x_{decyzyjny}$"). Reguły te muszą spełniać warunki minimalnej liczby przypadków i trafności. Zastosowany tu *One Rule* wybiera tylko jedną regułę, zaś *Zero Rule* konstruuje regułę przypisującą wszystkim przypadkom najliczniejszą klasę decyzyjną.

Drzewa decyzyjne Drzewa decyzyjne działają w sposób podobny do algorytmów regułowych, ale konstruują hierarchiczne zestawy reguł (reprezentowane jako struktury drzewiaste), pozwalające podzielić przypadki na podzbiory jak najbardziej jednolite pod względem wartości atrybutu decyzyjnego. Model konstruowany jest poprzez wielokrotny podział (w pierwszym kroku wszystkich przypadków, w następnych – otrzymanych wcześniej podzbiorów) pod względem wartości najlepiej porządkującego dany podzbiór atrybutu. Obydwa rodzaje algorytmów kodują uzyskane modele w sposób łatwy do interpretacji i zastosowania, przy czym model może nie brać pod uwagę wszystkich atrybutów opisujących dane.

Naiwny Klasyfikator Bayesowski Naiwny Klasyfikator Bayesowski klasyfikuje przypadki ze względu na łączne prawdopodobieństwo *a priori* wystąpienia w danej klasie posiadanych przez nie cech – wartości atrybutów niedecyzyjnych. Oparty jest na twierdzeniu Bayesa o prawdopodobieństwie warunkowym zdarzeń. Naiwny Klasyfikator Bayesowski posiada (uznawane za nierealistyczne) założenie o niezależności wartości atrybutów niedecyzyjnych, jest mimo to jednym z najczęściej stosowanych i posiadających najwyższą trafność algorytmów. Model otrzymywany przez algorytm stanowi lista prawdopodobieństw *a priori* wartości atrybutów opisujących przypadki

w każdej klasie i jest nieco trudniejszy do interpretacji. Zastosowanie tego algorytmu pozwala jednak także na stawianie hipotez na temat charakterystyki zebranych danych, gdyż osiąga maksima poprawności jedynie, gdy atrybuty są od siebie zupełnie niezależne, lub determinują się nawzajem (Rish, Hellerstein i Thathachar, 2001; Rish, 2001). W innych przypadkach może przeszacowywać informacje niesione przez atrybuty determinowane przez inne (Rish, 2001).

k-Nearest Neighbour Algorytmy klasyfikujące przypadek na podstawie atrybutu decyzyjnego najbardziej podobnych przypadków również mogą mieć wysokie wskaźniki poprawności, jednak są dużo trudniejsze do interpretacji. Biorą pod uwagę wszystkie atrybuty opisujące dane, przypisując im taką samą wagę w ocenie podobieństwa i nie konstruują modelu mogącego służyć do ponownej klasyfikacji. Gdy klasyfikator taki ma najwyższą poprawność ze wszystkich zastosowanych można interpretować to jako wskazówkę istnienia zależności między atrybutami niedecyzyjnymi, a decyzyjnym, które jednak nie mogą być ujęte w zadowalające zależności. Może to być sugestia, że zebrane dane nie oddają charakterystyk klasyfikacji i należy zebrać ich więcej lub zmienić zestaw analizowanych atrybutów. Gdy inne klasyfikatory konstruują zadowalające modele, *k-Nearest Neighbour* może zostać pominięty w analizie klasyfikacji. Taki sposób klasyfikacji przydaje się jednak przy zbiorach danych o wielu atrybutach, z których wszystkie są istotne dla opisu klasyfikacji.

Do przeprowadzenia opisanych poniżej klasyfikacji zastosowano następujące algorytmy:

1. *ZeroRule* (0R) – stosowany do określania minimalnego poziomu poprawności klasyfikacji (Witten i Frank, 2005, s. 88),
2. *OneRule* (1R),
3. drzewo decyzyjne J48 (nazywane też C 4.5), najbardziej klasyczne z drzew decyzyjnych przyjmujących atrybuty nominalne i liczbowe,
4. Naiwny Klasyfikator Bayesowski (NKB),
5. algorytm porównujący każdy przypadek do 3 najbliższych (najbardziej podobnych) IB3 (*3-Nearest Neighbour, 3-NN*).

Aby uniknąć przeszacowania poprawności klasyfikacji, zastosowano 10-krotną kros-walidację (krzyżowe sprawdzanie poprawności) jako metodę testowania zbieżności wyników klasyfikacji z wartościami oczekiwanymi. Jest to metoda stosowana dla zbiorów danych z małą liczbą przypadków, polega na 10-krotnym podziale zbioru danych na 10 podzbiorów, z których (w każdym kroku uczenia/testowania) 9 tworzy zbiór uczący, a 1 – testujący (Witten i

Tablica 8: *Atrybuty o najwyższym współczynniku korelacji z atrybutami decyzyjnymi suma operatorów oraz liczba operatorów globalnych, lokalnych i kataforycznych w tekście. Analiza przy pomocy algorytmu Simple Linear Regression.*

atrybut decyzyjny	atrybut niedecyzyjny	współczynnik korelacji	błąd standardowy
suma operatorów	liczba endofor	0,995*	6,351
liczba operatorów globalnych	liczna operatorów lokalizacyjnych	0,8766*	2,387
liczna operatorów lokalnych	liczba operatorów retorycznych	0,9299*	2,413
liczba operatorów kataforycznych	liczba operatorów lokalnych	0,9998	2,225

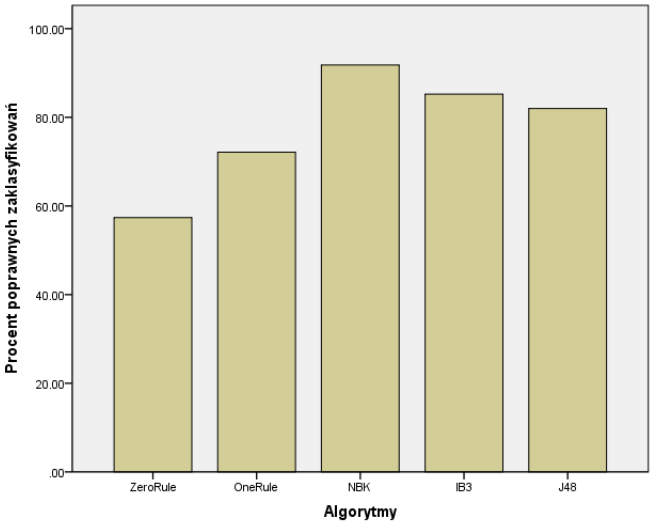
* korelacje istotne statystycznie w regresji liniowej

Frank, 2005, s. 149 – 151). Dzięki temu możliwe jest umieszczenie każdego przypadku w zbiorze testującym. Wskaźniki poprawności klasyfikacji liczone są jako średnia wskaźników z 10 przebiegów działania algorytmu (Witten i Frank, 2005, s. 150).

3.4.1 Regresja liniowa algorytmem Simple Linear Regression

Aby uzupełnić analizę regresji opisaną wcześniej skorzystano z algorytmu regresji liniowej dostępnego w oprogramowaniu Weka. Algorytm ten (*Simple Linear Regression*) wskazuje jedną zmienną niezależną o najwyższym wskaźniku korelacji ze zmienną zależną, przy czym jest on liczony nieco inaczej, niż r Pearsona – jako średni procent zbieżności z wartością oczekiwaną. Podsumowanie wyników znajduje się w tabeli 8. Większość znalezionych korelacji pokrywa się ze znalezionymi w regresji liniowej, jedynie dla argumentu decyzyjnego liczba katafor algorytm wskazał bardzo wysoką korelację która nie ma istotności statystycznej z w regresji liniowej. Blisko 100% poprawność klasyfikacji, jednak przy dość wysokim średnim błędzie może sugerować, że błędy w klasyfikacji wiązały się z dużą rozbieżnością dla nielicznych przypadków. Rozbieżność wielkości współczynników korelacji w regresji liniowej i omawianym tu algorytmie wynika z różnic jego obliczaniu oraz, przypuszczalnie, wielkości odchylenia od wartości oczekiwanej traktowanej jako brak błędu.

Wykres 1: *Poprawność klasyfikacji dla atrybutu decyzyjnego suma operatorów (klasyfikacja tekstów)*

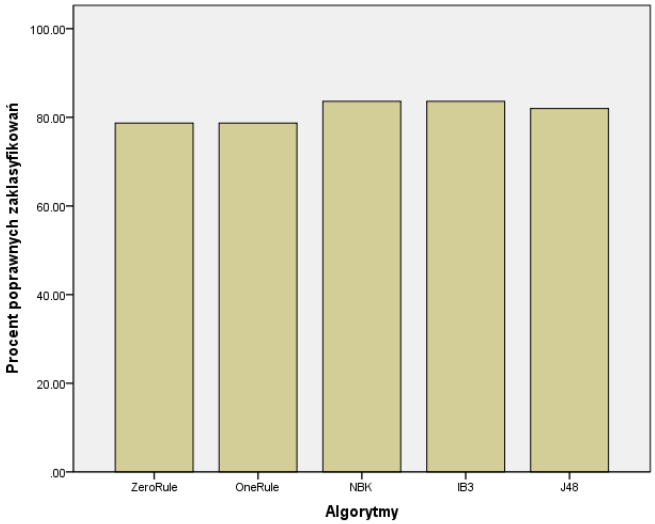


3.4.2 Klasyfikacja tekstów na podstawie operatorów

Przeprowadzono klasyfikacje tekstów na podstawie danych liczbowych o typach zawartych w nich operatorów. Atrybutami decyzyjnymi były zmienne grupujące opisane wyżej, przeprowadzono osobną analizę dla każdej zmiennej decyzyjnej.

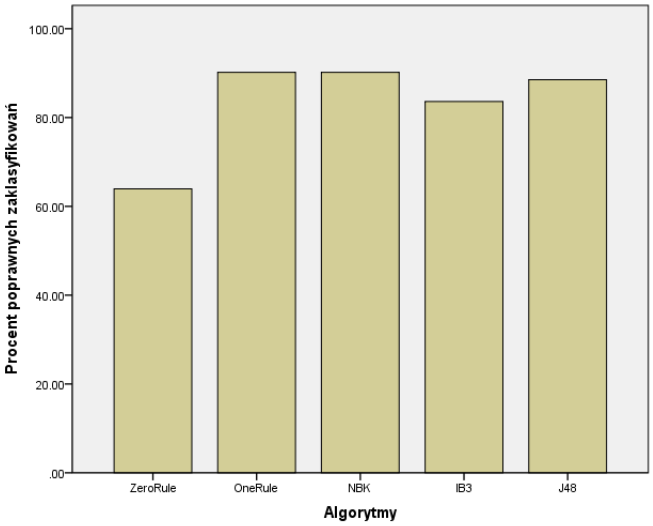
Dla wszystkich zmiennych decyzyjnych uzyskano dość wysokie miary poprawności klasyfikacji (także na poziomie podstawowym - 0R). Wyniki klasyfikacji tekstów dla wszystkich atrybutów decyzyjnych przedstawione są na wykresach 1 – 4. Z danych usuwano atrybuty, które determinowały atrybut decyzyjny (w tym przypadku np. liczba operatorów lokalnych i globalnych).Pod względem sumy operatorów teksty najlepiej przyporządkował do klas algorytm NKB – z poprawnością na poziomie 91,8%. AUC w przypadku tej klasyfikacji wynosi 0,985, jest więc w pełni zadowalająca (klasyfikacja w niemal 100% prawdopodobieństwo trafienia). Stosunkowo wysoki współczynnik poprawności J48 wskazuje, że atrybuty nie są od siebie niezależne (ale nie do końca dobrze opisują atrybut decyzyjny, o czym świadczy także wynik IB3), wysoki wynik NKB może sugerować, że przynajmniej niektóre wartości atrybutów determinowane są przez inne (Rish, 2001). Ze względu na atrybut decyzyjny silnie związany z innymi, zestaw danych był dużo mniejszy, niż w przypadku pozostałych klasyfikacji. Dane zawierały

Wykres 2: *Poprawność klasyfikacji dla atrybutu decyzyjnego liczba katafor (klasyfikacja tekstów)*

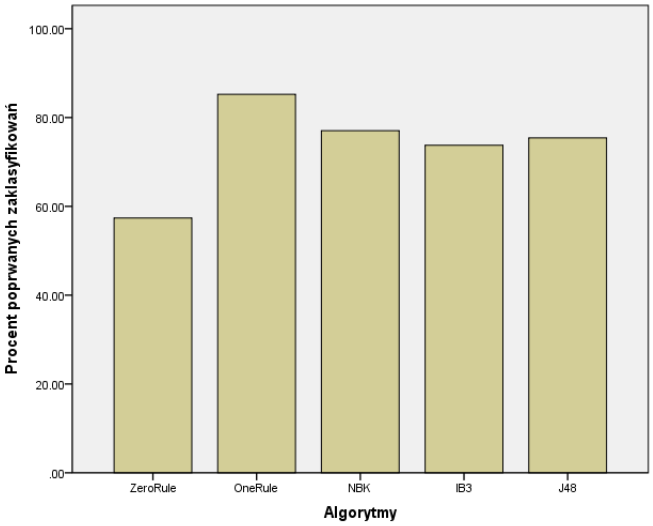


więcej atrybutów, niż poprzednio. do klas liczba katafor najlepiej przyporządkowały NKB i IB3 – 83,6% poprawności. Poziom poprawności jest tu niższy, niż w przypadku sumy operatorów. Opisujący klasyfikację zestaw danych nie jest wystarczający do dokładnej klasyfikacji, choć wartość klasy nie jest niezależna od innych atrybutów. Poszczególne atrybuty ani nie są niezależne, ani nie determinują się. Warto wspomnieć, że dość wysoki poziom poprawności osiągnął 1R (78,69%) posługując się regułą opartą na liczbie operatorów wskazówek synonimiczności w tekście (przy czym r^2 dla korelacji liczby katafor i wskazówek synonimiczności wynosi 0,261). W przypadku klas decyzyjnych opartych na liczbie operatorów lokalnych najlepszą klasyfikację uzyskały NKB i 1R – 90,16%. Algorytm regułowy oparł przyporządkowanie na liczbie operatorów retorycznych (powyżej/poniżej 10). Wyniki te pokrywają się w wynikami regresji liniowej wskazując na powiązanie pomiędzy klasami opartymi na zasięgu i formie operatora. Niższy wynik J48, który skonstruował 3-poziomowe drzewo może świadczyć o tym, że przy większym skomplikowaniu modelu następuje (nieznaczne) przeuczenie. Klasyfikacja ze względu na liczbę operatorów globalnych miała maksymalnie 85,25% poprawności, przy czym osiągnął ją algorytm 1R znajdując regułę oparta na liczbie operatorów lokalizacyjnych (powyżej albo poniżej 6,5 operatorów). Wynik J48 z 4-poziomowym drzewem wskazuje, że przy skomplikowaniu modelu nie przyrasta poprawność, ale następuje przeuczenie.

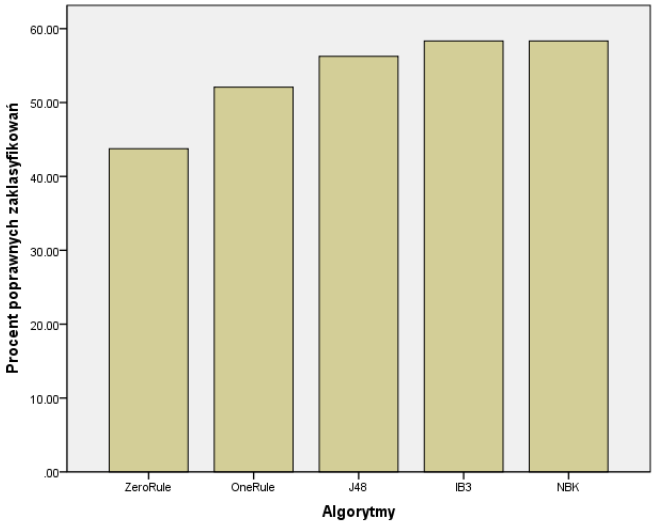
Wykres 3: *Poprawność klasyfikacji dla atrybutu decyzyjnego liczba operatorów lokalnych (klasyfikacja tekstów)*



Wykres 4: *Poprawność klasyfikacji dla atrybutu decyzyjnego liczba operatorów globalnych (klasyfikacja tekstów)*



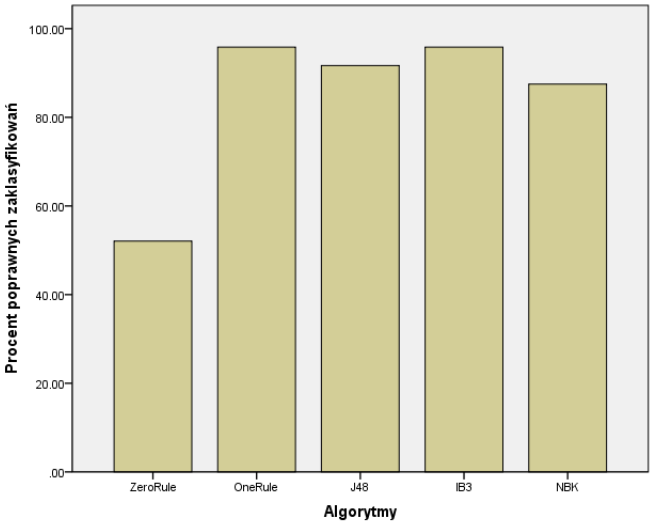
Wykres 5: *Poprawność klasyfikacji dla atrybutu decyzyjnego klasa operatorów tekstowych (klasyfikacja typów operatorów)*



3.4.3 Klasyfikacja typów operatorów – związki pomiędzy klasami operatorów

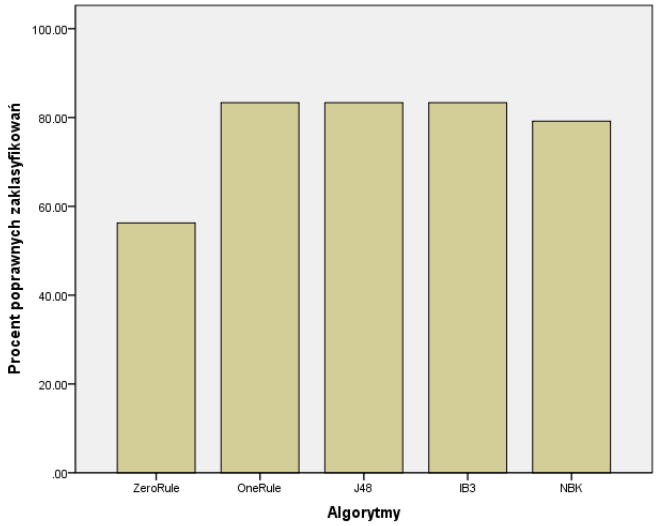
Przeprowadzono także klasyfikację operatorów wg danych zamieszczonych w Tabeli 2. Jako argumenty decyzyjne posłużyły przynależność do klas w analizowanych tu klasyfikacjach oraz występowanie w tekście jako katafora. Wyniki klasyfikacji przedstawiają wykresy 5 – 8. Klasyfikacja do odpowiednich klas operatorów tekstowych okazała się bardzo mało efektywna przy opracowanym zestawie danych. Najlepszą klasyfikacje uzyskano algorytmami IB3 i NKB – 58,3% poprawności. Algorytm 1R (52,08%) skonstruował regułę wskazującą na zawieranie się klasy operatorów endoforycznych w klasie operatorów lokalizacyjnych, zaś wskazówek wyliczenia – w klasie operatorów retorycznych. Reguła ta jest poprawna dla tych klas, ale nie uwzględnia pozostałych – stąd niska poprawność opartej na niej klasyfikacji. Wyniki sugerują, że pozostałe wartości atrybutu decyzyjnego nie wchodzą w regularne zależności z innymi cechami operatorów. Najlepszą klasyfikację operatorów lokalizacyjnych i retorycznych wskazały algorytmy 1R i NKB – 95,8% poprawności. Reguła znaleziona przez pierwszy z nich opierała się na przynależności do klas operatorów tekstowych (jeśli operator należy do endofor należy go zaklasyfikować do klasy lokalizacyjnych, w innych przypadkach – retorycznych), jedynie 2 operatory nie podlegały tej regule. Reguła ta nie

Wykres 6: *Poprawność klasyfikacji dla atrybutu decyzyjnego operator lokalizacyjny lub retoryczny (klasyfikacja typów operatorów)*

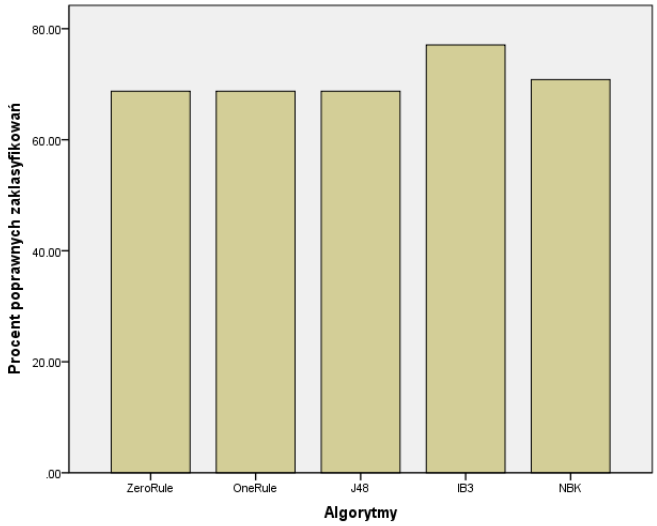


przynosi dodatkowych informacji w stosunku do tych wynikających z definicji klas. Gdy atrybutem decyzyjnym było występowanie w funkcji kataforycznej, najlepszym klasyfikatorem okazał się IB3 (77,08% poprawności), co przy porównywalnych wynikach klasyfikacji pozostałymi metodami sugeruje, że atrybuty niedecyzyjne są powiązane z decyzyjnym, jednak najprawdopodobniej zebrano w tym przypadku za mało danych, aby określić reguły powiązań. 1R uzyskał wysoką poprawność (68,75%), wskazał regułę prawidłową dla 41 z typów 48 operatorów opartą na liczbie ich wystąpień z zasięgiem globalnym. Niski wynik NKB może świadczyć o zbyt małej ilości danych, lub wskazywać na to, że atrybuty niedecyzyjne nie są w pełni niezależne. Dla argumentu decyzyjnego zasięg operatora największą poprawność uzyskano algorytmami 1R, IB3 i NKB (83,3%). Reguła znalezione przez 1R zaliczała do operatorów lokalnych retoryczne, zaś globalnych – lokalizacyjne (8 operatorów nie podlegało tej regule). Jest to zbieżne z poprzednimi analizami wskazującymi na związki pomiędzy zasięgiem i formą operatora.

Wykres 7: *Poprawność klasyfikacji dla atrybutu decyzyjnego występowanie w funkcji kataforycznej (klasyfikacja typów operatorów)*



Wykres 8: *Poprawność klasyfikacji dla atrybutu decyzyjnego zasięg operatora (klasyfikacja typów operatorów)*



4 Podsumowanie wyników badań

4.1 Operatory metatekstowe w krótkich tekstach w języku polskim

Spośród 48 znalezionych rodzajów operatorów operatorów najczęściej występującymi w przebadanych tekstach okazały się:

1. „wyżej" (endofora, częściej występująca z zasięgiem lokalnym, niż globalnym),
2. „pierwszy" (marker wyliczenia),
3. „wniosek" (spójnik logiczny),
4. „artykuł" (endofora o zasięgu globalnym),
5. „między innymi" (marker wyliczenia),
6. „po drugie" (jw.),
7. „po pierwsze" (jw.),
8. „zagadnienie" (marker topiku).

Wszystkie wyżej wymienione rodzaje operatorów mogły powtarzać się w pojedynczym tekście kilka razy (2 – 9).

Sześć ze znalezionych rodzajów operatorów wystąpiło tylko w jednym tekście. Wzięto je pod uwagę w analizie jako przykłady rzadko występującego metatekstu. Są to: „podsumowane", „poświęcony", „powiedziane", „vide", „wykonany", „znaczy to". Wyrażenia tego typu nie wydają się być charakterystyczne tylko dla pojedynczych autorów, można je znaleźć w innych tekstach, są jedynie wyjątkowo rzadkie w przebadanej próbie, wobec tego być może także w populacji krótkich tekstów pokonferencyjnych w języku polskim.

Funkcję kataforyczną miało jedynie średnio 10% operatorów w tekście. Najczęściej miały ją:

1. „artykuł",
2. „praca",
3. „część".

Wszystkie wyżej wymienione operatory są endoforami o zasięgu globalnym. Inne operatory przyjmujące funkcję kataforyczną (12 rodzajów) należały do klas:

- endofora (5),
- marker wyliczenia (3),
- marker aktu dyskursowego (1),
- spójnik logiczny (1),

- marker synonimiczności (1).

33 rodzaje operatorów nie przyjmowały funkcji kataforycznej.

Zasięg globalny miało średnio ok. 47% operatorów, najczęściej operatory lokalizacyjne. Spośród operatorów retorycznych zasięg globalny (w większości przypadków wystąpienia) miały:

1. „na podstawie",
2. „vide" .

Obydwa rodzaje operatorów metatekstowych należą do klasy wskazówek topiku, przy czym „vide" pojawiła się w tekstach tylko raz (w funkcji metatekstowej; „vide" lub „patrz" mogły pojawić się w więcej razy, ale jako wskazówki odniesień do literatury – w funkcji hipertekstowej).

Lokalny zasięg nadano w przebadanych tekstach średnio ok. 58% operatorom (procentowy udział operatorów o różnych zasięgach nie dodaje się do 100%, gdyż policzono średnią ze stosunków liczb operatorów w każdym tekście). Najczęściej miały go operatory retoryczne. Operatorami lokalizacyjnymi występującymi w większości z zasięgiem lokalnym były:

1. „opisany",
2. „poniżej",
3. „wspomniany",
4. „wykonany",
5. „wyżej".

Należy zauważyć, że endofory te nie zawierają odniesienia do konkretnego fragmentu tekstu, a jedynie orientacyjne w stosunku do położenia operatora. Wszystkie wyżej wymienione operatory (oprócz „wykonany", który wystąpił tylko raz) mogły przyjmowały także zasięg globalny.

4.2 Zależności między klasami operatorów

Podsumowanie analiz statystycznych i klasyfikacji poszczególnych klas operatorów metatekstowych pozwala na ocenę relacji pomiędzy tymi klasami. Wnioski te można odnosić tylko do operatorów metatekstowych w krótkich tekstach w języku polskim.

Po pierwsze, jeśli lokalizacyjność i retorycznosc zdefiniować w opisany wcześniej sposób, wyznaczone klasy pokrywają się z podziałem metatekstu tekstowego na endofory i markery wyliczenia w zbiorze „lokalizacyjne", i resztę klas (markery topiku, aktów dyskursowych i spójniki logiczne) w drugim.

Wobec tego, o ile podział taki może być przydatny do oceny stylu argumentacyjnego autora (Dahl, 2004), nie musi być w tym celu przeprowadzana dodatkowa klasyfikacja.

Nieco mniej jednoznaczne rysują się wnioski dotyczące zasięgu operatorów. Klasyfikacja metodami eksploracji danych opierała się, w przypadku reguł lub drzew decyzyjnych, na klasyfikowaniu operatorów lokalizacyjnych jako globalnych, zaś retorycznych – jako lokalnych. Przytoczone na początku tego rozdziały zestawienia dla pojedynczych rodzajów operatorów pokazują jednak, że nie tylko nie jest to reguła stuprocentowo trafna, ale redukcja tych dwóch podziałów do jednego sprawi, że utracona zostanie informacja o tych operatorach, które w pewnych przypadkach przyjmują zasięg niezgodny z tendencją w ich klasie. Interesujące wydają się operatory retoryczne przyjmujące zasięg globalny – dotyczy to kilku typów operatorów, być może w szczególnych sytuacjach komunikacyjnych. Operatory te mogą wymagać większego wysiłku poznawczego ze strony czytelnika, który będzie chciał porównać treść odległych fragmentów tekstu powiązanych relacją retoryczną wskazywaną przez operator (Lemarié et al., 2008). Przyjmowanie zasięgu lokalnego przez operatory lokalizacyjne było rzadsze, jednak też może stanowić istotną cechę niektórych rodzajów operatorów. Dokładna analiza zależności pomiędzy tymi klasami wymaga dokładniejszych badań.

Jeśli skupić się na podziale bardziej drobnoziarnistym – związków pomiędzy klasą tekstową operatora i jego zasięgiem,okaże się, że przytoczona wyżej reguła klasyfikacji nie dotyczy jedynie dwóch rodzajów wskazówek topiku oraz endofor niezawierających odniesienia do konkretnej części tekstu (niebędących etykietami (Lemarié et al., 2008, s. 33)). Wydaje się, że pozwala to na zrezygnowanie z podziału na metatekst retoryczny i lokalizacyjny w odniesieniu do operatorów metatekstowych, ale wskazuje na konieczność dodatkowego wymiaru opisu endofor (np. precyzja odniesienia (Lemarié et al., 2008, s. 38)) oraz wskazówek topiku, w tym celu jednak należałoby zebrać więcej danych o takich operatorach. Z drugiej strony, być może należałoby dodać jeszcze jedną klasę operatorów tekstowych (np. marker odniesienia) – niezawierających etykiety, ale odsyłających czytelnika do już przeczytanych fragmentów (anaforycznych). Należałoby się wówczas zastanowić, czy podział ten nie jest zbyt drobiazgowy, lub na siłę nie porządkuje różnych wymiarów opisu operatorów metatekstowych. na to pytanie, być może, pozwoliłoby badanie nad rozumieniem, kategoryzacją operatorów, lub uzupełnianiem ich w tekście na natywnych użytkownikach języka polskiego, w paradygmacie podobnym do badań Goldman i Murray'a (Goldman i Murray, 1992).

Najbardziej zróżnicowaną i najtrudniejszą do automatycznej klasyfikacji klasą operatorów były katafory. Klasa ta nie została wyróżniona w żadnej z cytowanych klasyfikacji mimo łatwości w zdefiniowaniu jej. Tym bardziej

interesujące wydaje się zestawienie klas operatorów z tendencją do nadawania im w tekstach funkcji kataforycznej – zapowiadania treści i struktury tekstu.

Wszystkie przebadane teksty posiadały abstrakty, prawie każdy ze znalezionych w streszczeniu operatorów był kataforyczny (i globalny). Prawdopodobnie z tej przyczyny analiza danych wskazywała często na związek pomiędzy tymi dwiema klasami. Odpowiedź na to pytanie wymaga jednak kolejnych badań uwzględniających w analizie różnicę pomiędzy operatorami umieszczanymi w zasadniczym tekście i w poprzedzającym go abstrakcie. Najczęściej funkcję tę miały endofory, potem wyliczenia, inne klasy operatorów tekstowych znacznie rzadziej występowały w tej funkcji, nie można jednak identyfikować katafor z metatekstem lokalizacyjnym. Wydaje się, że aby jednoznacznie stwierdzić, jakie relacje wiążą kataforyczność z innymi cechami operatora potrzebne jest dalsze badanie, uwzględniające więcej tekstów oraz kontrolujące położenie operatora w tekście.

5 Wnioski

Opisane tu badanie pozwoliło na wstępne scharakteryzowanie operatorów metatekstowych występujących w krótkich tekstach naukowych w języku polskim. Wykorzystano w nim klasyfikację metadyskursu Hylanda i Mur-Dueñas (Hyland, 1998; Hyland i Tse, 2004; Hyland, 2005; Mur-Dueñas, 2011, 2009) i oparte na niej klasyfikacje Dahl (2004) i Buntona (1999), opracowane dla języka angielskiego i hiszpańskiego, po raz pierwszy stosując je do tekstów w języku polskim. Analiza statystyczna wyników badania i eksploracja danych algorytmami klasyfikacyjnymi pozwoliła na przeanalizowanie relacji pomiędzy proponowanymi przez Hylanda , Mur-Dueñas, Dahl i Buntona klasami operatorów. Otrzymana w badaniu charakterystyka operatorów metatekstowych różnych klas pozwala na wyekstrahowanie ich jako określonej klasy obiektów metatekstowych z całości metatekstu (metadyskursu) na podstawie słów kluczowych oraz położenia i roli poszczególnych operatorów w tekstach. Przeprowadzenie klasyfikacji operatorów metatekstowych w tekstach określonego typu i analizy zależności między znalezionymi klasami operatorów może stanowić punkt wyjścia do dalszych badań zarówno nad częstością występowania metatekstu określonych typów, jak i jego wpływem na przetwarzanie tekstu.

Badanie Graessera i in. (Graesser, Jeon, Yan i Cai, 2007) wykazało, że teksty konstruowane na potrzeby badań nad czytaniem zawierają znacznie większą proporcję metatekstu, niż można znaleźć w tekstach, na przetwarzanie których rozciąga się wnioski z badań. Wyniki przedstawionych tu analiz wskazują wstępnie, jak można manipulować zawartością operatorów w tek-

ście, aby badania miały większą trafność zewnętrzną. Wśród przebadanych tekstów znaleziono jeden, który nie zawierał operatorów metatekstowych (nie znaczy to, że nie zawierał metatekstu w ogóle), zaś część tekstów zawierała niewielką liczbę operatorów. Można powiedzieć, że potwierdza to zewnętrzną trafność stosowania w badaniach jako próby kontrolnej tekstów nie zawierających elementów metatekstowych (McNamara et al., 1996; McNamara, 2001; T. Sanders, Land i Mulder, 2007b), przy czym wniosek ten można odnieść jedynie do operatorów metatekstowych, a nie elementów metatekstowych (metedyskursowych) w ogóle.

Zwiększenie kontroli nad tym, jakie elementy metatekstowe poddawane są manipulacji przez ograniczenie jej na przykład do operatorów metatekstowych zdefiniowanych jak w opisanym badaniu może pozwolić na systematyczne badanie wpływu spójności i wskazówek relacji retorycznych w tekście, odpowiadające w sposób kontrolowany w większym niż do tej pory stopniu (McNamara et al., 1996; McNamara, 2001; T. J. M. Sanders i Noordman, 2000) na pytanie, czy metatekst ułatwia czytelnikom rozumienie tekstu (Lemarié et al., 2008). Metodologia zastosowana w opisanym badaniu pozwalają na wyodrębnienie z całości metatekstu operatorów metatekstowych i skupienie się w przyszłych badaniach jedynie na nich, ale także zastosowanie wyników klasyfikacji operatorów podczas adaptacji tekstów do badań i analizę wpływu na proces czytania jedynie wybranych klas operatorów.

Wyniki badania mogą zostać wykorzystane w analizie porównawczej sposobu korzystania z operatorów metatekstowych w tekstach pisanych po polsku przez przedstawicieli innych dyscyplin naukowych, lub przedstawicieli nauk poznawczych, ale piszących w innych językach, na wzór badań Mur-Dueñas (2009, 2011) i licznych prac Hylanda (na przykład 2005, gdzie zebrał wiele badań tego typu). Analiza częstości występowania w tekstach metatekstu różnych typów stanowi także podstawę stawiania hipotez co do celu stosowania środków metatekstowych przez autorów przebadanych prac, na wzór badań Hylanda (2005, 1998) i innych badaczy (Mur-Dueñas, 2011, 2009; Saz Rubio, 2011; Abdi et al., 2010).

W szerszej perspektywie wyniki opisanego badania i kolejnych, w takim paradygmacie badawczym, mogą także stanowić podstawę konstrukcji algorytmów automatycznego przetwarzania języka naturalnego na potrzeby generowania streszczeń oraz wydobywania struktury retorycznej tekstu, jak w pracach Marcu, Knotta i Dale'a (Marcu, 1997; Knott i Dale, 1993, 1996). Oparte między innymi na analizie metatekstu metody automatycznego odczytywania struktury topiku i subtopików tekstu – związków pojawiających się w tekście idei z jego głównym tematem (Zwaan i Radvansky, 1998) – mogą ponadto ułatwić wyszukiwanie tekstów na dany temat (jako uzupełnienie przeszukiwania baz za pomocą słów kluczowych). Mogłyby także sta-

nowić podstawę konstrukcji algorytmów wspomagających uczenie się z tekstu poprzez wyróżnianie najbardziej kluczowych fragmentów (na przykład tych, z którymi wiąże się wiele dalszych) oraz monitorowanie procesu zapamiętywania i rozumienia tekstu w oparciu o wydobytą strukturę.

Literatura

Abdi, R., Rizi, M. T. i Tavakoli, M. (2010). The cooperative principle in discourse communities and genres: A framework for the use of metadiscourse. *Journal of Pragmatics*, *42*.

Aijmer, K. i Simon-Vandenbergen, A.-M. (2004). A model and a methodology for the study of pragmatic markers: the semantic field of expectation. *Journal of Pragmatics*, *36*(10), 1781-1805.

Bunton, D. (1999). The use of higher level metatext in ph.d theses. *English for Specific Purposes*, *18*(1), 41-56.

Dahl, T. (2004). Textual metadiscourse in research articles: a marker of national culture or of academic discipline? *Journal of Pragmatics*, *36*(10), 1807-1825.

Degand, L. i Sanders, T. (2002). The impact of relational markers on expository text comprehension in l1 and l2. *Reading and Writing*, *15*(7).

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, *31*(7), 931-952.

Goldman, S. R. i John A. Rakestraw, J. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson i R. Barr (Eds.), *Handbook of reading research. vol. 3* (p. 311-336). Lawrence Erlbaum Associates.

Goldman, S. R. i Murray, J. D. (1992). Knowledge of connectors as cohesion devices in text: A comparative study of native-english and english-as-a-second-language speakers. *Journal of Educational Psychology*, *84*(4), 504-519.

Graesser, A. C., Jeon, M., Yan, Y. i Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journa*, *15*(3), 199-213.

Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics*, *30*(4), 437-455.

Hyland, K. (2005). *Metadiscourse: exploring interaction in writing.* wyd. Continuum.

Hyland, K. i Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistic*, *25*(2), 156-177.

Knott, A. i Dale, R. (1993). *Using linguistic phenomena to motivate a set of rhetorical relations* (Tech. Rep.). Discourse Processes.

Knott, A. i Dale, R. (1996). Choosing a set of coherence realtions for text generation: a data-driven approach. *Lecture Notes in Computer Science*, *1036*(1036), 47-67.

Lemarié, J., Robert F. Lorch, J., Eyrolle, H. i Virbel, J. (2008). Sara: A text-based and reader-based theory of signaling. *Educational Psychologist*, *43*(1), 27-48.

Louwerse, M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistic*, *12*(3), 291-315.

Louwerse, M. M. i Mitchell, H. H. (2003). Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse Processes*, *35*(3), 199-239.

Marcu, D. (1997). *The rhetorical parsing of natural language texts.* Association for Computational Linguistics.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, *55*(1), 51-62.

McNamara, D. S., Kintsch, E., Butler Songer, N. i Kintsch, W. (1996). Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1-43.

Mur-Dueñas, P. (2009). Logical markers in l1 (spanish and english) and l2 (english) business research articles. *English Text Construction*, *2*(2), 246-264.

Mur-Dueñas, P. (2011). An intercultural analysis of metadiscourse features in research articles written in english and in spanish. *Journal of Pragmatics*, *43*(12), 3068-3079.

Rish, I. (2001). An empirical study of the naive bayes classifier. In *Ijcai 2001 workshop on empirical methods in artificial intelligence.*

Rish, I., Hellerstein, J. i Thathachar, J. (2001). *An analysis of data characteristics that affect naive bayes performance* (Tech. Rep.). Technical Report RC21993, IBM T.J. Watson Research Center.

Sanders, T., Land, J. i Mulder, G. (2007a). Linguistic markers of coherence improve text comprehension in functional contexts. *Information Design Journal*, *15*(3), 219-235.

Sanders, T., Land, J. i Mulder, G. (2007b). Linguistic markers of coherence improve text comprehension in functional contexts. *Information Design Journal*, *15*(3), 219-235.

Sanders, T. J. M. i Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, *29*(1), 37-60.

Saz Rubio, M. M. del. (2011). A pragmatic approach to the macro-structure

and metadiscoursal features of research article introductions in the field of agricultural sciences. *English for Specific Purposes*, *30*(4), 258-271.

Winiarska, J. (2001). *Operatory metatekstowe w dialogu telewizyjnym*. Wyd. Universitas.

Witten, I. H. i Frank, E. (2005). *Data mining: Practical machine learning tools and techniques, second edition*. wyd. Morgan Kaufmann.

Zwaan, R. A. i Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162-185.