**Elżbieta Hajnicz**

# Eksperymenty z zakresu klasyfikacji czasowników w semantycznym słowniku walencyjnym polskiego

**Nr 1021**

Warszawa, maj 2011

**Streszczenie**

Niniejszy raport opisuje wstępne eksperymenty dotyczące klasyfikacji syntaktyczno-semantycznej czasowników polskich. Wpierw omówione zostały istniejące prace z tej dziedziny, dotyczące głównie języka angielskiego. Następnie opisana została gradacyjna analiza odpowiedniości i skupień, która została użyta do klasyfikacji. Potem przedstawiony został semantyczno-syntaktyczny słownik walencyjny będący źródłem danych do klasyfikacji. Na koniec zaprezentowane były właściwe eksperymenty dotyczące klasyfikacji wraz z ich ewaluacją.

**Słowa kluczowe**: lingwistyka komputerowa, semantyka leksykalna, słowniki walencyjne, klasyfikacja czasowników, preferencje selekcyjne, wordnet

**Abstract**

**Experiments on classifying verbs in a semantic dictionary of Polish**

The present report describes initial experiments on syntactic-semantic classification of Polish verbs. First, the existing works on this subject were discussed, mainly concerning English. Second, Grade Correspondence-Cluster Analysis used in experiments was described. Next, syntactic-semantic valence dictionary of Polish verbs being a source of data for experiments was presented. Finally, actual experiments were discussed and evaluated.

**Keywords**: computational linguistics, lexical semantics, electronic valence dictionaries, verb classification, selectional preferences, wordnet

# 1 Wstęp

Klasyfikacja czasowników budzi duże zainteresowanie wśród badaczy. Klasy czasowników są istotne nie tylko ze względu na uwypuklenie własności czasowników podczas ich formalnego badania, lecz także z przyczyn praktycznych, gdyż wspólny opis ułatwia operowanie czasownikami rzadszymi, zmniejszając negatywny wpływ rozproszenia danych. Klasyfikacja czasowników bywa tworzona ręcznie (Levin, 1993; Baker *et al.*, 1998; Fillmore *et al.*, 2003). Ostatnimi czasy jednak coraz więcej prac poświęcanych jest metod automatycznej klasyfikacji. Metody takie mogą być wykorzystywane jako element w rozlicznych działach przetwarzania języka naturalnego, np. znakowania rolami semantycznymi (Swier i Stevenson, 2004; Zapirain *et al.*, 2008), ujednoznacznia znaczenia wyrazów (Dang, 2004), parsingu (Shi i Mihalcea, 2005), i in.

Klasyfikacja czasowników ze względu na podobieństwo ich charakterystyki syntaktyczno-semantycznej jest jednak interesującym tematem sama w sobie. Może też zostać uznana za ostatni etap tworzenia semantycznego słownika walencyjnego, i taką funkcję pełni również w ciągu eksperymentów, których dotyczy niniejszy raport. W omówionych tu eksperymentach każdy schemat został potraktowany osobno, czyli charakterystyka syntaktyczna była ustalona.

# 2 Przegląd metod klasyfikacji czasowników

## 2.1 Podejście oparte na zależnościach syntaktycznych

W dwóch kolejnych artykułach (Stevenson i Merlo, 1999; Merlo i Stevenson, 2001) zaproponowana została metoda klasyfikacji czasowników będąca jednocześnie metodą wykrywania alternacji w oparciu o pochodzącą od Levin (1993) ideę polegającą na tym, że czasowniki należące do tej samej klasy podlegają jednocześnie tym samym alternacjom i *vice versa*.

Autorki rozważają trzy klasy czasowników zgodnych z klasyfikacją Levin (1993):

- nieergatywne (ang. *unergative*)

  (1) a. *The horse raced past the barn.*
  b. *The jockey raced the horse past the barn.*

- nieakuzatywne (ang. *unacusative*)

  (2) a. *The butter melted in the pan.*
  b. *The cook melted the butter in the pan.*

- zaniku dopełnienia (ang. *object-drop*)

(3) a. *The boy washed the hall.*
b. *The boy washed.*

Pierwsze dwie klasy podlegają alternacji kauzatywnej, zaś trzecia — alternacji zaniku dopełnienia. Tak więc z punktu widzenia wykrywania alternacji rozróżnienie pomiędzy pierwszymi dwoma jest nieistotne i rozróżniające je cechy mogą zostać zignorowane.

Obie wspomniane alternacje objawiają się istnieniem przechodniego i nieprzechodniego schematu walencyjnego, w których jednak argumenty pełnią różne role semantyczne. Podstawowa różnica pomiędzy czasownikami ze wszystkich trzech rozważanych klas polega na tym, że w pierwszych dwóch dopełnienie realizacji przechodniej staje się podmiotem realizacji nieprzechodniej (co stanowi o alternacji kauzatywnej), zaś w trzeciej dopełnienie realizacji przechodniej po prostu znika (alternacja zaniku dopełnienia).

Różnica pomiędzy klasami nieergatywną i nieakuzatywną kryje się w rolach semantycznych pełnionych przez dopełnienie w realizacji przechodniej. W wypadku pierwszej jest to kontragens, który staje się agensem realizacji nieprzechodniej (podmiot wolicjonarny); w wypadku drugiej jest to obiekt (ogólniej temat; podmiot niewolicjonarny).

Różnice te powodują odmienną dystrybucję czasowników różnych klas pomiędzy oba wzorce walencyjne oraz odmienną dystrybucję rzeczowników pomiędzy pozycją podmiotu i dopełnienia względem wzorców. W szczególności, czasowniki nieergatywne rzadziej występują w postaci przechodniej (por. Stevenson i Merlo, 1997), dzięki czemu alternacja kauzatywna będzie zachodziła najczęściej dla czasowników nieakuzatywnych, co umożliwi ich wyróżnienie. Dla czasowników gubiących dopełnienie nie powinna zachodzić wcale. Tak więc zachodzenie alternacji kauzatywnej jest istotną cechą służącą do rozróżnienia klas.

Za kolejną istotną cechę autorki uznają dystrybucję pomiędzy czynne i bierne występowanie czasowników. Za szczególnie charakterystyczne uznają wystąpienie bierne w zredukowanej (pozbawionej zaimka względnego) postaci względnej.[1]

Ostatecznie, we wcześniejszej z prac Stevenson i Merlo (1999) rozważają cztery cechy: przechodniość, strona, opozycja czas przeszły prosty — postać względna zredukowana i kauzatywność.

W późniejszej pracy Merlo i Stevenson (2001) zauważają, że w roli agensa bądź kontragensa częściej występują rzeczowniki ożywione niż w roli obiektu

[1] *The horse raced past the barn fell.*; w języku polskim taka konstrukcja nie istnieje.

(tematu). Przyjmując założenie, że czasowniki nieakuzatywne często posiadają realizację nieprzechodnią, dla tych właśnie czasowników ożywiony podmiot będzie rzadszy niż dla czasowników z pozostałych dwóch klas.

Eksperymenty zostały przeprowadzone na wersji korpusu Browna oznakowanej klasami gramatycznymi oraz korpusie WSJ, a połączony korpus zawierał ok. 65 mln. wyrazów. Korpus ten został następnie sparsowany za pomocą parsera statystycznego opracowanego przez Collinsa (1997). Frekwencja cech wyliczana była dla poszczególnych czasowników w następujący sposób:

- INTR (nieprzechodniość)
  Fraza nominalna następująca tuż za czasownikiem uznawana była za potencjalne dopełnienie, a wystąpienie czasownika zliczane było jako przechodnie, wpp. wystąpienie czasownika zliczane było jako nieprzechodnie.[2]
- ACT (strona)
  Czasownik główny oznakowany jako `VBD` (występujący w stronie czynnej) zliczane były jako czynne, podobnie czasowniki oznakowane jako `VBN` poprzedzone wystąpieniem czasownika *have* (`HV`), zaś czasowniki oznakowane przez `VBN` poprzedzone wystąpieniem czasownika *be* (`BE`) zliczane były jako bierne.
- VBD (imiesłów)
  Zliczane są wystąpienia czasowników oznakowane jako `VBD/VBN`. Tak więc cecha ta jest nierozłączna z poprzednią.
- CAUS (kauzatywność)
  Cecha ta jest najbardziej złożona i dlatego wyliczana była w kilku krokach. Wpierw zliczane były frekwencje rzeczowników znajdujących się na pozycji podmiotu w realizacji nieprzechodniej i pozycji dopełnienia w realizacji przechodnich. Następnie dla wszystkich rzeczowników występujących na obu pozycjach wybierana była większa z tych frekwencji. Frekwencje były sumowane po rzeczownikach, a współczynnik liczony był względem frekwencji zsumowanej dla obu pozycji. Tak więc CAUS $= \frac{\sum_{n \in r_1 \cap r_2} \max(\mathfrak{f}_1(n), \mathfrak{f}_2(n))}{\sum_{n \in r_1 \cup r_2} \mathfrak{f}_1(n) + \mathfrak{f}_2(n)}$. McCarthy (2001) zauważa, że proporcja ta jest zaniżana ze względu na fakt sumowania frekwencji, i jej maksymalna wartość dla równych wielozbiorów wynosi $0,5$. Można to zmienić przez zamianę mianownika tej równości na $\sum_{n \in r_1 \cup r_2} \max(\mathfrak{f}_1(n), \mathfrak{f}_2(n))$.

Dodanej w późniejszej pracy (Merlo i Stevenson, 2001) cechy ożywioności autorki nie szacują na podstawie jakiejś zewnętrznej wiedzy typu tezaurusa, lecz opierają na poniższej heurystyce:

[2] Należy więc podejrzewać, że zastosowany parser był powierzchniowy.

- ANIM (ożywioność)
  Na bazie założenia, że wystąpienia zaimków osobowych *I, we, you, she, he* są ożywione,[3] zliczane są w korpusie wszystkie wystąpienia zaimków na pozycji podmiotu dla nieprzechodnich realizacji rozważanych czasowników.

Frekwencje były normalizowane względem liczby wystąpień czasownika.[4]

Do eksperymentów Merlo i Stevenson wybrały po 20 czasowników z każdej z badanych klas, wybranych z klas Levin (1993), dla których wyliczone zostały odpowiednie wektory cech. Zauważmy, że zakres badanych wariantów syntaktycznych znany jest z góry. W szczególności, nie są rozważane czasowniki nie należące do żadnej z tych klas, np. posiadających wyłącznie schemat przechodni lub nieprzechodni.[5]

Punkt odniesienia stanowi losowy dobór klas (jednej z trzech), dla którego trafność (ang. *accuracy*) wynosi 33%.

### 2.1.1 Wcześniejsze eksperymenty

Wcześniejsze eksperymenty, opisane w pracy (Stevenson i Merlo, 1999), przeprowadzone zostały dla wektorów cech postaci:

[verb, VBD, ACT, INTR, CAUS]

oraz ich trzyelementowych podzbiorów (nie licząc czasownika) w celu zbadania, które cechy najlepiej wyznaczają klasy.

**Uczenie bez nadzoru** Pierwsza zastosowana metoda, algorytm grupowania hierarchicznego (ang. *hierarchical clustering algorithm*) należy do metod automatycznego uczenia się bez nadzoru. Najwyższa trafność — 66%, a więc 2-krotnie lepsza od punktu odniesienia, uzyskana została dla wektora cech [VBD, INTR, CAUS], wyniki dla pozostałych wektorów wahają się od 45% do 54%. Pokazuje to, jak bardzo dobór cech może wpłynąć na uzyskiwane wyniki.

**Uczenie z nadzorem** Do uczenia pod nadzorem został użyty algorytm C5.0 generowania drzew decyzyjnych. Ponadto ze zbioru drzew decyzyjnych automatycznie generowany był zbiór reguł. Obie metody stosowane były dla tych

[3] W języku polskim zaimki *on, ona* nie sugerują żywotności. Z kolei *they* rozbijane jest na męskoosobową wskazującą ożywienie formę *oni* oraz niemęskoosobową *one*.

[4] Jak rozumiem, z wyłączeniem CAUS, która sama w sobie stanowi współczynnik ułamkowy.

[5] Takie czasowniki łatwo odróżnić od poprzednich, gdyż wartością cechy INTR jest w ich wypadku 0 lub 1. Nie wiadomo jednak, jak pozostałe cechy wpłynęłyby na ich klasyfikację. Prawdziwy kłopot pojawiłby się jednak dopiero w momencie, gdy zamiast dwóch nowych klas dodalibyśmy jedną — inne.

samych wektorów cech, przy czym dokonywano 10-krotnej walidacji krzyżowej (ang. *cross-validation*). Tym razem najlepsze wyniki uzyskano dla wektora złożonego ze wszystkich 4 cech, dla drzew decyzyjnych trafność wynosiła 64,2% (przy odchyleniu standardowym 1,7%), zaś dla reguł 64,9% (1,6%). Dla pozostałych wektorów cech trafność wahała się od 54,4% do 60,9% dla drzew oraz od 55,7% do 62,3% dla reguł, przy odchyleniu standardowym poniżej 1,5%. Tak więc zazwyczaj ekstrakcja reguł poprawia nieco wyniki, są one jednak nieco gorsze od uzyskanych metodą bez nadzoru.

### 2.1.2 Późniejsze eksperymenty

W eksperymentach z (Merlo i Stevenson, 2001) użyty został rozszerzony wektor cech

[verb, VBD, ACT, INTR, CAUS, ANIM].

Uznawszy, że wszystkie badane metody (autorki prowadziły także eksperymenty dla dwóch rodzajów sieci neuronowych) dają podobne wyniki, Merlo i Stevenson zdecydowały skupić się na wymienionych powyżej metodach uczenia pod nadzorem. Podobnie jak poprzednio, rozważany jest pełen wketor cech oraz wszystkie jego, tym razem czteroelementowe, podzbiory. W artykule opisane zostało kilka różnych eksperymentów wraz z szeroką analizą wyników. Metoda 10-krotnej walidacji krzyżowej dała najlepsze wyniki (69,8%) dla pełnego wektora cech oraz dla wektora bez cechy ACT. Usunięcie tej cechy dało we wcześniejszym eksperymencie poprawę wyników, więc wyraźnie nie ma ona korzystnego wpływu na działanie algorytmu. Dla pozostałych wektorów trafność waha się od 61,6% do 67,3%. Warto zauważyć, że dla rozważanego poprzednio wektora [verb, VBD, ACT, INTR, CAUS] tym razem trafność jest nieco niższa, 63,2%. Różnica wynika z faktu, że eksperymenty były przeprowadzane ponownie, w szczególności tym razem 10-krotna walidacja krzyżowa przeprowadzana była dla 50-ciu różnych podziałów losowych zbioru wektorów. Tak więc wynik ten jest bardziej wiarygodny, w szczególności odchylenie standardowe wynosiło za drugim razem jedynie 0,5–0,6%.

Autorki podkreślają, że wektory cech są o tyle niezależne od siebie, że każdy czasownik występuje jedynie w jednym z nich. Jaki więc by nie był podział danych na treningowe i testowe, czasowniki występujące w jednym są nieobecne w drugim, co jest szczególnie trudnym przypadkiem klasyfikacji.

Merlo *et al.* (2002) pokazują, że ich metoda może zostać zastosowana dla innych języków (włoskiego, chińskiego). Co więcej, pokazują, że podobieństwa i różnice w realizacji argumentów tego samego czasownika w dwóch różnych językach (wydobyte z niezależnych korpusów, nie z korpusu równoległego) mogą usprawnić proces klasyfikacji tego czasownika dla jednego z tych języków.

### 2.1.3 Uogólnienie podejścia

Rozważone powyżej rozwiązania opierały się na precyzyjnym określeniu wektora cech dla znanego z góry, niewielkiego zbioru schematów oraz alternacji zachodzących dla czasowników występujących w tych schematach. Joanis i Stevenson (2003); Joanis *et al.* (2008) pokazali jak uogólnić te koncepcje dla dowolnych schematów i alternacji, na których m.in. oparta jest klasyfikacja czasowników Levin (1993).

Idea polega na zastąpieniu niewielkiego, starannie dobranego wektora cech ogólną przestrzenią cech, gdyż cechy nie mogą być dopasowane do konkretnego zestawu alternacji. Należy rozważyć wszystkie cechy mogące różnicować grupy czasowników (i jednocześnie determinować zachodzenie alternacji) na wszystkich pozycjach syntaktycznych. Pozwala to przy okazji ograniczyć pracę lingwistów niezbędną przy określaniu wektora cech właściwego dla danego zbioru alternacji.

Przestrzeń cech zawierała:

- frekwencję współwystępowania danego wymagania z czasownikiem;[6]
- stopień zachodzenia na siebie poszczególnych wymagań w parach schematów czasownika wskazujących na potencjalną możliwość zachodzenia alternacji (rozważana powyżej cecha CAUS jest przypadkiem takiej zależności);
- ożywioność (ANIM) liczona dla wszystkich wymagań; poza zaimkami zliczane były frazy NP oznaczone w korpusie jako osobowe;
- zgodnie z Levin uznającą za alternację sformułowania rozpoczynające się od pustych semantycznie *It/There* względem sformułowań ze standardowym podmiotem,[7] zliczane są wystąpienia *it* na pozycji podmiotu i dopełnienia bliższego oraz *there* na pozycji podmiotu.

Ponadto zliczane są częstości występowania czasownika ze znacznikami pochodzącymi z tagsetu Penn będącym zmodyfikowaną wersją tagsetu Korpusu Browna. Nie jest jasne, czy zliczenia te dokonywane są łącznie czy też dla każdego schematu oddzielnie.

Łączna liczba cech wynosi 220.

W eksperymencie opisanym w (Joanis i Stevenson, 2003) autorzy rozważali 5 par i dwie trójki klas czasownikowych, z których jedna dotyczyła dwóch klas

[6] Mniej częste za to bliskie znaczeniowo przyimki; np. *between, in between, among, amongs, amid, amids* były rozważane łącznie.

[7] Podany przykład to *A problem developed / There developed a problem* (pol. *Pojawił się problem.*)

podlegających różnym alternacjom, cztery dotyczyły klas, z których tylko jedna zawiera alternujące czasowniki, zaś dwie klas różniących się sposobem rozdziału ról semantycznych między argumenty.[8] Pary te były więc bardzo zróżnicowane.

Sam eksperyment realizowany był w sposób podobny do opisanego powyżej. Dla par klas trafność wahała się od 72,7% do 84,6%, jednak jest to zadanie prostsze od opisanego w powyższych sekcjach (punkt odniesienia 50%). Dla dwóch trójek (punkt odniesienia 33%) osiągnięto trafność 65,9% i 72,1%, co jest wynikiem porównywalnym do uzyskanego dla specjalnie dobranego wektora cech. Wreszcie dla wszystkich 13 klas (jedna z rozważanych klas była podklasą innej) otrzymana trafność 46.4 %, co jest znakomitym wynikiem biorąc pod uwagę trudność zadania (punkt odniesienia 7.7%).

## 2.2 Podejście oparte na algorytmie EM

Gildea (2002) omawia kwestię wykorzystania zjawiska alternacji do klasyfikacji czasowników za pomocą statystycznego modelu opartego na algorytmie EM (Dempster *et al.*, 1977). Autor odwołuje się do koncepcji grupowania zaimków zaproponowanej przez Rootha i in. (1999). Użyli oni algorytmu EM opartego na prostym modelu probabilistycznym mówiącym, że rzeczownik $n$ na pozycji $r$ czasownika $v$ jest niezależny względem ukrytej zmiennej $\rho$ oznaczającej skupienie (ang. *cluster*).[9]

$$P(v,r,n) = \sum_{\rho} P(\rho)\, P(v,r|\rho)\, P(n|\rho). \tag{4}$$

Gildea wskazuje, że powyższe założenie nie jest takie oczywiste. W wypadku, gdy pomiędzy dwoma schematami (w pewnym uproszczeniu: pozycjami syntaktycznymi) czasownika zachodzi alternacja, niezależność względna nie ma miejsca. Aby uwzględnić takie sytuacje, autor dodaje kolejną ukrytą zmienną mającą reprezentować rolę semantyczną pełnioną przez rzeczownik, którą będziemy oznaczać przez $a$.

Dane do eksperymentów wydobywane były z korpusu BNC oznakowanego za pomocą statystycznego parsera Collinsa (1997). Jako że podmiot oznaczany był jako argument zewnętrzny w stosunku do czasownika, wyszukiwane były konstrukcje, w których podmiot NP1 był dominowany przez zdanie S wraz z

[8] W Joanis *et al.* (2008) rozważona została jeszcze jedna para klas.

[9] W algorytmach grupowania typowe jest używanie symbolu $c$ dla oznaczenia skupienia, jednak symbol $c$ używany jest do oznaczania klas semantycznych (w wypadku wordnetu synsetów). Różnica między użytymi czcionkami jest zbyt mała, by pomogła w interpretacji symbolu, dlatego posługuję się mniej oczywistym symbolem $\rho$.

frazą czasownikową VP, która z kolei dominowała dopełnienie NP2 (w wypadku schematów przechodnich). Następnie znajdowane były centra takich fraz, więc reprezentację można uznać za zbliżoną do wykorzystywanych w naszych eksperymentach rozbiorów zredukowanych (por. Hajnicz i Woliński, 2009; Hajnicz, 2009b, 2011).

Centra dopełnień zostały opatrzone wybranymi znacznikami funkcyjnymi (z dodanym znacznikiem null) z wersji 2 banku drzew Penn (Marcus, 1994; Marcus *et al.*, 1994) w celu odróżnienia rzeczywistych dopełnień od okolicznków, np. czasu. Sposób doboru takich znaczników został wytrenowany na fragmencie banku Penn zawierającym korpus WSJ w następujący sposób:

$$P(f|v,n) \quad = \quad \begin{cases} \hat{P}(f|v,n) & \text{dla } (v,n) \text{ współwystępujących} \\ & \text{w danych,} \\ \frac{1}{2}\hat{P}(f|v) + \frac{1}{2}\hat{P}(f|n) & \text{wpp.,} \end{cases}$$

gdzie $f$ to znacznik funkcyjny. Zauważmy brak pozycji $r$ w powyższej definicji wynikający z faktu, że rozważana jest wyłącznie pozycja dopełnienia. Jedynie wyrazy, które uzyskały w ten sposób znacznik null uznawane były za faktyczne dopełnienia.

Gildea (2002) rozważa następujące modele mogące służyć do grupowania czasowników:

1. **Potrójna zależność względem skupienia** — jednostki zależą wyłącznie od skupienia; zaproponowane w (Hoffman i Puzicha, 1998):

$$P_\rho(v,r,n) \quad = \quad P(\rho)\, P(v|\rho)\, P(r|\rho)\, P(n|\rho),$$

2. **Zależność czasownik–wymaganie** — opisana na wstępie propozycja Rootha i in. (1999):

$$P_{\rho_{v,r}}(v,r,n) \quad = \quad P(\rho)\, P(v,r|\rho)\, P(n|\rho),$$

3. **Zależność rzeczownik–wymaganie**:

$$P_{\rho_{n,r}}(v,r,n) \quad = \quad P(\rho)\, P(v|\rho)\, P(n,r|\rho),$$

4. **Alternacja**:

$$P_{alt}(v,r,n) \quad = \quad P(\rho)\, P(v|\rho)\, P(r|\rho)\, P(a|r,\rho)\, P(n|a,\rho).$$

Aby uzyskać prawdopodobieństwo $P(v, r, n)$, należy dokonać sumowania po zmiennych ukrytych. W szczególności, dla przypadku 2 jest to zapisane w równości (4). Dla najbardziej interesującego przypadku 4 opisującego alternacje zachodzi zaś następująca równość:

$$P(v, r, n) = \sum_{\rho, a} P(\rho)\, P(v|\rho)\, P(r|\rho)\, P(a|r, \rho)\, P(n|a, \rho). \tag{5}$$

Dla modeli 1–3 zmienna $\rho$ mogła przyjmować 128 wartości, dla modelu 4 — 64 wartości, co daje 128 rozkładów dla rzeczowników ($P(n|o, \rho)$), lecz tylko 64 dla czasowników ($P(v|\rho)$).

Ponieważ rozkładów prawdopodobieństw $P(\rho)$, $P(v|\rho)$, $P(r|\rho)$, $P(a|r, \rho)$ i $P(n|a, \rho)$ nie da się wydobyć bezpośrednio z danych, Gildea stosuje algorytm EM do wytrenowania parametrów tych rozkładów poprzez maksymalizację prawdopodobieństwa przy zadanych $\rho$ i $a$. Najlepsze wyniki pod względem rozwiązywania zadania grupowania czasowników dawał, wbrew wątpliwościom Gildei, zaproponowany przez Rootha i in. (1999), model 2.

Rzeczowniki nie są wstępnie w żaden sposób grupowane. Proponowane rozwiązanie ma jednak za zadanie nie tylko łączenie czasowników w skupienia, lecz także grupowanie rzeczowników występujących wspólnie na obu pozycjach. Rola $a$ reprezentuje zbiór rzeczowników współwystępujących na określonych pozycjach struktury argumentowej czasowników. Pojawienie się tej samej roli w dwóch wzorcach walencyjnych czasownika oznacza wystąpienie alternacji.

## 2.3 Grupowanie za pomocą preferencji selekcyjnych

Rozwiązując zadanie automatycznej klasyfikacji czasowników, Schulte im Walde (2000) także bierze pod uwagę ich zdolność do podlegania alternacjom. Jako podstawę swoich badań traktuje klasyfikację Levin (1993), uwzględniając występujące tam informacje o charakterze semantycznym czasowników (np. czasowniki ruchu), ich syntaktycznych realizacjach oraz wiążących je zależnościach o charakterze alternacji.

Dane wejściowe do eksperymentów stanowiło 5,5 mln. zdań wydobytych z BNC sparsowanego za pomocą parsera statystycznego opracowanego przez Carrolla i Rootha (1998), który tworzy las drzew wykorzystując do tego gramatykę bezkontekstową języka angielskiego, a następnie za pomocą leksykalnego modelu probabilistycznego wybiera z niego najbardziej prawdopodobne rozbiory. Ze zdań wydobywane są schematy walencyjne wraz z centrami fraz, czyli w naszej terminologii rozbiory zredukowane. Przykład takiej reprezentacji znajduje się w (6); `pn` oznacza nazwę własną.

(6) Sammut handled the plaudits during the awards ceremony.
(Sammut zyskał uznanie podczas ceremonii wręczania nagród.)
`handle subj*pn*sammut obj*plaudit pp*during*ceremony.`

Na tej podstawie Schulte im Walde wyszczególniła 88 schematów walencyjnych, które wystąpiły w danych przynajmniej 2 tys. razy (z dowolnym czasownikiem). Następnie policzyła wystąpienia czasowników z każdym spośród tych schematów, co stanowi ich syntaktyczną charakterystykę.

W celu udoskonalenia subkategoryzacji w stosunku do czystych schematów walencyjnych, autorka wykorzystała preferencje selekcyjne oznaczane w pewnej taksonomicznej hierarchii hierarchii klas i zdefiniowaną przez Resnika (1993) miarę powiązania selekcyjnego $A$ czasownika $v$ względem klasy semantycznej $c$ ma pozycji $r$

$$A(v,r,c) = \frac{P(c|v,r)\log\frac{P(c|v,r)}{P(c)}}{\sum_{c'} P(c'|v,r)\log\frac{P(c'|v,r)}{P(c')}}, \tag{7}$$

proponując jedynie prostszy sposób estymacji prawdopodobieństw:

$$\hat{p}(c|v,r) = \frac{\mathfrak{f}(v,c,r)}{\mathfrak{f}(v,r)}, \tag{8}$$

$$\hat{p}(c,r) = \frac{\mathfrak{f}(c,r)}{\sum_{c'\in\mathscr{C}}\mathfrak{f}(c',r)} = \frac{\mathfrak{f}(c,r)}{\mathfrak{f}(r)}, \tag{9}$$

$$\mathfrak{f}(c,r) = \sum_{n\in c}\frac{\mathfrak{f}(n,r)}{|\mathcal{C}_n|}, \tag{10}$$

gdzie $\mathscr{C}$ oznacza zbiór klas semantycznych, zaś $\mathcal{C}_n$ oznacza zbiór klas zawierających sensy rzeczownika $n$. Oszacowanie frekwencji klasy na pozycji $r$ z równości (10) zostało zaproponowane w (Ribas, 1994, 1995a,b).

Jak większość autorów, Schulte im Walde (2000) przeprowadzała swoje eksperymenty na hierarchii hipo-/hiperonimii WordNetu. Zamiast jednak wykorzystywać pełną strukturę hierarchii, rozważała jedynie 23 synsety. Było to 10 spośród 11-tu synsetów szczytowych. Synset Entity uznała za zbyt ogólny i zastąpiła jego 13-toma bezpośrednimi hiponimami. Tak więc nie rozważa ona tak naprawdę struktury hierarchicznej, lecz płaską kategoryzację, podobną do znakowania kategoriami semantycznymi w *Słowosieci*, por. Tabela 1.

Kryterium doboru czasowników do grupowania było z jednej strony zjawisko polisemii, a z drugiej ich frekwencja (wysoka vs. niska). Schulte im Walde wybrała do grupowania 153 czasowniki sklasyfikowane przez Levin (1993) posiadających 226 sensów, przynależnych do 30 różnych klas semantycznych.

Zadanie grupowania autorka realizuje za pomocą dwóch algorytmów, każdy z nich w oparciu o:

A. czystą informację syntaktyczną dotyczącą schematów syntaktycznych,

B. informację dotyczącą schematów syntaktycznych wraz z danymi na temat zachodzenia alternacji, reprezentowaną przez kombinację klas $C \subseteq \mathscr{C}$ preferowaną selekcyjnie na pozycji $r$ schematu $g$.

### 2.3.1 Algorytm iteracyjny

Pierwsze podejście wykorzystuje iteracyjny algorytm grupowania zaproponowany przez Hughesa (1994). Na początku każde skupienie zawiera pojedynczy czasownik. Iteracyjnie liczone były miary podobieństwa między skupieniami, a najbardziej podobne skupienia były łączone.

Każdy czasownik $v$ ma przypisany rozkład nad zbiorem schematów walencyjnych $G$. W zależności od zadania, rozważane prawdopodobieństwa to:

A.

$$\hat{p}(g|v) = \frac{\mathfrak{f}(v,g)}{\mathfrak{f}(v)}, \tag{11}$$

B.

$$p(g,C|v) \stackrel{def}{=} p(g|v)\,p(C|v,g), \tag{12}$$

$$\text{gdzie} \quad p(C|v,g) \stackrel{def}{=} \frac{\sum_{c\in C}\prod_{r\in R} A(v,c,r)}{\sum_{c'\in\mathscr{C}}\prod_{r\in R} A(v,c',r)}.$$

Schulte im Walde wykorzystywała entropię względną jako miarę podobieństwa dla tych rozkładów. Najbliższe skupienia były łączone, przy czym w następnym kroku obliczany był rozkład prawdopodobieństwa dla każdego skupienia jako średnia ważona jego elementów. Liczba iteracji była wyznaczana eksperymentalnie. Aby zapobiec powstawaniu zbyt dużych skupień, autorka ograniczyła liczbę przynależnych do nich czasowników do czterech. Jeśli dobrze rozumiem, faktyczna liczba ich elementów może być wyższa, o ile są to schematy tego samego czasownika. Chociaż takie ograniczenie wydaje się uzasadnione, dziwić może jego niski poziom.

### 2.3.2 Ukryta analiza skupień

Drugie podejście wykorzystuje metodę analizy ukrytych skupień opartą na algorytmie EM zaproponowaną w (Rooth, 1998). Wejście dla algorytmu stanowiły frekwencje czasowników z określonym schematem walencyjnym. Empirycznie zostało ustalone, że 80 skupień najlepiej modeluje semantyczne klasy czasowników. Zauważmy, że jest to liczba ponad 2,5 razy większa niż 30 klas, do których te czasowniki należą. Ponieważ jednak Schulte im Walde przyjęła, by dla zgodności z modelem iteracyjnym liczba czasowników w skupieniu nie przekraczała 4, mniejsza liczba skupień byłaby trudna do zaakceptowania ze względu na sumaryczną liczbę czasowników, gdyż „na siłę" równałaby ich liczebność do czterech.

Przy tych założeniach, wykonywane bez nadzoru 200 iteracji algorytmu EM grupuje pary $\langle v, g\rangle$ w 80 skupień $\rho$ za pomocą iteracyjnie estymowanych prawdopodobieństw:

A.

$$\hat{p}(v,g) \;=\; \sum_{\rho} \hat{p}(\rho,v,g) \;=\; \sum_{\rho} \hat{p}(\rho)\,\hat{p}(v|\rho)\,\hat{p}(g|\rho), \tag{13}$$

B.

$$\hat{p}(v,g,C) \;=\; \sum_{\rho} \hat{p}(\rho,v,g,C) \;=\; \sum_{\rho} \hat{p}(\rho)\,\hat{p}(v|\rho)\,\hat{p}(g,C|\rho). \tag{14}$$

Niestety, zaprezentowane przez autorkę wyniki ewaluacji pokazują, że dodanie informacji dotyczących preferencji selekcyjnych wpływa negatywnie na ich klasyfikację.

## 2.4 Grupowanie przy użyciu reguły Bayesa

Lapata i Brew (1999, 2004) stwierdzają, że chociaż aż 76% czasowników sklasyfikowanych przez Levin (1993) należy do pojedynczej klasy, ich frekwencja w korpusie BNC wynosi tylko 52% frekwencji wszystkich sklasyfikowanych czasowników. Wyciągają z tego wniosek, że czasowniki wieloznaczne są częściej używane w języku.

Odwołując się do pracy Lapaty (1999) wskazują, że 19 wzorców z klasyfikacji Levin zakłada występowanie schematu NP1 V NP2 NP3, 22 — schematu NP1 V NP2 *to* NP3, zaś 14 — schematu NP1 V NP2 *for* NP3. Nie dziwi też fakt, że stopień wieloznaczności czasownika rośnie wraz z liczbą realizowanych

przezeń schematów syntaktycznych. W pracy (Lapata i Brew, 2004) rozważali też schematy NP1 V NP2 *at* NP3 oraz NP1 V NP2.

Podstawowy model zakłada, że wybór klasy $\rho$ dla potencjalnie wieloznacznego czasownika $v$ przy zadanym schemacie $g$ polega na maksymalizacji ich łącznego prawdopodobieństwa:

$$p(\rho, v, g) = p(v)\, p(g|v)\, p(\rho|v, g). \tag{15}$$

Ponadto przyjęli założenie, że to klasa narzuca schematy walencyjne, tzn. czasowniki należące do tej samej klasy posiadają identyczne schematy.

$$p(\rho|v, g) \approx p(\rho|g). \tag{16}$$

Stosując z kolei regułę Bayesa, uzyskujemy

$$p(\rho|g) = \frac{p(g|\rho)\, p(\rho)}{p(g)}, \tag{17}$$

co po podstawieniu do (15) daje

$$p(\rho, v, g) = \frac{p(v)\, p(g|v)\, p(g|\rho)\, p(c)}{p(g)}. \tag{18}$$

Podobnie jak we wspomnianej pracy Lapaty, eksperymenty prowadzone były na danych wydobytych z korpusu BNC za pomocą parsera Gsearch (Keller *et al.*, 1999). Na podstawie tego dokonywana jest estymacja prawdopodobieństw:

$$\hat{p}(v) = \frac{\mathfrak{f}(v)}{\sum_{v' \in V} \mathfrak{f}(v')}, \qquad \hat{p}(g|v) = \frac{\mathfrak{f}(v, g)}{\sum_{g' \in G} \mathfrak{f}(v, g')},$$

$$\hat{p}(g) = \frac{\mathfrak{f}(g)}{\sum_{g' \in G} \mathfrak{f}(g')}, \qquad \hat{p}(g|\rho) = \frac{\mathfrak{f}(\rho, g)}{\sum_{g' \in G} \mathfrak{f}(\rho, g')},$$

$$\hat{p}(\rho) = \frac{\mathfrak{f}(\rho)}{\sum_{\rho'} \mathfrak{f}(\rho')}.$$

Ostatnia z tych równości sugeruje, że zbiór klas jest znany z góry.

Ze względu na brak danych dotyczących frekwencji $\mathfrak{f}(\rho, g)$, autorzy założyli, że wszystkie schematy występujące w danej klasie są równie prawdopodobne. Pozostałe frekwencje, których nie dało się bezpośrednio obliczyć na podstawie danych korpusowych, estymowane były jako:

$$\begin{aligned}
\hat{\mathfrak{f}}(\rho) &= \sum_{v \in V} \mathfrak{f}(v, \rho), \\
\hat{\mathfrak{f}}(g, \rho) &= \sum_{v \in V} \mathfrak{f}(v, g, \rho), \\
\hat{\mathfrak{f}}(v, \rho) &= \mathfrak{f}(v)\, \hat{p}(\rho|v).
\end{aligned}$$

Dla czasowników jednoznacznych, frekwencja $\mathfrak{f}(v, g, \rho) = \mathfrak{f}(v, g)$, więc może zostać uzyskana bezpośrednio z korpusu. Ze względu na brak korpusu znakowanego klasami Levin, czasowniki wieloznaczne są równomiernie dystrybuowane między klasy.

$$\hat{\mathfrak{f}}(v, g, \rho) = \frac{\mathfrak{f}(v, g)}{|\Upsilon_{v,g}|},$$

gdzie $\Upsilon_{v,g}$ jest to zbiór klas, do których należy czasownik $v$ ze schematem $g$. Nie jest jednak dla mnie jasne, skąd autorzy biorą informację o tym zbiorze lub choćby jego liczności. Podobnie,

$$\hat{p}(\rho|v) = \frac{\mathfrak{f}(v)}{|\Upsilon_v|},$$

gdzie rzecz jasna $\Upsilon_v = \sum_{g \in G} \Upsilon_{v,g}$.

Autorzy prezentują także bardziej wyrafinowane metody estymacji $\hat{\mathfrak{f}}(v, g, \rho)$ i $\hat{p}(\rho|v)$.

Jako punkt odniesienia, autorzy wykorzystali metodę, w której brak jest jakichkolwiek informacji syntaktycznych, tzn. maksymalizowane jest prawdopodobieństwo $p(v, \rho) = p(v)\, p(\rho|v)$.

Lapata i Brew (2004) przeprowadzili dwa eksperymenty. Pierwszy dotyczył czasowników, dla których schemat syntaktyczny jednoznacznie wskazywał klasę, do której dany czasownik należał. Przy założeniu równomiernej dystrybucji czasowników między klasy, trafność wahała się od 83,3% dla schematu NP1 V NP2 *at* NP3 (2,4% dla algorytmu stanowiącego punkt odniesienia) do 98,2% dla schematu NP1 V NP2 *for* NP3 (63,6%), a łącznie dla wszystkich schematów 93,9% (55,8%).

Drugi eksperyment przeprowadzony został dla 67 czasowników, które mimo posiadania jednego schematu walencyjnego przynależą do kilku klas wedle klasyfikacji Levin. Znów, przy założeniu równomiernej dystrybucji czasowników między klasy, trafność wahała się od 68,8% dla schematu NP1 V NP2 *to* NP3 (43,8%) do 100% dla schematów NP1 V NP2 *for* NP3 (0,00%)[10] i NP1 V NP2 *for* NP3 (100,00%), przy czym łącznie dla wszystkich schematów 74,6% (46,2%). Trudniejsze zadanie dało średnio gorsze wyniki. Idealny wynik dla dwóch klas mógł wynikać z ich niskiej liczebności.

# 3 Gradacyjna analiza odpowiedniości i skupień

W omawianych badaniach zdecydowałam się na wykorzystanie do grupowania czasowników gradacyjnej analizy skupień.

Gradacyjna analiza danych to gałąź szybko rozwijającej się dziedziny „wydobywania informacji danych" (ang. *data mining*). Stanowi ona istotny krok na drodze do integracji eksploracji danych, statystyki, taksonomii i teorii pomiaru oraz do jednolitego traktowania danych ciągłych i dyskretnych. Zastosowania metod gradacyjnych występują w szczególności tam, gdzie poszukuje się trendów (ukrytych struktur jednorodnych), skupień oraz elementów odstających od zaobserwowanych trendów (ang. *outliers*).

Termin *gradacyjny* pojawia się w literaturze probablistycznej w teorii kopuł w wyrażeniu *przekształcenie gradacyjne*. Pojęcie to zostało zdefiniowane w przystępny a zarazem ścisły sposób w (Kolev *et al.*, 2006; Niewiadomska-Bugaj i Kowalczyk, 2005), gdzie także można znaleźć definicję kopuły. Przekształcenie gradacyjne odwzorowuje zbiór dystrybuant wektora losowego $\mathbf{X}$ mającego $n$ dyskretno-ciągłych składowych o wartościach rzeczywistych w zbiór dystrybuant $n$-wymiarowych kopuł. Rozkłady brzegowe kopuł są jednostajne na przedziale $[0, 1]$; przy $n = 1$ dystrybuanta dowolnej dyskretno-ciągłej zmiennej $X$ zostaje przekształcona na dystrybuantę rozkładu jednostajnego na przedziale $(0, 1)$.

Kopuła, którą przekształcenie gradacyjne przyporządkowuje rozkładowi wektora $\mathbf{X}$, nosi nazwę rozkładu gradacyjnego tego wektora, zaś parametry powstałe z parametrów wektora $\mathbf{X}$ po tym przekształceniu nazywają się gradacyjnymi. Rozkład gradacyjny jest ciągły; parametrem gradacyjnym jest więc w szczególności gęstość gradacyjna wektora $\mathbf{X}$, czyli gęstość rozkładu przyporządkowanej mu kopuły. Gdy $n = 2$ i gdy rozkład pary $\langle X, Y \rangle$ jest dyskretny

[10] Przepisane z tabeli 10 str. 58 z (Lapata i Brew, 2004); jest to prawdopodobnie literówka.

z macierzą prawdopodobieństwa $P_{m\times k}[p_{ij},\ i=1,\ldots,m,\ j=1,\ldots,k]$, gęstość gradacyjna $h(u,v)$, gdzie $\langle u,v\rangle \in (0,1)^2$ jest stała na prostokątach $R_{ij}$, na które dzieli się kwadrat jednostkowy:

$$\begin{aligned} h(u,v) &= \frac{p_{ij}}{p_{i+}\,p_{+j}}, \\ \langle u,v\rangle &\in R_{ij} = [S_{i-1},S_i)\times[T_{j-1},T_j), \end{aligned} \tag{19}$$

gdzie rozkłady brzegowe $p_{i+}=\sum_{j=1}^{k} p_{i,j}$ i $p_{*j}=\sum_{i=1}^{m} p_{i,j}$ oraz zachodzi

$$\begin{aligned} S_i &= \sum_{s=1}^{i} p_{s+}, \\ T_j &= \sum_{t=1}^{j} p_{+t}. \end{aligned} \tag{20}$$

Gdy $X$ i $Y$ są niezależne, $p_{ij}=p_{i+}\,p_{+j}$, a więc gęstość gradacyjna jest wówczas tożsamościowo równa 1. Zatem gęstość gradacyjna w punkcie $\langle u,v\rangle$ może być interpretowana jako wskaźnik lokalnej nadreprezentacji rozkładu pary $\langle X,Y\rangle$ wobec rozkładu niezależnego o tych samych rozkładach brzegowych, i termin taki funkcjonuje. Wówczas wykres gęstości gradacyjnej (wyrażony na kwadracie jednostkowym za pomocą barw) nosi nazwę *mapy nadreprezentacji* pary $\langle X,y\rangle$. Nadreprezentacja może przyjmować dowolne wartości nieujemne; wartości z przedziału $[0,1)$ bywają nazywane *niedoreprezentacją*.

Ważnym przykładem parametru niezmienniczego ze względu na przekształcenie gradacyjne pary $\langle X,Y\rangle$ o rozkładzie dyskretno-ciągłym jest korelacja gradacyjna $\rho^\star$ (zwana inaczej współczynnikiem korelacji Spearmana): $\rho^\star(X,Y)$ przechodzi na współczynnik korelacji rozkładu gradacyjnego (będący w kopule jednocześnie współczynnikiem korelacji Spearmana) zachowując wartość.

Wskaźnik $\rho^\star$ dla tablicy prawdopodobieństwa $P_{m\times k}$ jest zadany wzorem

$$\rho^*(P_{m\times k}) = 3\sum_{i=1}^{m}\sum_{j=1}^{k} p_{i,j}\,(2S_{i-1}+p_{i+}-1)\,(2S_{j-1}+p{+}j-1) \tag{21}$$

gdzie $p_{i*}$, $p_{*j}$, $S_i$ oraz $T_j$ zdefiniowane są jak powyżej.

Wynik przekształcenia gradacyjnego pary $\langle X,Y\rangle$ zmienia się zazwyczaj w wyniku permutowania wierszy i/lub kolumn tablicy prawdopodobieństw. Parze $\langle X,Y\rangle$ można więc przyporządkować tak zwaną *maksymalną gradacyjną*

*korelację* $\rho^{\star}_{\max}(X,Y)$ uzyskaną poprzez maksymalizację wartości gradacyjnej korelacji $\rho^{\star}$ w zbiorze wszystkich rozkładów o permutowanych wierszach i/lub kolumnach.

Definicję przekształcenia gradacyjnego można rozszerzyć w taki sposób, by objęła ona zbiór dwudzielnych tablic kontyngencji $[M_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,k]$. Zbiór ten przekształcamy w zbiór tablic prawdopodobieństwa kładąc

$$p_{ij} \quad = \quad \frac{M_{ij}}{\sum_{s=1}^{m}\sum_{t=1}^{k} M_{st}}$$

Ogólniej, przekształcenie gradacyjne staje się narzędziem analizy danych, ilekroć dane przyjmują postać tablicy o wartościach nieujemnych, które można w sensowny sposób sumować po wierszach i kolumnach, tak jak to ma na ogół miejsce dla tablic kontyngencji.

Do metod gradacyjnych należy *gradacyjna analiza odpowiedniości* (ang. *Grade Correspondence Analysis*, GCA), wprowadzona w pracy (Ciok *et al.*, 1995). Jest to technika gradacyjna przekształcająca tablicę kontyngencji w tablicę (zestaw tablic) o tak spermutowanych wierszach i/lub kolumnach, by $\rho^{\star}$ przyjmowało wartość maksymalną $\rho^{\star}_{\max}$.

W praktyce, GCA polega na równoczesnym przestawianiu w tablicy danych jej wierszy i kolumn w taki sposób, by w wynikowej tablicy cecha porządkująca wiersze była jak najsilniej dodatnio zależna od cechy porządkującej kolumny. Powoduje to równocześnie poprawę regularności monotonicznej zależności w macierzy danych.

GCA została zaimplementowana w stworzonym w Instytucie Podstaw Informatyki PAN programie GradeStat przez dr. Olafa Matyję (2003). Algorytm przekształca początkową tablicę danych aż do uzyskania największej wartości $\rho^{\star}$. Jest to procedura typu Monte-Carlo, więc maksymalizuje wartość $\rho^{\star}$ w sposób przybliżony i jej wyniki nie zawsze są identyczne przy kolejnych uruchomieniach dla tych samych danych.

GCA służy najczęściej do podziału wierszy i/lub kolumn wynikowej tablicy na skupienia. Liczbę skupień wierszy i liczbę skupień kolumn określa osoba analizująca dane. Liczby te mogą być determinowane z góry ze względu na charakter danych, często jednak wykonuje się wiele kolejnych podejść, przeglądając uzyskane wyniki. Procedura podziału na skupienia nosi nazwę *gradacyjna analiza odpowiedniości i skupień* (ang. *Grade Correspondence-Cluster Analysis*, GCCA). GCCA dzieli wiersze i/lub kolumny na rozłączne skupienia zawierające wyłącznie elementy sąsiadujące. Dokonany podział ma tę własność, że

| skrót | nazwa | skrót | nazwa | skrót | nazwa |
|---|---|---|---|---|---|
| cech | cecha | ksz | kształt | sbst | substancja |
| cel | motyw | msc | miejsce | st | stan |
| czas | czas | os | osoba | umy | poznanie |
| czc | ciało | por | komunikacja | wytw | wytwór |
| czuj | odczucie | pos | posiadanie | zdarz | zdarzenie |
| czy | czynność | prc | proces | zj | zjawisko |
| grp | grupa | rsl | roślina | zwr | zwierzę |
| il | ilość | rz | obiekt fiz. | zwz | związek |
| jedz | jedzenie | | | | |

Tabela 1: Kategorie semantyczne rzeczowników

po agregacji wierszy i kolumn każde skupienie tworzy tablicę o maksymalnym wskaźniku $\rho^\star$ w zbiorze wszystkich podziałów z zadaną liczbą (parą liczb) skupień. Oznacza to, że dokonany podział zapewnia najmniejszą możliwą utratę siły i regularności zależności monotonicznej.

Więcej informacji na temat GCA można znaleźć w (Ciok *et al.*, 1995; Kowalczyk *et al.*, 2004), zaś bardziej przystępny opis wraz zastosowaniami w (Książyk *et al.*, 2005).

# 4 Syntaktyczno-semantyczny słownik walencyjny

Twórcy opisanych w sekcji 2 metod klasyfikacji czasowników wykorzystywali do realizacji tego zadania wielu zestawów cech, syntaktycznych i semantycznych. Punktem wyjścia eksperymentów opisanych w następnej sekcji był eksperymentalny słownik syntaktyczno-semantyczny zawierający 32 czasowniki posiadające bogaty zestaw schematów syntaktycznych.

Słownik został utworzony automatycznie w oparciu o korpus rozbiorów zredukowanych (Hajnicz, 2009b, 2011, rozdz. 11) oznakowany semantycznie kategoriami pochodzącymi z polskiego wordnetu zwanego *Słowosiecią* (Piasecki *et al.*, 2009). Zestaw kategorii semantycznych rzeczowników widnieje w Tabeli 1, zaś czasowników w Tabeli 2. Istnieją dwie wersje słownika: *słownik prosty* $\mathfrak{D}$ złożony z ram prostych, w którym każdy argument czasownika opatrzony jest kategorią (Hajnicz, 2009a, 2011, sekcja 12.1) oraz *słownik zagregowany* $\widetilde{\mathfrak{D}}$, w którym powiązane semantycznie ramy proste zostały połączone w ramy złożone, mające w założeniu reprezentować pojedyncze znaczenie czasownika.

| skrót | nazwa | skrót | nazwa | skrót | nazwa |
|---|---|---|---|---|---|
| cczuj | emocje | cumy | poznanie | pst | percepcja |
| cjedz | konsumpcja | cwytw | wytwarzanie | ruch | ruch |
| cpor | komunikacja | dtk | dotyk | sp | społeczne |
| cpos | posiadanie | hig | higiena | wal | rywalizacja |
| cst | statyczne | pog | pogoda | zmn | zmiana |

Tabela 2: Kategorie semantyczne czasowników

Formalnie rzecz ujmując, $\mathfrak{D} = \{\langle v, g, f\rangle, m_{v,g}, k_{v,g,f}\}$, gdzie $\langle v, g\rangle \in \mathcal{D}$ jest schematem syntaktycznym czasownika $v$, $f \in F_g$ jest jedną z jego ram semantycznych, $m_{v,g}$ oznacza frekwencję schematu $g$ z czasownikiem $v$, zaś $k_{v,g,f}$ oznacza frekwencję ramy dla tego schematu. Oczywiście $\sum_{f \in F_g} k_{v,g,f} = m_{v,g}$. Natomiast $\tilde{\mathfrak{D}} = \{\langle v, g, \tilde{f}\rangle, m_{v,g}, k_{v,g,\tilde{f}}\}$, przy czym dla dowolnej $\langle v, g, f\rangle \in \mathfrak{D}$ istnieje dokładnie jedna $\langle v, g, \tilde{f}\rangle \in \tilde{\mathfrak{D}}$ taka, że $f \in \tilde{f}$ oraz $k_{v,g,\tilde{f}} = \sum_{f \in \tilde{f}} k_{v,g,f}$. Przykładowe hasło słownika prostego $\mathfrak{D}$ dla schematu `np:dat np:nom` czasownika *minąć* została przedstawiona w (22), zaś odpowiadające jej hasło słownika zagregowanego $\tilde{\mathfrak{D}}$ w (23).

(22) minąć `np:dat` `np:nom` 63
dat: grp; nom: czas 5
dat: grp; nom: czy 1
dat: grp; nom: zdarz 2
dat: os; nom: czas 17
dat: os; nom: czy 8
dat: os; nom: prc 1
dat: os; nom: zdarz 5

(23) minąć `np:dat` `np:nom` 63
dat: grp,os; nom: cech 6
dat: grp,os; nom: czas,czy,prc,zdarz 39
dat: os; nom: czuj,umy 18

# 5 Zastosowanie GCCA do klasyfikacji czasowników polskich

Czasowniki klasyfikowane były w oparciu o zbliżone preferencje selekcyjne reprezentowane przez poszczególne ramy proste. Grupowanie czasowników przeprowadzone zostało dla obu słowników. W wypadku słownika prostego $\mathfrak{D}$ wszystkie ramy proste schematu rozważane były łącznie.W wypadku słownika zagregowanego $\tilde{\mathfrak{D}}$ każda rama zagregowana została rozpatrzona oddzielnie.

W obu przypadkach utworzona została tablica kontyngencji. W zależności od schematu, każdy wiersz zawierał frekwencje wszystkich par $a = \langle r, c\rangle$, przy czym $g = \langle \ldots, r, \ldots\rangle$. Formalnie rzecz ujmując, dla słownika prostego $\mathfrak{D}$ komórka macierzy kontyngencji

$$M_{v,a} \quad = \quad \sum_{\langle v,g,f\rangle \in \mathfrak{D}_{v,g}} k_{v,g,f}, \qquad \text{gdzie} f = \langle \ldots, a, \ldots\rangle.$$

Natomiast dla słownika zagregowanego $\tilde{\mathfrak{D}}$ oraz dowolnej $\langle v, g, \tilde{f}\rangle \in \tilde{\mathfrak{D}}$

$$M_{\tilde{f},a} \quad = \quad \sum_{f \in \tilde{f}} k_{v,g,f}, \qquad \text{gdzie} f = \langle \ldots, a, \ldots\rangle.$$

Zauważmy, że w pierwszym wypadku liczba wierszy jest równa liczbie czasowników posiadających dany schemat. Natomiast w drugim wypadku jest to liczba ram zagregowanych takich czasowników, czyli znacznie, znacznie więcej.

W Tabeli 3 przedstawiona jest najprostsza tablica kontyngencji, utworzona dla rozdystrybuowanych ram czasowników schematu `<np:nom  że>`. Ponieważ schemat ten zawiera tylko jedną frazę nominalną `np:nom`, w Tabeli pominięta została informacja o typie frazy, tzn. zamiast 'nom: cech', 'nom: czy' itd. w Tabeli widnieje samo 'cech', 'czy' itd. Warto zwrócić uwagę, że jeszcze przed uporządkowaniem widać, że jest to tablica regularna i w stosunku do rozmiaru rzadka.

Uzyskane tablice kontyngencji[11] poddane zostały grupowaniu za pomocą metody GCCA (por. sekcja 3). Ustalona liczba skupień wynosiła 12, gdyż tyle kategorii semantycznych przypisanych w *Słowosieci* posiadały rozważane czasowniki (por. Tabela 2).

[11] Po usunięciu zbyt nielicznych tablic przebadane zostało 8 schematów dla niepodzielonych czasowników i 23 schematy dla ram zagregowanych.

| lemat | cech | czy | grp | il | msc | os | por | umy | wytw | zdarz | zw |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mówić | 0 | 7 | 46 | 5 | 0 | 387 | 9 | 20 | 93 | 4 | 2 |
| pisać | 1 | 0 | 2 | 0 | 0 | 10 | 0 | 0 | 8 | 0 | 0 |
| powiedzieć | 0 | 0 | 7 | 0 | 0 | 297 | 0 | 0 | 2 | 0 | 0 |
| powtarzać | 0 | 0 | 3 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 0 |
| powtórzyć | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| proponować | 0 | 0 | 10 | 0 | 0 | 7 | 0 | 0 | 1 | 0 | 1 |
| przyjmować | 0 | 0 | 6 | 0 | 0 | 5 | 0 | 0 | 1 | 0 | 0 |
| przyjąć | 0 | 0 | 12 | 0 | 2 | 4 | 0 | 0 | 0 | 0 | 0 |
| widzieć | 0 | 0 | 2 | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 0 |

Tabela 3: Przykładowa tablica kontyngencji (dla schematu `<np:nom że>`)

W wypadku $M_{v,g,a}$ zawierającej poniżej 15 wierszy liczba skupień była obligatoryjnie zmniejszana proporcjonalnie do jego liczności, by niejako wymusić skupienia kilkuelementowe.[12]

(24) a. `<:np:acc: :np:nom:>`

1. { interesować }
2. { odnieść, spotkać, uderzyć }
3. { przechodzić, przejść }
4. { rozpoczynać, rozpocząć, zacząć }
5. { kończyć, powtórzyć, trzymać, zaczynać }
6. { bronić, mówić, robić }
7. { odnosić, widzieć }
8. { lubić }
9. { powiedzieć, przyjmować, skończyć }
10. { postawić, powtarzać }
11. { kupić, pisać, przygotować, przygotowywać }
12. { przyjąć }

b. `<:np:acc: :np:nom: :prepnp:do:gen:>`

1. { mówić, powiedzieć }
2. { pisać }
3. { odnosić }
4. { postawić, przygotować }
5. { odnieść, przygotowywać }

[12]Należałoby w zasadzie sprawdzić, jakie kategorie przypisane zostały czasownikom posiadającym dany schemat, ale niewiele by to miało wspólnego z automatycznym przetwarzaniem danych.

W (24) można zobaczyć przykładowy podział czasowników na skupienia dla kilku schematów. Kolejność skupień jest znacząca: bardziej podobne znajdują się bliżej siebie.

W wypadku grupowania ram zagregowanych sprawa jest oczywiście o wiele bardziej złożona ze względu na fakt, że naszym celem jest grupowanie czasowników, a nie ich ram. Oczywiście teoretycznie rama zagregowana reprezentuje pojedyncze znaczenie czasownika, jednak do takiej interpretacji liczba przypisywanych im kategorii semantycznych jest zbyt mała (a w konsekwencji także liczba skupień). Z drugiej strony, wykorzystanie w tym celu synsetów mija się z celem przy tak niewielkiej liczbie czasowników: tylko synonimy niektórych z nich znajdują się w zbiorze ChVlist (i to nie dla wszystkich reprezentujących ich znaczenia jednostek leksykalnych).

Dlatego podjęte zostały następujące kroki upraszczające wyniki grupowania.

1. Dwie ramy zagregowane występujące w tym samym skupieniu były łączone ze względu na przyjęcie założenia, że reprezentują to samo znaczenie czasownika.
2. Skupienia będące podzbiorami innych skupień były usuwane w celu preferowania większych skupień.

Wynik grupowania dla przykładowej tablicy $M_{\tilde{f},a}$, już po dokonaniu łączenia ram w kroku 1, znajduje się w (25). Skupienia usunięte w kroku 2 oznaczone są przez −.

(25) `<:np:acc: :np:inst: :np:nom:>`

1. { skończyć,zakończyć } —
2. { kończyć,kupić,skończyć,uderzyć,zakończyć }
3. { kończyć,skończyć,zacząć,zakończyć }
4. { kończyć,robić,rozpoczynać }
5. { kończyć,rozpocząć,skończyć,zaczynać,zacząć }
6. { zakończyć } —
7. { pisać,robić,rozpoczynać,uderzyć,widzieć,zakończyć }
8. { kończyć,rozpoczynać,rozpocząć,zacząć }
9. { zakończyć } —
10. { kończyć,rozpoczynać,rozpocząć } —
11. { kończyć,robić,rozpocząć,zacząć,zakończyć }
12. { widzieć,zakończyć } —

Cały opisany w (Hajnicz, 2011) proces tworzenia semantycznego słownika semantycznego oparty był na operacjach dokonywanych dla każdego czasownika

niezależnie. Dlatego przyjęta do eksperymentów liczba 32 czasowników wydawała się całkowicie wystarczająca. Grupowanie czasowników w najoczywistszy sposób własności tej nie posiada, wręcz przeciwnie, odbywa się dla wszystkich czasowników jednocześnie i może w wyrazisty sposób zależeć od ich liczności i doboru.[13] Dlatego wyniki poniższej ewaluacji należy przyjmować z niejakim sceptycyzmem.

## 5.1 Ewaluacja

Do ewaluacji grupowania czasowników posłużę się wskaźnikiem jednorodności skupień (ang. *cluster purity*) (Kim *et al.*, 2004; Zhao *et al.*, 2001). Niech więc $\mathcal{C}_v$ będzie zbiorem kategorii semantycznych czasownika $v$. Ponadto niech $\mathcal{C}^g$ będzie zbiorem skupień schematu $g$, $\mathcal{C}$ będzie zbiorem wszystkich skupień, $C_j^g$ będzie $j$-tym skupieniem schematu $g$, zaś $C_j^{g,c}$ będzie jego podzbiorem zawierającym wyłącznie czasowniki posiadające kategorię $c$. Przy ustalonym schemacie $g$ będziemy pisać $C_j$ i $C_j^c$, odpowiednio. Wówczas miara jednorodności skupień wyraża się wzorem:

$$\mathcal{P}(C_j) = \frac{1}{|C_j|} \max_c(|C_j^c|), \tag{26}$$

$$\mathcal{P}(E) = \frac{1}{k} \sum_{j=1}^{k} \mathcal{P}(C_j), \qquad \text{przy czym} \quad E = \bigcup_{j=1}^{k} C_j. \tag{27}$$

Wskaźnik ten jednak musi zostać poddany pewnym modyfikacjom ze względu na konieczność rozwiązania następujących problemów:

- czasowniki nie są jednoznacznie kategoryzowane (mogą posiadać kilka kategorii),
- grupowanie ram zagregowanych sprowadzić można do grupowania wielu egzemplarzy tego samego czasownika, czyli jeden czasownik może należeć do kilku skupień,
- grupowanie nie było przeprowadzane w jednym zbiorze, lecz dla każdego schematu walencyjnego oddzielnie.

W celu uwzględnienia tych kwestii należało rozróżnić dwa czynniki.

- Sposób zliczania kategorii czasowników, co wpływa na wartość wskaźnika jednorodności pojedynczego skupienia:

[13] Warto jednak zauważyć, że zawsze będą pojawiać się rzadkie schematy, powiązane z niewielką liczbą czasowników.

– Zliczanie w proporcji do liczby kategorii czasownika:

$$\mathcal{P}(C_j) = \frac{1}{|C_j|} \max_{c} \sum_{v \in C_j} \frac{\mathbf{1}_{\mathcal{C}_v}(c)}{|\mathcal{C}_v|}, \tag{28}$$

gdzie $\mathbf{1}_A$ jest funkcją charakterystyczną zbioru $A$.

– Uwzględnianie tylko tych kategorii czasowników niejednoznacznych, które są dominujące w skupieniu dla czasowników jednoznacznych:

$$\begin{aligned} \mathcal{P}(C_j) &= \frac{1}{|C_j|} \max_{c \in \mathcal{C}_j^{\max}} |C_j^c \cup \{v \in C_j - \overline{C_j} \colon c \in \mathcal{C}_v\}|, \\ \mathcal{C}_j^{\max} &= \arg\max_{c} |C_j^c|, \end{aligned} \tag{29}$$

gdzie $\overline{C_j} \subseteq C_j$ jest podzbiorem czasowników jednoznacznych w $C_j$, zaś $C_j^c \subseteq C_j$ jest zbiorem czasowników skupienia $j$ posiadających jednoznacznie kategorię $c$.

- Sposób wyliczania sumarycznego wskaźnika jednorodności skupień:
  – Wyliczanie średniej jednorodności po jej obliczeniu dla każdego schematu oddzielnie

$$\mathcal{P}(\mathfrak{C}) = \frac{1}{|G|} \sum_{g \in G} \frac{1}{|\mathcal{C}^g|} \sum_{C_j^g} \mathcal{P}(C_j). \tag{30}$$

  – Wyliczanie jednorodności dla wszystkich skupień wszystkich schematów łącznie

$$\mathcal{P}(\mathfrak{C}) = \frac{1}{|\mathcal{C}|} \sum_{g \in G} \sum_{C_j^g} \mathcal{P}(C_j). \tag{31}$$

Łącząc te dwa czynniki, uzyskujemy cztery wskaźniki oceniające w nieco odmienny sposób jednorodność skupień czasowników:

A. średnia jednorodność zbioru przy proporcjonalnym zliczaniu jednorodności skupienia (równości (28) i (30);

B. średnia jednorodność zbioru przy preferowaniu dominujących kategorii semantycznych dla czasowników niejednoznacznych (29) i (30);

C. łączna jednorodność zbioru przy proporcjonalnym zliczaniu jednorodności skupienia (równości (28) i (31);

D. łączna jednorodność zbioru przy preferowaniu dominujących kategorii semantycznych dla czasowników niejednoznacznych (29) i (31).

Wyniki ewaluacji zostały zaprezentowane w Tabeli 4. Pierwsza kolumna identyfikuje zestawy skupień: *czasowniki* oznacza grupowanie pojedynczych czasowników, *ramy* oznacza grupowanie ram zagregowanych, *maksymalne* oznacza zestaw skupień uzyskany po usunięciu podzbiorów. W drugiej kolumnie znajduje się średnia liczba czasowników w skupieniu w danym zestawie, pozostałe kolumny zawierają właściwe wyniki ewaluacji.

| zestaw | średnia | A | B | C | D |
|---|---|---|---|---|---|
| czasowniki | 1,68 | 61,23 | 55,63 | 62,91 | 58,18 |
| ramy | 2,64 | 60,91 | 58,49 | 61,06 | 58,64 |
| maksymalne | 4,76 | 50,05 | 57,65 | 54,96 | 57,48 |

Tabela 4: Ewaluacja grupowania czasowników

Biorąc pod uwagę średni rozmiar skupienia, wynik grupowania pojedynczych czasowników należy uznać za zły, gdyż jest on nieznacznie jedynie lepszy od losowego. Należy jednak wziąć pod uwagę ograniczenia samego zbioru danych, niezależne od zastosowanej metody. Natomiast wyniki grupowania ram zagregowanych wydają się zaskakująco dobre, jeśli weźmie się pod uwagę poziom skomplikowania problemu, szczególnie dla zestawu z usuniętymi podskupieniami.

# 6 Podsumowanie

Niniejszy raport zawiera opis wstępnych eksperymentów dotyczących klasyfikowania czasowników na podstawie ich semantycznej charakterystyki. Mały rozmiar (32 czasowniki) słownika dostępnego do eksperymentów stanowi uzasadnienie słabej jakości uzyskanych wyników. Dlatego eksperymenty przy wykorzystaniu innych metod nie były już prowadzone.

Badania nad automatyczną klasyfikacją czasowników polskich będą kontynuowane. Trwają bowiem pracę nad tworzeniem większego słownika o lepszej jakości metodą ręczną i półautomatyczną. Wydaje się, że już słownik rozmiaru 100-200 czasowników nadaje się do eksperymentów.

Planowane eksperymenty mają obejmować nie tylko szerszy zestaw metod klasyfikacji, ale przede wszystkim bardziej zaawansowane struktury danych. Punktem wyjścia ma być semantyczny słownik walencyjny w miejsce słownika

syntaktyczno-semantycznego. Każde hasło takiego słownika definiuje znaczenie czasownika, a jego ramy są ciągiem czysto semantycznych argumentów (wraz z ich preferencjami selekcyjnymi). Z hasłem powiązane są wszystkie schematy syntaktyczne realizujące dane znaczenie czasownika. Tak ustrukturyzowane dane są znacznie bogatszym źródłem wiedzy leksykalnej, jednak rzecz jasna trudniej jest nimi operować.

# Bibliografia


ARPA (1994) *Proceedings of the ARPA Human Language Technology Workshop*, Morgan Kaufmann, Princeton, NJ.

C. F. Baker, C. J. Fillmore, J. B. Lowe (1998) *The Berkeley FrameNet Project*, w: *Proceedings of COLING-ACL'98*, s. 86–90, Montreal, Kanada.

G. Carroll, M. Rooth (1998) *Valence Induction with a Head-Lexicalized PCFG*, w: *Proceedings of (EMNLP-1998)*, s. 36–45, Granada, Hiszpania.

A. Ciok, T. Kowalczyk, E. Pleszczyńska, W. Szczesny (1995) *Algorithms of grade correspondence-cluster analysis*, Archiwum Informatyki Teoretycznej i Stosowanej, t. 7, nr 1–4, s. 5–22.

M. J. Collins (1997) *Three generative, lexicalised models for statistical parsing*, w: *Proceedings of (ACL'97)*, s. 16–23, Madrid, Spain.

H. T. Dang (2004) *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation*, Rozprawa doktorska, Computer and Information Science Department, University of Pennsylvania.

A. P. Dempster, N. M. Laird, D. B. Rubin (1977) *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, series B, t. 39, s. 185–197.

C. J. Fillmore, C. R. Johnson, M. R. L. Petruck (2003) *Background to FrameNet*, International Journal of Lexicography, t. 16, nr 3, s. 235–250.

D. J. Gildea (2002) *Probabilistic model of verb-argument structure*, w: *Proceedings of (CoNNL-2002)*, s. 308–314, Taipei, Tajwan.

E. Hajnicz (2009a) *Problems with Pruning in Automatic Creation of Semantic Valence Dictionary for Polish*, w: V. Matoušek, P. Mautner (red.), *Proceedings of TSD 2009*, t. 5729 serii *Lecture Notes in Artificial Intelligence*, s. 131–138, Springer-Verlag, Pilzno, Czechy.

— (2009b) *Semantic annotation of verb arguments in shallow parsed Polish sentences by means of EM selection algorithm*, w: M. Marciniak, A. Mykowiecka (red.), *Aspects of Natural Language Processing*, t. 5070 serii *Lecture Notes in Computer Science*, s. 211–240, Springer-Verlag.

— (2011) *Proces tworzenia semantycznego słownika walencyjnego*, Inżynieria Lingwistyczna, Akademicka Oficyna Wydawnicza Exit, Warszawa.

E. Hajnicz, M. Woliński (2009) *How Valence Information Influences Parsing Polish with* Świgra, w: M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, K. Trojanowski (red.), *Recent Advances in Intelligent Information Systems*, Challenging Problems in Science: Computer Science, s. 193–206, Akademicka Oficyna Wydawnicza Exit, Warszawa.

T. Hoffman, J. Puzicha (1998) *Statistical Models for co-occurrence data*, Memo.

J. Hughes (1994) *Automatically Acquiring Classification of Words*, Rozprawa doktorska, School of Computer Studies, University of Leeds, Leeds.

E. Joanis, S. Stevenson (2003) *A general feature space for automatic verb classification*, w: *Proceedings of (EACL-2003)*, s. 163–170, Budapeszt, Węgry.

E. Joanis, S. Stevenson, D. James (2008) *A general feature space for automatic verb classification*, Natural Language Engineering, t. 14, nr 3, s. 337–367.

F. Keller, M. Corlay, S. Corlay, M. W. Crocker, S. Trevin (1999) *Gsearch: a tool for syntactic investigation of unparsed corpora*, w: *Proceedings of EACL-1999*, s. 56–63, Bergen, Norwegia.

M. Kim, H. Yoo, R. S. Ramakrishna (2004) *Cluster Validation for High-Dimensional Datasets*, w: C. Bussler, D. Fensel (red.), *Proceedings of the 11th International Conference on Artificial Intelligence Methods, Systems and Applications*, t. 3192 serii *Lecture Notes in Computer Science*, s. 178–187, Springer-Verlag, Berlin / Heidelberg.

N. Kolev, B. Vaz de Mendes, U. dos Anjos (2006) *Copulas: a Review and Recent Developments*, Stochastic Models, t. 22, nr 4, s. 617–660.

T. Kowalczyk, E. Pleszczyńska, F. Ruland (red.) (2004) *Grade Models and Methods for Data Analysis. With Applications for the Analysis of Data Populations*, Studies in Fuzziness and Soft Computing, Springer-Verlag, Berlin Heidelberg New York.

J. Książyk, O. Matyja, E. Pleszczyńska, M. Wiech (2005) *Analiza danych medycznych i demograficznych przy użyciu programu GradeStat*, Instytut Podstaw Informatyki, Polska Akademia Nauk Instytut „Pomnik — Centrum Zdrowia Dziecka", Warszawa.

M. Lapata, C. Brew (1999) *Using subcategorization to resolve verb class ambiguity*, w: *Proceedings of the Joint SIGDAT Conference on Empirical methods in NLP and Very Large Corpora*, s. 266–274, College Park, MD.

— (2004) *Verb class disambiguation using informative priors*, Computational Linguistics, t. 30, nr 1, s. 45–73.

B. Levin (1993) *English verb classes and alternation: a preliminary investigation*, University of Chicago Press, Chicago, IL.

M. P. Marcus (1994) *The Penn TreeBank: A revised corpus design for extracting predicate-argument structure*, w: ARPA.

M. P. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, B. Schasberger (1994) *The Penn Treebank: Annotating predicate argument structure*, w: ARPA, s. 114–119.

O. Matyja (2003) *Smooth Grade Correspondence Analysis and Related Computer System*, Rozprawa doktorska, Instytut Podstaw Informatyki, Polska Akademia Nauk.

D. McCarthy (2001) *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*, Rozprawa doktorska, University of Sussex.

P. Merlo, S. Stevenson (2001) *Automatic verb classification based on statistical distributions of argument structure*, Computational Linguistics, t. 27, nr 3, s. 373–408.

P. Merlo, S. Stevenson, V. Tsang, G. Allaria (2002) *A multilingual paradigm for automatic verb classification*, w: *Proceedings of (ACL'02)*, s. 207–214, Philadelphia, PA.

M. Niewiadomska-Bugaj, T. Kowalczyk (2005) *On grade transformation and its implications for copulas*, Brazilian Journal of Probability and Statistics, t. 19, s. 125–137.

M. Piasecki, S. Szpakowicz, B. Broda (2009) *A Wordnet from the Ground Up*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

P. Resnik (1993) *Selection and Information: A Class-Based Approach to Lexical Relationships*, Rozprawa doktorska, University of Pennsylvania, Filadelphia, PA.

F. Ribas (1994) *An Experiment on Learning Appropriate Selectional Restrictions from Parsed Corpus*, w: *Proceedings of (COLING-1994)*, s. 769–774, Kioto, Japonia.

— (1995a) *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy*, Rozprawa doktorska, University of Catalonia.

— (1995b) *On Learning More Appropriate Selectional Restrictions*, w: *Proceedings of (EACL'95)*, s. 112–118, Dublin, Irlandia.

M. Rooth (1998) *Two-Dimensional Clusters in Grammatical Relations*, Rap. tech. 3, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

M. Rooth, S. Riezler, D. Prescher, G. Carroll, F. Beil (1999) *Inducing a semantically annotated lexicon via EM-based clustering*, w: *Proceedings of (ACL'99)*, s. 104–111, College Park, MA.

L. Shi, R. Mihalcea (2005) *Putting Pieces Together: Combining FramNet, VerbNet and PropBank for Robust Semantic Parsing*, w: A. Gelbukh (red.), *Proceedings of CICLing-2005*, t. 3406 serii *Lecture Notes in Computer Science*, s. 100–111, Springer-Verlag, Heidelberg.

S. Stevenson, P. Merlo (1997) *Lexical structure and parsing complexity*, Language and Cognitive Process, t. 12, nr 2/3, s. 349–399.

— (1999) *Automatic Verb Classification Using Distributions of Grammatical Features*, w: *Proceedings of (EACL'99)*, s. 45–52, Bergen, Norwegia.

R. S. Swier, S. Stevenson (2004) *Unsupervised semantic role labelling*, w: *Proceedings of (EMNLP-2004)*, s. 95–102, Barcelona, Hiszpania.

S. Schulte im Walde (2000) *Clustering verbs semantically according to their alternation behaviour*, w: *Proceedings of (COLING-2000)*, s. 747–753, Saarbrücken, Niemcy.

B. Zapirain, E. Agirre, L. Màrquez (2008) *Robustness and generalization of role sets: PropBank vs. VerbNet*, w: *Proceedings of (ACL'08)*, s. 550–558, Columbus, OH.

Y. Zhao, , G. Karypis (2001) *Criterion Functions for Document Clustering: Experiments and Analysis*, Rap. tech. #01–40, University of Minnesota, Minneapolis, MN.

Pracę zgłosił Adam Przepiórkowski


| | |
|---|---|
| Adres autorki: | Elżbieta Hajnicz<br>Instytut Podstaw Informatyki PAN<br>ul. Ordona 21<br>01-237 Warszawa<br>Polska |
| e-mail: | Elzbieta.Hajnicz@ipipan.waw.pl |



Symbol klasyfikacji rzeczowej: CR: I.2.7


Na prawach rękopisu
Printed as manuscript


Nakład 100 egzemplarzy. Papier kserograficzny klasy III. Oddano do druku w maju 2011. Wydawnictwo IPI PAN. ISSN 0138-0648.