

Projekt nowoczesnego polsko < > ukraińskiego słownika elektronicznego

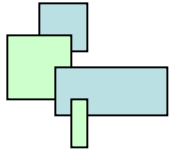
- Natalia Kotsyba

Instytut Slawistyki, Polska Akademia Nauk

- Magdalena Turska

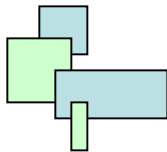
Uniwersytet Warszawski

<http://corpus.domeczek.pl>



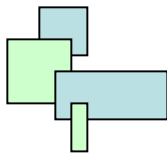
Polsko-ukraiński korpus równoległy

- nie istnieje,
- a przecież mógłby być materiałem wyjściowym do
 - nauki języka
 - konstrukcji słowników
 - badań językowych



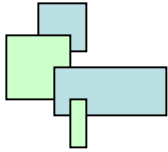
Plan

- Korpus II
- Konkordancer II
- Środowisko sieciowe do grupowej pracy nad słownikiem



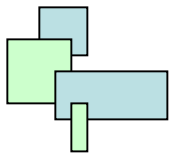
A w „międzyczasie”

- Analiza frekwencyjna
- Analiza walencji
- Analiza związków frazeologicznych i innych jednostek językowych



Korpus

- Problematyczne jest pozyskanie odpowiedniej ilości materiałów równoległych
- Konieczność posilkowania się korpusami jednojęzycznymi, np. IPI PAN i ULIF – dodatkowe komplikacje techniczne i organizacyjne



Korpus równoległy - technicznie

- Gromadzenie materiałów źródłowych
- Przetworzenie do sformalizowanej postaci elektronicznej (TEI XML)
- Analiza morfologiczna i alignment (anotacja)
- zapis do postaci umożliwiającej efektywne przeszukiwanie oraz ochronę praw autorskich

Polsko-ukraiński korpus równoległy - Microsoft Internet Explorer

Plik Edycja Widok Ulubione Narzędzia Pomoc

Wstecz Wyszukaj Ulubione

Adres http://localhost/corpus/index.php?option=korpus

Przejdź Łącz

Polsko-ukraiński korpus równoległy

nie przecze mebel piękny na pewno He pobralis',
nie znalis',
A on roztrinaot'
a jeszcze ples pasowy to juz cywilizacja A on dobrot'
Wdovu za podushie, a sina kutyot',
Chodzie wszystkie stany Lita ml m'odli
Kolorowi, biali, czarni
Chodzie zwlaszcza wy, ludko
Przez na ościez bramy ... A blode pto mnyot'
Opuhla dityna - golodnes mre,

[Korpus](#)

[Autorki](#)

[Publikacje](#)

[Linki](#)

[Podziękowania](#)

[XXX](#)

/kra[ww][s]/ Szukaj

	Oryginał	Tłumaczenie
...	1 AIESEC	✗ AIESEC
...	7 czygier_op.xml	✗ czygier_tu.xml
...	9 downar_op.xml	✗ downar_tu.xml
...	5 Drul	✗ Drul
...	11 eurowybor_ou.xml	✗ eurowybor_tp.xml
...	13 finberg_ou.xml	✗ finberg_tp.xml
...	14 hrycak_ou.xml	✗ hrycak_tp.xml
...	16 jaworski_op.xml	✗ jaworski_tu.xml
...	29 kasjanow_ou.xml	✗ kasjanow_tp.xml
...	30 krawczyk_op.xml	✗ krawczyk_tu.xml
...	32 krym_ou.xml	✗ krym_tp.xml
...	27 kubisz_op.xml	✗ kubisz_tu.xml
...	3 List Burakowski	✗ List Burakowski
...	18 natoukr_ou.xml	✗ natoukr_tp.xml
...	20 nowystaw_op.xml	✗ nowystaw_tu.xml
...	34 obolonczyk_ou.xml	✗ obolonczyk_tp.xml
...	36 oniszczyk_ou.xml	✗ oniszczyk_tp.xml
...	37 pashawer_ou.xml	✗ pashawer_tp.xml
...	23 popowycz_ou.xml	✗ popowycz_tp.xml
...	39 riabczyk_ou.xml	✗ riabczyk_tp.xml
...	41 tozsamosc_op.xml	✗ tozsamosc_tu.xml
...	43 wolyn_ou.xml	✗ wolyn_tp.xml
...	45 wozniak_ou.xml	✗ wozniak_tp.xml
...	47 zacharow_ou.xml	✗ zacharow_tp.xml

© tuurma '05

Gotowe Lokalny intranet

Start C:\WINDOWS\s... Polsko-ukrai... Polsko-ukrai... Microsoft Power... 11:36

Polsko-ukraiński korpus równoległy

nie przecze mebel piękny na pewno *Не побрались, не знали,*
А ои розпинають
 a jeszcze pies rasowy to już cywilizacja *Воду за подушине, а сина кують,*
Літа ти могоді
 Kolorowi, biali, czarni *Мені, єдині, дитини*
 Chodźcie zwłaszcza wy, ludko *... А вноє під тинюм*
 Przez na oścież bramy *Опухла дитина - голоднес мре,*



Korpus

Autorki

Publikacje

Linki

Podziękowania

XXX

Орест Друль

Украина vs українськості

Все дуже просто.

Простота вищого рівня – цивілізаційного масштабу: виклик глобалізації породжує посилення регіональної ідентифікації; посилення ідентифікації експлікується сепаристичними (у сенсі шлюбної сепарації) тенденціями, рухами [1], розмовами, статтями тощо; іншими словами – Європа додифундувала таки до Галичини у формі гасла "Європа регіонів".

Простота інша – на державну скалю: тенденція поглиблення не те що перманентного, але консеквентного невдоволення соціально-економічною ситуацією на тлі державної інституалізації України компрометує ідентифікацію з цією аутсайдерною країною; іншими словами – одержавлення тендітного поняття "Україна" як ідеалу всього, чого тільки можна уявити доброго: від мови (другої у світі за мітичним, але авторитетним французьким конкурсом) через кохання (ті ж Бітлз співали про ukrainian girls – послухайте "Back in USSR" з т.зв. "White Album") до вершин демократії (тих, хто ще пам'ятає неспадкувану її конотацію, відсилаю до прецеденту конституції Орлика) – призвело, прогнозовано, зрештою, до настільки повної трансформації семантики слова, що утотожнюватися з новим його значенням стало вищою дзен-практикою. Зрозуміло, що вищі ступені цієї практики (як от, що рух – це застиглість) досяжні лише немоногой щепті – тож питання вибору іншої ідентичності для неочудненої решти конгруентне kwestії часу.

Простота ще інакшіша – регіональна, прецінь галицька: Боже! а ми ж

Orest Drul

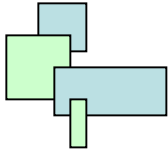
Ukraina vs. ukraińskość

To bardzo proste.

Prostota wyższa to prostota na skalę cywilizacyjną: wyzwania globalizacji rodzą nasilenie patriotyzmu regionalnego; nasilenie regionalnego patriotyzmu tłumaczy separatystyczne (w znaczeniu separacji małżeńskiej) tendencje, ruchy [1], rozmowy, artykuły itd.; innymi słowy, Europa dodyfundowała ostatecznie do Galicji pod postacią hasła „Europa regionów”.

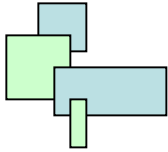
Inny rodzaj prostoty to prostota w skali państwowej: tendencja pogłębiania się nie dość że permanentnego, ale też konsekwentnego niezadowolenia z sytuacji społecznej i ekonomicznej na tle państwowej instytucjonalizacji Ukrainy kompromituje poczucie związku z tym krajem-outsiderem; innymi słowy – upaństwowienie delikatnego pojęcia „Ukraina” jako iдеалу wszelkiego dobra, jakie tylko można sobie wyobrazić: od języka (drugiego na świecie według mitycznego, ale szanowanego francuskiego rankingu) przez miłość (nawet Beatlesi śpiewali o ukrainian girls – proszę posłuchać „Back in USSR” z tzw. „White Album”) do szczytów demokracji (pamiętających jeszcze pierwotne, czyste konotacje tego słowa odsyłam do precedensu konstytucji Filipa Orlika) – doprowadziło, zgodnie zresztą z przewidywaniami, do tak całkowitej przemiany semantyki słowa, że utożsamianie się z jego nowym znaczeniem to wyższa szkoła zen. Wiadomo, że wyższe arkana tej szkoły (jak np. to, że ruch jest bezruchem) dostępne są tylko nielicznej garstce – więc problem wyboru innej tożsamości dla niedouczzonej reszty jest kongruentny z kwestią czasu.

Jeszcze inna prostota to prostota regionalna, wszakże galijska: Boże! a



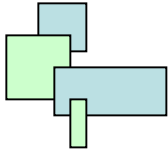
Korpus

Pozwala na natychmiastowe korzystanie z
bazy ekwiwalentów tłumaczeniowych –
wystarczy dobry równoległy konkordancer



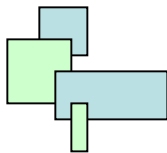
Metodyka tworzenia słownika

- Utworzenie **makrostruktury słownika** na podstawie **listy frekwencyjnej** słów występujących w korpusie oraz list haseł z popularnych słowników (w przypadku ograniczonego rozmiaru korpusu)
- Ustalenie **znaczeń haseł** na podstawie analizy **konkordancji** i istniejącej mikrostruktury, zgodnie ze spójną klasyfikacją
- Ustalenie **ekwiwalentów przekładowych** na podstawie analizy **konkordancji równoległych**, a ekwiwalentów opisowych na podstawie definicji



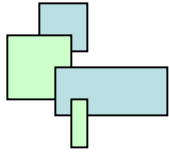
Wirtualne laboratorium

- Internetowe środowisko pracy umożliwiające dostęp zdalny i współpracę dużej grupie redaktorów
- Zasady pracy zbliżone do **Wikipedii**



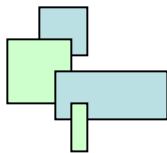
Wirtualne laboratorium 2

- Klasyfikacja lingwistyczna
- Ekwiwalenty tłumaczeniowe
- Ekwiwalenty opisowe
- Klasyfikacja dziedzinowa
- Automatyzacja niektórych zadań
(wyszukiwanie walencja, związki frazeologiczne itd.)
- ...



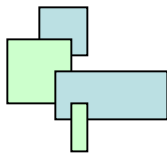
Natalia Kotsyba

- Makrostruktura słownika, objętość, sposoby doboru haseł
- Mikrostruktura słownika, problem ekwiwalencji



Wiarygodność korpusów

- (Fillmore 1992) badanie mające na celu porównanie haseł słownikowych angielskiego wyrazu *risk* z danymi uzyskanymi korpusu
- hasła z 10 słowników vs rezultat, uzyskany poprzez analizę przykładów z 25-milionowego korpusu
- jedynie dwa słowniki odnotowywały wszystkie trzy znaczenia wyrazu
- żaden z nich nie odnotowywał wszystkich istotnych kolokacji. (źródło - Wójtowicz, 170)



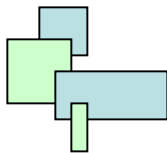
Wiarygodność korpusów

- **Podejść – podchodzić**

„mieć stosunek do czegoś” np. *podejść do problemu* 20% użyć w Korpusie IPI PAN, nie zanotowano w SJP Doroszewskiego, SJP PWN

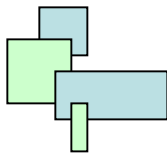
- **Kwiatek**

„nieoczekiwany wynik”, iron.
10% użyć, też nie zanotowane



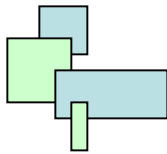
Integrowany materiał językowy

- korpus przedstawia układ syntagmatyczny
- słownik/gramatyka – układ paradygmatyczny
- korpusy + słowniki = uniwersalny bank danych językowych
- im bogatsza anotacja, tym prostsze jest automatyczne przejście między układami
- integracja danych syntagmatycznych i paradygmatycznych ułatwia dalszą automatyczną anotację



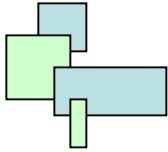
Makrostruktura

- objętość korpusu równoległego, nawet przy najbardziej optymistycznych prognozach nie wystarcza, co oznacza potrzebę dodatkowych źródeł:
- korpusy jednojęzyczne: IPI PAN, ULIF
- słowniki jednojęzyczne: SUM, PWN
- zgoda autorów/ właścicieli praw autorskich



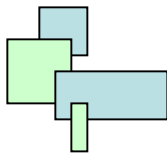
Mikrostruktura

- wielomodułowa, z możliwością wyświetlania informacji, których potrzebujemy w danym momencie
- informacje są przechowywane i pobierane z relacyjnej bazy danych



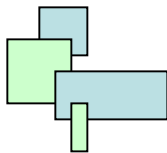
Mikrostruktura

- fonetyka
- morfologia
- **semantyka** (największe kłopoty)
- pragmatyka -||-
- ilustracja (reprezentacja z korpusu)



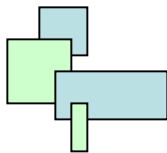
Mikrostruktura - semantyka

- Piotrowski (1994: 161): „[...] ekwiwalenty mają dwie funkcje: po pierwsze, mają one określić znaczenie wyrażenia wyjściowego, a po drugie, mają być odpowiednikami tłumaczeniowymi. A odpowiedniki tłumaczeniowe [...] powinno dać się podstawić za właściwe wyrażenie wyjściowe w określonym kontekście.”
- (cyt. za Wójtowicz, 2005)



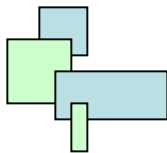
Mikrostruktura - semantyka

- według anglojęzycznej terminologii ze względu na sposób doboru ekwiwalentów istnieją słowniki: *encoding* (odpowiedniki tłumaczeniowe) i *decoding* (objaśnienia haseł)
- w ideale są identyczne (terminologia)
- różne potrzeby odbiorców (opisowe – dla zrozumienia tekstu w języku obcym, odp. tłum. – dla tworzenia tekstów)



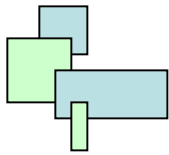
Mikrostruktura - semantyka

- konkordancje równoległe są najlepszym źródłem **odpowiedników tłumaczeniowych** (to powinno wystarczyć tłumaczom, szukającym odpowiedniego wyrazu, tzw. typologizacja kontekstów)
- praktyka leksykograficzna często działa na tych samych zasadach, podając synonimiczny rząd jako ekwiwalent, odbiorca musi sam ustalać w jakich kontekstach którego synonimu użyć
- takie ekwiwalenty są przeważnie **przybliżone**



Mikrostruktura - semantyka

- **objaśnienia** znaczeń są bardziej ścisłymi odpowiednikami, ale:
- dla osiągnięcia odpowiedniego poziomu dokładności objaśnienia mogą być tak rozbudowywane, że niepotrzebnie dublują definicje w języku wyjściowym
- mało nadają się przy tworzeniu tekstów w języku obcym
- kosztem objętości

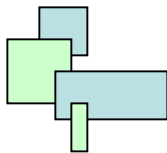


Mikrostruktura - semantyka

„Idealem, do którego w rekonstrukcji znaczeń języka naturalnego należy dążyć, jest posługiwanie się wyłącznie elementami absolutnie prostymi, przynależącymi do semantycznej bazy. To jednak na obecnym etapie zaawansowania teorii semantycznej nie jest możliwe...”

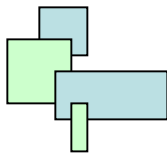
(Bogusławski & Danielewicz, 2005, s. 20)

- całkowita dekompozycja



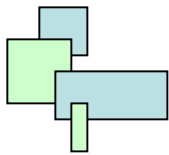
Mikrostruktura - semantyka

- połączenie obu rodzajów ekwiwalencji
- oprócz OT – zwięzły opis znaczenia przy pomocy kategoryzacji i **minimalnej** dekompozycji semantycznej (wystarczającej do rozróżnienia)
 - w ideale oparty na opisie struktury hasła w słowniku jednojęzycznym, inaczej redaktor słownika dwujęzycznego jest skazany na robienie potrójnej pracy
- dla słowników dwujęzycznych istotna jest kategoryzacja **względna**, która pokazuje różnice między wyrazami, podanymi jako OT



Mikrostruktura - semantyka

- ścisły, możliwie formalny opis ma być oparty na jednej spójnej klasyfikacji, która by działała jako uniwersalny filtr semantyczny
- taka nie istnieje

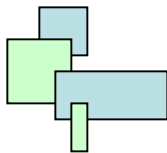


Mikrostruktura - klasyfikacje

- klasyfikacja „grubsza”, czasowniki –
Zołotowa, 1998

Zalety:

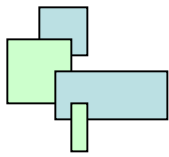
- pozwala uwzględnić **wszystkie** czasowniki
- konsekwentny podział znaczeń na gramatyczne (półpredykatywne) i leksykalne (*pełnopredykatywne)



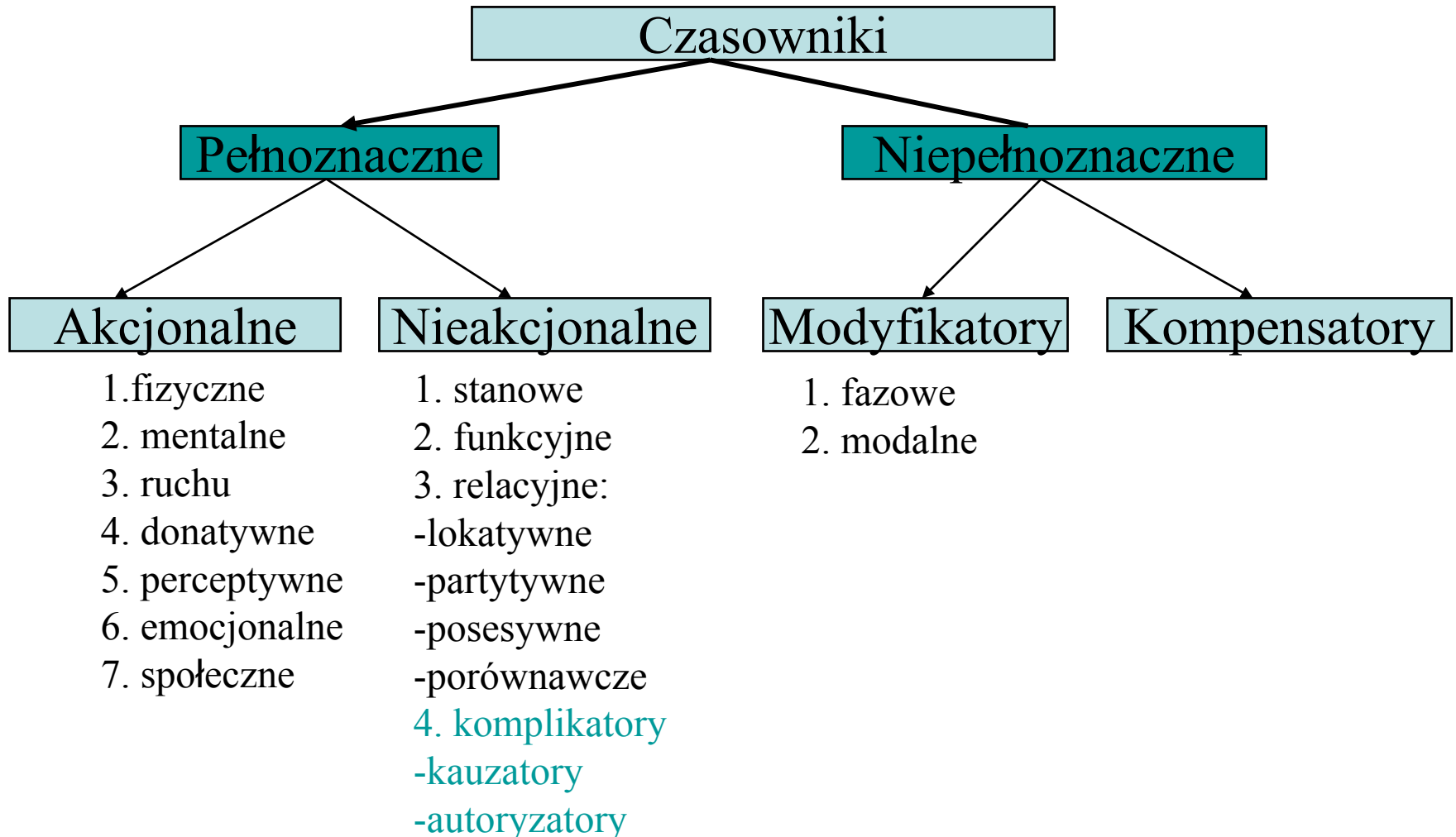
Mikrostruktura - klasyfikacje

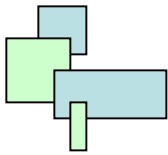
Słabe strony:

- tylko czasowniki
- holistyczne podejście do leksyki, brak dekompozycji semantycznej nie pozwala rozróżniać czasowników
*jednopredykatywnych od polipredykatywnych (np. wszystkie czasowniki, zawierające element kauzatywny są w jednej grupie bez względu na typ głównego predykatu)
- nie uwzględnia informacji stylistycznej, pragmatycznej, struktury predykatowo-argumentowej, itd.

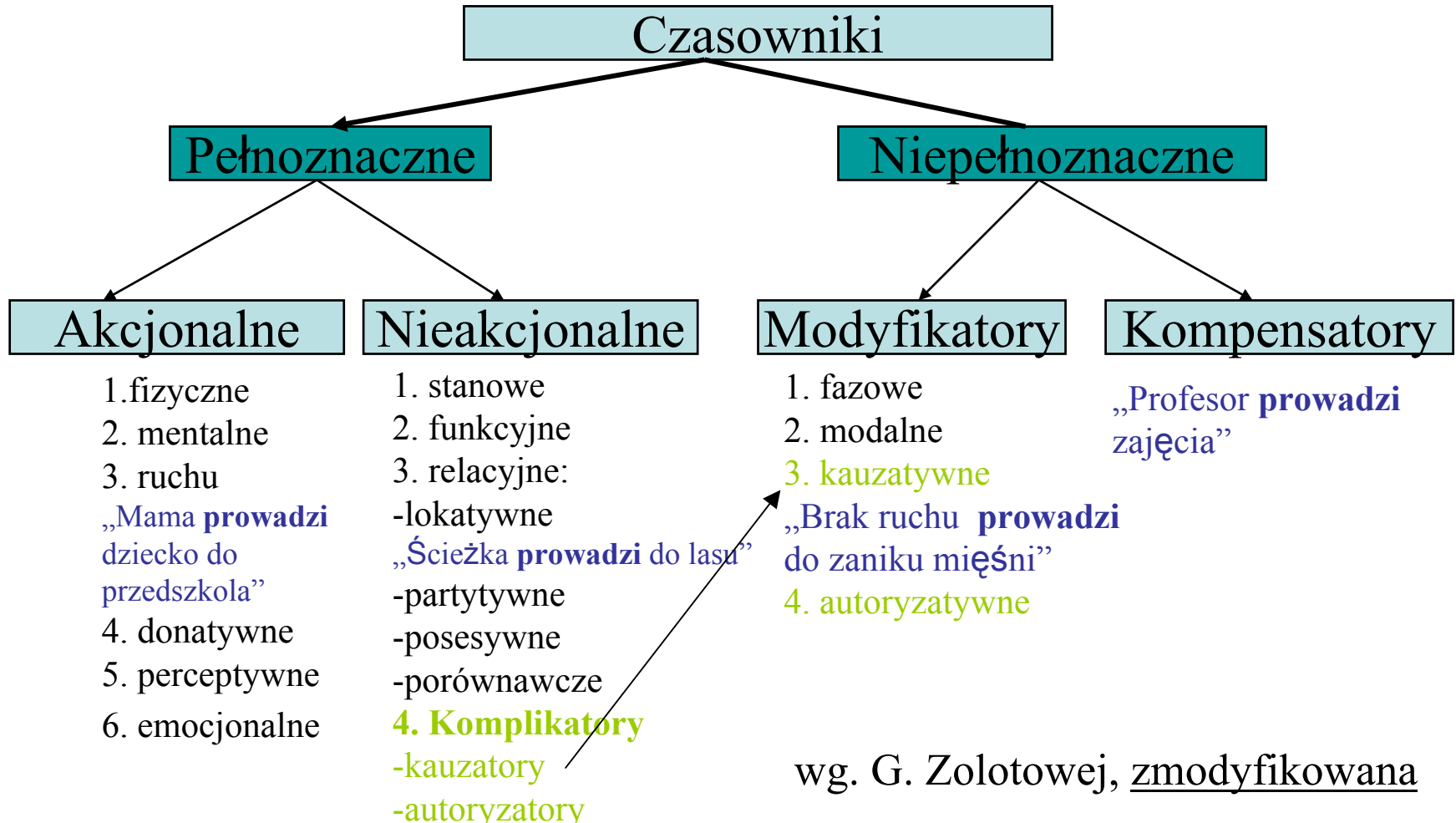


Klasyfikacja Złołotowej

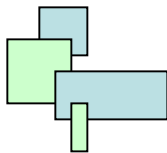




Klasyfikacja Zołotowej, zmodyfikowana

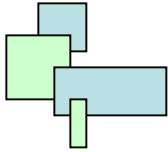


wg. G. Zołotowej, zmodyfikowana



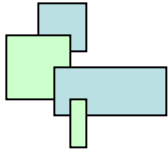
Przykłady czasowników **półpredykatywnych**

- **wyróżniać się** talentem = być utalentowanym (kompensator)
- **przewodzić** badania = badać (kompensator)
- **oceniać, rozpatrywać, uważać za** (autoryzatory)
- **zaciągać** kredyt, **zachodzić** w ciążę (modyfikator fazowy)



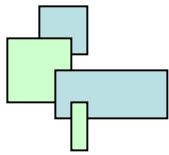
Przykłady czasowników **polipredykatywnych**

- **zaistnieć** = zacząć być (nieakc. relacyjny + modyfikator fazowy)
- **prowadzić** x = powodować, że x się porusza (akcjon. ruchu + kauzacja)



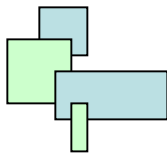
Przykład analizy hasła ze zgramatyzalizowanymi znaczeniami

- **prowadzić** (słownik PWN)
- 1. «wieść kogoś, coś do jakiegoś miejsca, do celu; przeprowadzać kogoś, kto nie zna drogi albo sam iść nie może, wskazywać komuś drogę» - **pełnoznaczny, akcjonalny ruchu + kauzacja (akcjonalny ruchu czy nieakcjonalny komplikator???) dk odprowadzić, zaprowadzić - вести**
- 2. «stawać się przyczyną czegoś; powodować, wywoływać coś, pociągać za sobą coś» - **pełnoznaczny, nieakcjonalny komplikator (kauzatywny) dk doprowadzić/ zaprowadzić - вести**
- 3. «kierować pojazdem mechanicznym, statkiem, końmi» =1 **brak dk (бути) за кермом, за штурвалом, верхи**
- 4. «kierować partnerem w tańcu» =1 **brak dk - вести**
- 5. «sprawować nad czymś nadzór, zarządzać czymś, zajmować się, trudnić się czymś; kierować czymś» - **niepełnoznaczny, kompensator dk poprowadzić, przeprowadzić проводити (заняття)**
- 6. «w połączeniu z rzeczownikiem będącym dopełnieniem oznacza: realizować, kontynuować to, co jest wyrażone w dopełnieniu» =5
- 7. «ciągnąć się w jakąś stronę, stanowić dojście lub przejście do czegoś» - **pełnoznaczny, nieakcjonalny, relacyjny, lokatywny doprowadzić**
- 8. «o zawodnikach, drużynach sportowych: być pierwszym w klasyfikacji, mieć przewagę nad przeciwnikiem; przodować» - **pełnoznaczny, nieakcjonalny, relacyjny, porównawczy**



Inne sposoby opisu leksyki

- połączenie filtrów „grubszego” i „cieńszego”
- np. teoria **generatywnego leksykonu**
Pustejovskiego, 1995
- na razie dość szkicowa, chociaż aktywnie rozwijana
- próba podejścia ponadkategorialnego – opisuje czasowniki, rzeczowniki i przymiotniki
- wysoki stopień formalizacji opisu



Generatywny leksykon Pustejovskiego

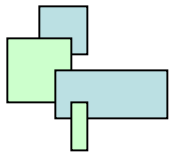
Znaczenie słowa opiera się na czterech generatywnych faktorach, tzw. *qualia*, które pokazują jak ludzie rozumieją przedmioty i relacje w świecie i dają minimalne objaśnienie zachowania lingwistycznego jednostek leksykalnych

FORMALNE: podstawowa kategoria, która wyróżnia przedmiot w większej grupie (~hiperonimia)

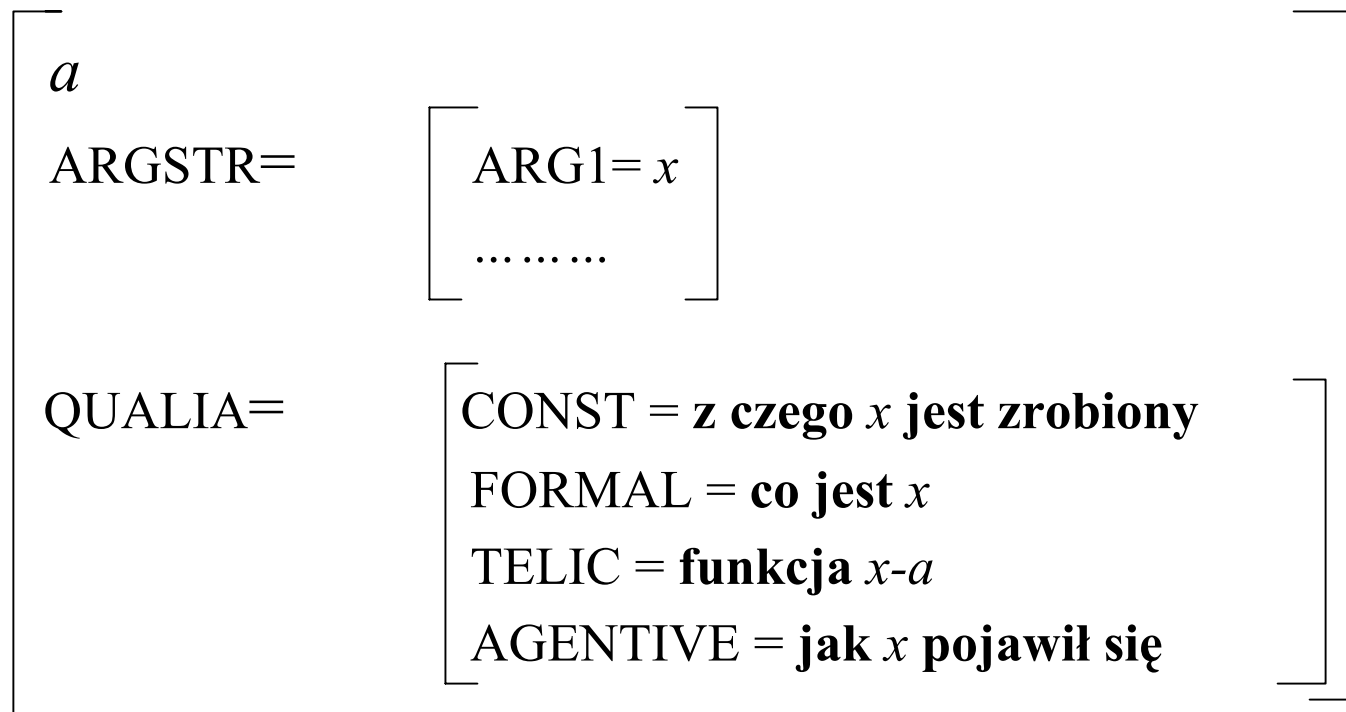
KONSTYTUTYWNE: relacje pomiędzy przedmiotem i jego częściami (~meronimia)

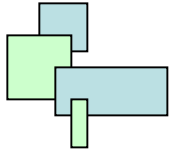
TELICZNE: funkcja / zastosowanie przedmiotu

AGENTYWNE: faktory, mające wpływ na powstanie przedmiotu



Opis schematyczny jednostki leksykalnej *a* wg Pustejovskiego



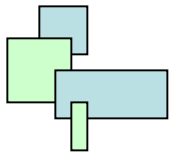


Schematyczna prezentacja - *food*

$$\left[\begin{array}{l} \mathbf{food} \\ \text{ARGSTR} = [\text{ARG } 1 = x:\text{substance}] \\ \text{QUALIA} = \left[\begin{array}{l} \text{FORMAL} = x \\ \text{TELIC} = \text{eat}(e, y, x) \end{array} \right] \end{array} \right]$$

Interpretacja rzeczownika *food* daje następujące wyrażenie logiczne:

$$\lambda x[\text{substance}_F(x) \wedge \lambda y \lambda e [\text{eat}_T(e, y, x)]]$$

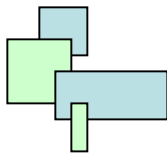


Schematyczna prezentacja - *book*

book					
ARGSTR =	<table><tr><td>ARG1 = y:information</td></tr><tr><td>ARG2 = x:phys_obj</td></tr></table>	ARG1 = y:information	ARG2 = x:phys_obj		
ARG1 = y:information					
ARG2 = x:phys_obj					
QUALIA =	<table><tr><td>information•phys_obj</td></tr><tr><td>FORM = hold(x,y)</td></tr><tr><td>TELIC = read(e,w,x•y)</td></tr><tr><td>AGENT = write(e,v,x•y)</td></tr></table>	information•phys_obj	FORM = hold(x,y)	TELIC = read(e,w,x•y)	AGENT = write(e,v,x•y)
information•phys_obj					
FORM = hold(x,y)					
TELIC = read(e,w,x•y)					
AGENT = write(e,v,x•y)					

Wyrażenie logiczne:

$$\lambda x \bullet y \exists e \bullet \exists v [book(x:physobj \bullet y:info) \wedge hold(x,y) \wedge \lambda w \lambda e [read(e,w,x \bullet y) \wedge [write(e;v,x \bullet y)]]]$$

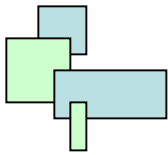


Różna walencja ekwiwalentów

- szczotkować – чистити (щіткою)

szczotkować	[arg1= x:agent]	]
	[arg2 = y:object]	

чистити	[arg1= x:agent]	]
	[arg2 = y:object]	
	[arg3 (opt) =	
z:instrument]		



Bardziej subtelne rozbieżności w znaczeniach ekwiwalentów

- **zabrać – забрати**

zabrać kogoś do miasta/ klubu; dziecko do cyrku/kina

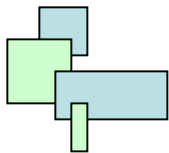
**забрати когось до міста/ клубу; дитину до цирку/ в кіно*

Próbka Korpusu IPI PAN - [base="zabrać"][orth="do"]:

Wkrótce na miejscu zdarzenia zjawiała się policja, a pijaka **zabrano do** izby wytrzeźwień. **akcjonalny, ruchu + kausacja** **завезли**

... rozdzielono mnie z bratem, którego **zabrano do** Brzezinki. **забрали**

Mechanika **zabrano do** szpitala po zawinięciu do Los Angeles. Wymieniono też część załogi. **завезли**



Schematy struktury znaczeniowej *zabrać – забрати*

- **wspólny** dla pol. i ukr.

arg1 = x:anim ◦ loc1

arg2 = y:anim ◦ loc2

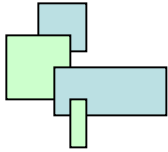
y CAUSE x (loc1 → loc2)

- **tylko w jęz. pol.**

arg1 = x:anim ◦ loc1

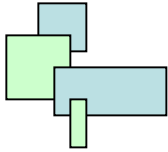
arg2 = y:anim ◦ loc1

y CAUSE x^y (loc1 → loc2)



Inne aspekty opisu

- informacje stylistyczne i pragmatyczne
- przykład „Sond Słownikowych”
A. Bogusławskiego i współautorów
- słaba baza teoretyczna



Bibliografia

- Fillmore, J. 1992. *“Corpus Linguistics” or “Computer-aided armchair linguistics”*. W: Directions in Corpus Linguistics. Proceedings of Nobel Symposium....
- Bogusławski, A., Danielewiczowa M. 2005 *Verba polona abscondita. Sonda słownikowa III*. Warszawa.
- Piotrowski, T. 1994. *Z zagadnień leksykografii*. Warszawa: Wydawnictwo Naukowe PWN.
- Pustejovsky, J. 1995. *The generative Lexicon*. Cambridge, MA: MIT Press.
- Wójtowicz, B. 2005. *Ogólnolingwistyczne i leksykograficzne podstawy organizacji słownika suahilijsko-polskiego*, nieopublikowana rozprawa doktorska.
- Золотова Г., Онипенко Н., Сидорова М. 1998 *Коммуникативная грамматика русского языка*, Москва, ИРЯ - РАН