# Polsko-Ukraiński Korpus

Autorzy:

Natalia Kotsyba,

Instytut Slawistyki PAN

Magdalena Turska,

Uniwersytet Warszawski

# Czym jest korpus?

- Zbiór tekstów albo zapisanych wypowiedzi, wykorzystywany dla celów analizy lingwistycznej

(*The American Heritage® Dictionary of the English Language*)

# Czym jest korpus równoległy?

- Zbiór par tekstów (eng. *bitexts*) w dwóch językach

Запропонована модель виглядає доволі переконливо...

Proponowany model wygląda dość przekonująco...

# Czemu służą korpusy równoległe

- Baza danych odpowiedników słów i wyrażeń oraz ich kontekstów dla tłumaczy
- Baza danych dla konfrontatywnej analizy lingwistycznej
- Punkt wyjścia do konstrukcji wiarygodnych słowników dwujęzycznych

# Jak tworzymy korpusy?

- Dobieramy obszerny zbiór tekstów reprezentujących różne style językowe

(np. artykuł, notatka prasowa, korespondencja, dokumentacja techniczna, literatura piękna)

- Znakujemy „surowiec” za pomocą odpowiednich narzędzi

(lematyzator, analizator morfologiczny, aligner)

# Narzędzia do znakowania

- Analizator morfologiczny

  wyznacza formę lub formy (w przypadku homonimii) podstawowe słowa oraz jego charakterystyki gramatyczne na podstawie bazy słownikowej oraz reguł gramatycznych

- **miał**

  1. „mieć”, czasownik, 3os lp, cz. przeszły, tr. orzekający...

  2. „miał”, rzeczownik, mianownik lp. r.męski...

# Istniejące analizatory j. polskiego

- Marcin Woliński et al.,  IPI PAN
- Krzysztof Szafran, MIM UW
- M. Gajęcki et al. , AGH

# Istniejące analizatory j. ukr

- ULIF (Ukraiński Fundusz Lingwistyczno-Informacyjny) istnieje słownik fleksyjny

# Lematyzator

- dokonuje analizy morfologicznej w oparciu o heurystyki
- jest przydatny w przypadku braku analizatora morfologicznego, wyniki wymagają weryfikacji

# Aligner

- służy do „wyrównywania” tekstów równoległych, przyporządkowując nawzajem odpowiadające sobie ich fragmenty
- typowymi fragmentami są zdania lub akapity
- *Problem: jak wyróżnić i dopasować zdania?*

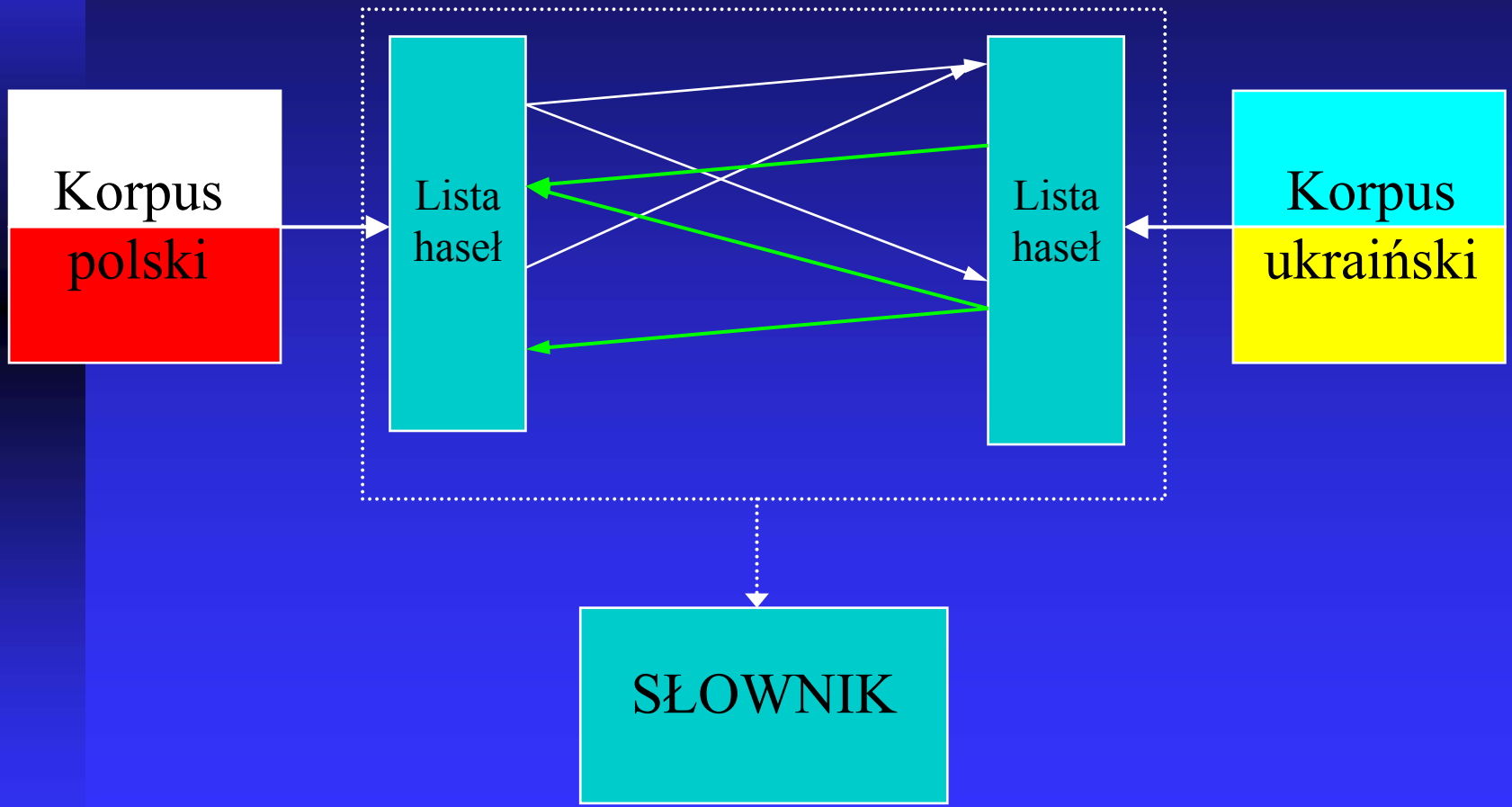
# Zastosowania korpusu (powtórka):

- Narzędzie pracy tłumaczy
- Podstawa do tworzenia słowników
- Materiał do analizy lingwistycznej
- ...

# Słownik dwujęzyczny:

- Analiza morfologiczna i wyznaczenie listy form podstawowych (haseł)
- Redakcja haseł (jako zbioru znaczeń) oraz przyporządkowanie sobie znaczeń pomiędzy językami na podstawie analizy konkordancji
- Selekcja haseł (opcjonalnie) pod kątem konkretnej edycji (np. ograniczenie objętości, wybór dziedziny)

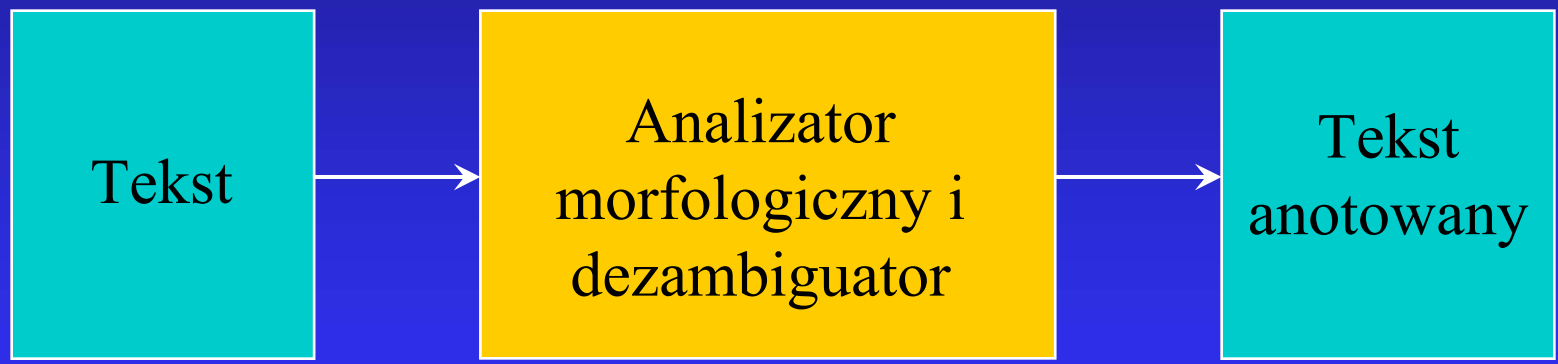
# Proces tworzenia słownika



# Analiza morfologiczna:

- Każdy wyraz w korpusie sprowadzamy do jego formy podstawowej
- Lista form podstawowych jest listą wyjściową haseł słownika

# Przebieg analizy morfologicznej

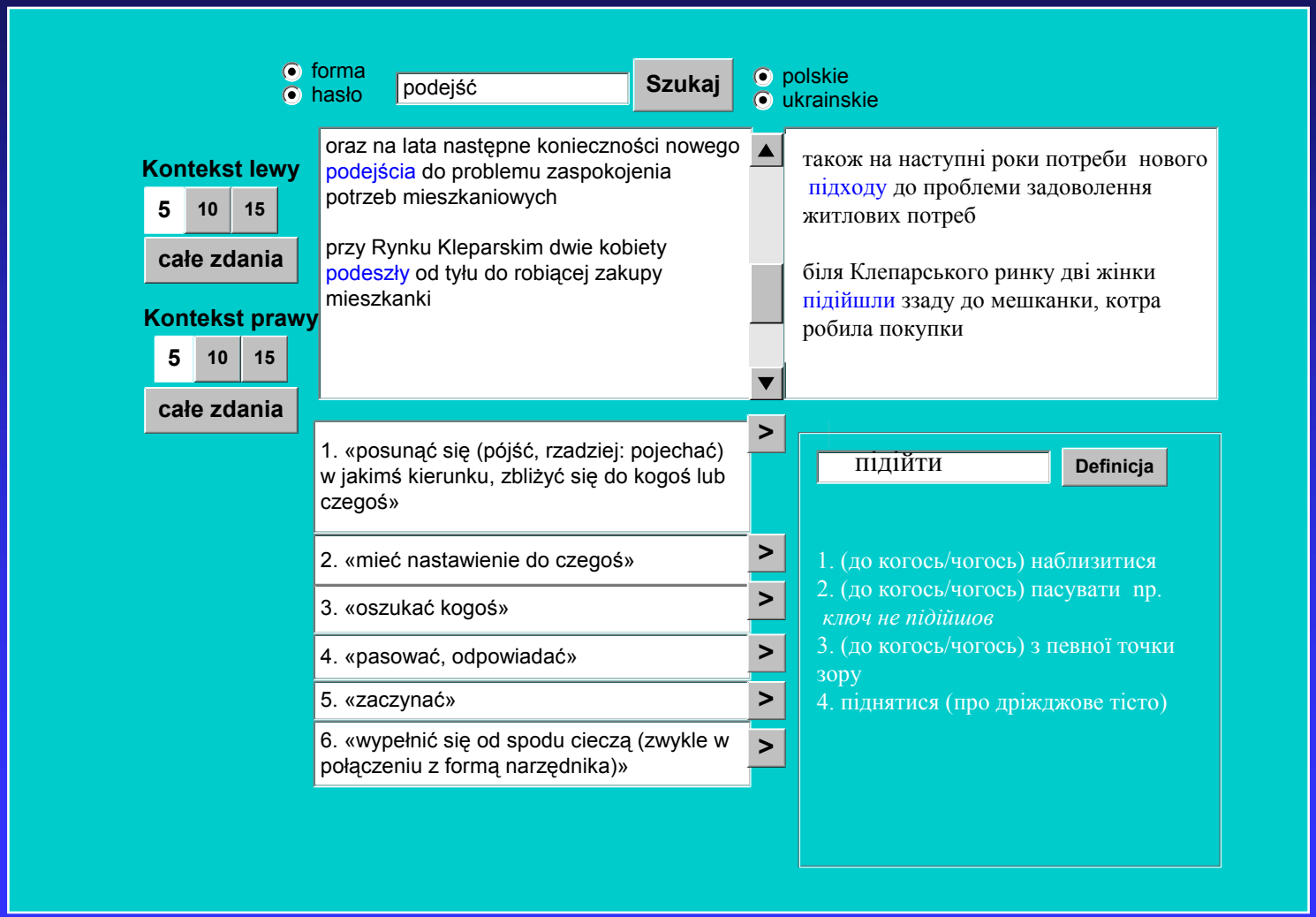


# Realizacja analizy morfologicznej

- Modułowa budowa projektu pozwala na „wpięcie” dowolnego analizatora

  (pod warunkiem zastosowania wrappera dostosowującego formaty wejścia i wyjścia)

# Interfejs narzędzia redakcyjnego



# Edycja hasła – punkt wyjścia

- Opis z istniejącego słownika jednojęzycznego (dla j.polskiego np. słownik rzędu Doroszewskiego lub PWN, dla j. ukraińskiego – słownik pod red. Biłodida)
- Zbiór konkordancji z korpusu równoległego

# Zalety konkordancji równoległych

- Korzystamy z twórczych rozwiązań tłumaczy dot. opisu realiów językowych, frazeologizmów, terminologii:
- Neologizmy czy okazjonalizmy?
- Opisy/ kolokacje (stąd m.in. asymetria słowników dwujęzycznych)
- Synonimy

# Pułapki konkordancji

- Tłumaczy się nie odrębne wyrazy, lecz sytuacje

Przykłady z nagłówków gazet:

- *Czesław Miłosz nie żyje (* ...jest martwy)*
- *Помер Чеслав Мілош (*... не живе, ...є мертвий),* szyk – sposób na wyrażenie znaczenia Perfektu
- *Czeslaw Milosz is dead (*...does not live)*

# Pułapki konkordancji

- Tłumacz ma wolną rękę co do zamiany hiperonimiczno-hiponimicznej – uogólnia lub uszczegółnia wyrazy w zależności od realiów językowych i według własnego uznania
- np. *goździk – kwiatek*

# Słownik objaśniający – idealny

- Opisuje wszystkie występujące i tylko występujące znaczenia
- Znaczenia są podawane w kolejności, zgodnej z częstotliwością używania

# Słownik objaśniający – w rzeczywistości

- Opisywane znaczenia są często przestarzałe (wyodrębnione i egzemplifikowane na podstawie utworów klasycznych)
- Znaczenia są niepotrzebnie powielane
- Znaczenia są niekonsekwentnie pogrupowane
- Kolejność opisywanych znaczeń nie zawsze odzwierciedla częstotliwość ich występowania w mowie

# Słownik objaśniający – w rzeczywistości c.d.

- To utrudnia proces kojarzenia znaczeń haseł, które są podobnie używane w spokrewnionych językach, ale w różny sposób opisane w słownikach objaśniających + faux amis
- *umowa – угода* ‘agreement’
- *warunek – умова* ‘condition’

# Edycja hasła – ustalenie znaczeń

- Wystąpienia hasła w konkordancjach przyporządkowujemy do kategorii semantycznych
- To pozwala w konsekwentny sposób rozróżniać odrębne znaczenia hasła
- Ostateczna decyzja o dołączeniu bądź usunięciu znaczenia należy do redaktora

# Edycja hasła – kojarzenie znaczeń

- Przyporządkowujemy sobie znaczenia pomiędzy dwoma (lub więcej) językami, na podstawie równoległej analizy konkordancji w obu językach, istniejących definicji w słownikach jednojęzycznych oraz, ostatecznie, arbitralnej decyzji redaktora

# Przykładowe hasło polskie

- PODEJŚĆ - PODCHODZIĆ (za słownikiem PWN)
- **1.** «posunąć się (pójść, rzadziej: pojechać) w jakimś kierunku, zbliżyć się do kogoś lub czegoś»
- **2.** «posunąć się pod górę; wspiąć się»
- **3.** zwykle *dk* «postąpić wobec kogoś podstępnie, chytrze, zdradziecko; oszukać kogoś»
- **4.** częściej *ndk* «zbliżać się do kogoś lub czegoś ostrożnie, ukradkiem, zwykle w celu dokonania napaści lub podpatrzenia; tropić»
- **5.** «wypełnić się od spodu cieczą (zwykle w połączeniu z formą narzędnika)»

# Przykładowe hasło - najbliższy odpowiednik ukraiński

- **ПІДІЙТИ** док. - **ПІДХОДИТИ** недок. (za słownikiem Biłodida)

1. - ідучи, наближатися до кого-, чого-небудь;

   - наближатися підїжджаючи, підпливаючи, підлітаючи і т.ін. до кого-, чого-небудь;

   - прибувати куди-небудь;
2. - приступати до чого-небудь, братися за яку-небудь справу;

   - виявляти своє ставлення до чого-небудь, оцінюючи;
3. - уміти привернути, прихилити кого-небудь до себе, завоювати довір'я

   - звернутися до кого-небудь з проханням, пропозицією, вимогою і т.ін
4. наближатися, наставати (про час, події, явища і т.ін.)
5. розміщуватися близько чого-небудь, бути в безпосередньому сусідстві з чимсь, межувати з ним

# Przykładowe hasło - najbliższy odpowiednik ukraiński c.d.

6. - бути придатним, прийнятним, відповідаючи яким-небудь вимогам
   - бути відповідним
   - личити
   - пристосовуватися, підроблятися
7. переміщатися, підніматися догори
8. ідучи, пройти яку-небудь відстань
9. збільшуючись в об’ємі, підійматися (про тісто)
10. насичуватись чим-небудь *Сніг підійшов водою.*

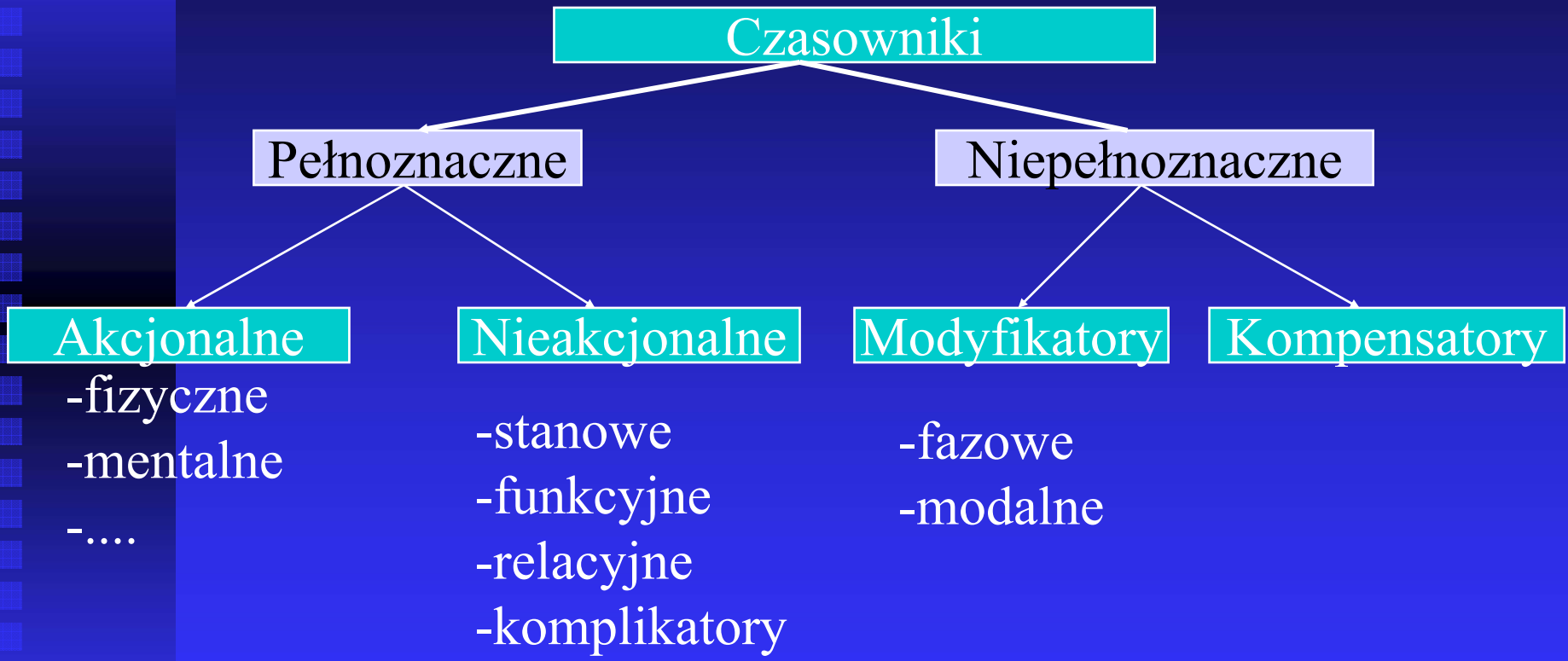
# Analiza konkordancji

- Zapytanie do korpusu IPI PAN **[base=”podejść”] meta created>1960** zwróciło 244 próbki tekstów.
- Zdecydowana większość użyć (ok. 80%) była w znaczeniu 1; w najbliższej dystrybucji wskazywano docelowe miejsce ruchu „podejść do + nazwa przedmiotowa” (np. *do furtki, do koszyka, do nas; bliżej* = do mówiącego).
- Podobnie z zapytaniem **[base=”podchodzić”] meta created>1960** - 203 próbki, proporcje te same

# Analiza konkordancji 2

- 1. Grzegorz Kaliciak: - Mieliśmy grać to co zawsze. I choć Pogoń bardzo poważnie **podeszła** do **meczu**, to my jesteśmy lepiej przygotowani. Mieliśmy wprawdzie drobny kryzys po meczu w Parmą, ale to już przeszłość. => NOWE ZNACZENIE „mieć nastawienie" (nieakcjonalny, relacyjny) – ok. 20% użyć
- 2. W okolicy miasta Mechelen musieli awaryjnie lądować. **Podchodzący do lądowania** samolot nie uszedł uwagi belgijskich żołnierzy pełniących służbę na znajdującym się nieopodal posterunku granicznym. => NOWE ZNACZENIE „zaczynać" (niepełnoznaczny, modyfikator fazowy)
- 3. Z drugiej zaś strony Nathan, Michael, Shawn i Wanya z powodzeniem wcielają w czyn swoje indywidualne pomysły; na przykład **podchodzące** pod śpiew a cappella wokalizy. Innymi słowy: Boyz II Men nagrali sprytny, inteligentny, dość dobry album, który świadczy o tym <Dziennik Polski, X. 2000> => 2
- 4. Na 5 pytań odpowiedział gładko, zapewniwszy sobie gwarantowany 1000 zł. Niestety, pytanie za 2000 zł, dotyczące El Greco, "nie **podeszło** ". => NOWE ZNACZENIE, „pasować, odpowiadać" (nieakcjonalny, relacyjny, porównawczy)

# Klasyfikacja semantyczna




wg. G. Zolotowej

# Klasyfikacja - przykład

- **prowadzić**
- **1.** «wieść kogoś, coś do jakiegoś miejsca, do celu; przeprowadzać kogoś, kto nie zna drogi albo sam iść nie może, wskazywać komuś drogę» - **akcjonalny predykat ruchu (kauzacja przemieszczenia się)**
- **2.** «stawać się przyczyną czegoś; powodować, wywoływać coś, pociągać za sobą coś» - **nieakcjonalny kauzatywny predykat (komplikator)**
- **3.** «kierować pojazdem mechanicznym, statkiem, końmi» **=1**
- **4.** «kierować partnerem w tańcu» **=1**
- **5.** «sprawować nad czymś nadzór, zarządzać czymś, zajmować się, trudnić się czymś; kierować czymś» - **niepełnoznaczny, kompensator**
- **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» - **nieakcjonalny predykat lokalizujący**
- **8.** «o zawodnikach, drużynach sportowych: być pierwszym w klasyfikacji, mieć przewagę nad przeciwnikiem; przodować» - **nieakcjonalny predykat porównawczy**
- **9.** *muz.* «uwypuklać melodię w kompozycji albo w jej wykonaniu» - **akcjonalny predykat, kauzacja relacji** , t. zn. „robiąc coś, sprawiać, że P(x)”
- **10.** *ogr.* «podpierać, przycinać, wycinać itp. gałęzie, pędy lub korzenie rośliny w celu nadania roślinie określonego kształtu, określonej formy»

# Różne zachowanie składniowe

- Mama prowadzi dziecko do przedszkola – Dziecko jest prowadzone przez mamę – Prowadzenie dziecka przez mamę – ?Mama zaczyna prowadzić dziecko do przedszkola – Mama powoduje, że dziecko idzie...

**akcjonalny predykat ruchu + kauzacja ruchu**

- Tata prowadzi samochód – Prowadzenie przez tatę samochodu – ?Tata zaczyna prowadzić samochód – Tata powoduje, że samochód się porusza

**akcjonalny predykat ruchu + kauzacja ruchu; nieakcjonalny stanowy**

- Profesor prowadzi zajęcia – Zajęcia są prowadzone przez profesora – Prowadzenie zajęć/ profesora – Profesor zaczyna/chce prowadzić zajęcia – *Profesor powoduje, że...

**niepełnoznaczny kompensator**

- Ścieżka prowadzi do lasu – *Prowadzenie przez ścieżkę do lasu – *Ścieżka zaczyna prowadzić do lasu – * Ścieżka powoduje...

**nieakcjonalny predykat lokalizujący**

(są też ograniczenia na użycie form imperatywnych, imiesłowu przysłówkowego „*prowadząc”, 1 i 2 osoby, formy dokonanej)

# Przykładowe hasło po analizie

PODEJŚĆ - PODCHODZIĆ

(za słownikiem PWN)

**1.** «posunąć się (pójść, rzadziej: pojechać) w jakimś kierunku, zbliżyć się do kogoś lub czegoś»

**2.** «posunąć się pod górę; wspiąć się»

**3.** zwykle *dk* «postąpić wobec kogoś podstępnie, chytrze, zdradziecko; oszukać kogoś»

**4.** częściej *ndk* «zbliżać się do kogoś lub czegoś ostrożnie, ukradkiem, zwykle w celu dokonania napaści lub podpatrzenia; tropić»

**5.** «wypełnić się od spodu cieczą (zwykle w połączeniu z formą narzędnika)»

- PODEJŚĆ - PODCHODZIĆ
  (analiza konkordancji i klasyfikacja semantyczna)

- **1.** «posunąć się (pójść, rzadziej: pojechać) w jakimś kierunku, zbliżyć się do kogoś lub czegoś» *akcjonalny ruchu*
- **2.** «mieć nastawienie do czegoś» *nieakcjonalny relacyjny*
- **3.** «oszukać kogoś» *złożony predykat akcjonalny mentalny + kauzacja*
- **4.** «pasować, odpowiadać» *nieakcjonalny, relacyjny*
- **5.** «zaczynać» *niepełnoznaczny modyfikator fazowy*
- **6.** «wypełnić się od spodu cieczą (zwykle w połączeniu z formą narzędnika)» *nieakcjonalny stanowy*

# Przykładowe hasło po analizie

- **ПІДІЙТИ** док. - **ПІДХОДИТИ** недок. (za słownikiem Biłodida)

1. - ідучи, наближатися до кого-, чого-небудь;
   - наближатися підїжджаючи, підпливаючи, підлітаючи і т.ін. до кого-, чого-небудь;
   - прибувати куди-небудь;
2. - приступати до чого-небудь, братися за яку-небудь справу;
   - виявляти своє ставлення до чого-небудь, оцінюючи;
3. - уміти привернути, прихилити кого-небудь до себе, завоювати довір'я
   - звернутися до кого-небудь з проханням, пропозицією, вимогою і т.ін
4. наближатися, наставати (про час, події, явища і т.ін.)

- **ПІДІЙТИ** док. – **ПІДХОДИТИ** недок. (za słownikiem Biłodida)

1. наближатися – **akcjonalny ruchu**
2. мати ставлення– **nieakcjonalny relacji**
3. пасувати – **nieakcjonalny rel.**
4. починати(ся) – **niepełnoznaczny modyfikator fazowy**
5. бути близько чого-небудь – **nieakcjonalny relacyjny lokalizujący**
6. збільшуючись в об'ємі, підійматися (про тісто) – **nieakcjonalny stanowy**
7. насичуватись чим-небудь (рідиною) - **nieakcjonalny stanowy**

# Przykładowe hasło po analizie 2

5. розміщуватися близько чого-небудь, бути в безпосередньому сусідстві з чимсь, межувати з ним
6. - бути придатним, прийнятним, відповідаючи яким-небудь вимогам

   - бути відповідним

   - личити

   - пристосовуватися, підроблятися
7. переміщатися, підніматися догори
8. ідучи, пройти яку-небудь відстань
9. збільшуючись в об’ємі, підійматися (про тісто)
10. насичуватись чим-небудь *Сніг підійшов водою.*

# Kojarzenie haseł pol. > ukr.

- **PODEJŚĆ - PODCHODZIĆ** (analiza konkordancji i klasyfikacja semantyczna)

- **1.** «posunąć się w jakimś kierunku, zbliżyć się do kogoś lub czegoś» → **підійти - підходити** 1
- **2.** «mieć nastawienie do czegoś» *nieakcjonalny, relacyjny* → **підійти - підходити** 2
- **3.** «oszukać kogoś» = *złożony predykat akcjonalny mentalny, z kauzacją propozycji* → **ошукати – ошукувати 1, надурити – надурювати 1**
- **4.** «pasować, odpowiadać» *nieakcjonalny, relacyjny* → **підійти - підходити** 3
- **5.** «zaczynać» *niepełnoznaczny modyfikator fazowy* → **починати**
- **6.** «wypełnić się od spodu cieczą (zwykle w połączeniu z formą narzędnika)» → **підійти - підходити** 5

# Kojarzenie haseł ukr. > pol.

- **ПІДІЙТИ** док. – **ПІДХОДИТИ** недок. (za słownikiem Biłodida)

1. наближатися *akcjonalny ruchu* → podchodzić 1
2. мати ставлення *nieakcjonalny relacji* → podchodzić 2
3. пасувати *nieakcjonalny rel.* → podchodzić 4
4. починати(ся) *niepełnoznaczny modyfikator fazowy* → zaczynać się
5. бути близько чого-небудь *nieakcjonalny relacyjny lokalizujący* → znajdować się blisko
6. збільшуючись в об’ємі, підійматися (про тісто) *nieakcjonalny stanowy* → rosnąć 7
7. насичуватись чим-небудь (рідиною) *nieakcjonalny stanowy* → podchodzić 6

# Realizacja

- Zespół redaktorski
- Podział pracy nad hasłami tematyczny

# Implementacja

- Elektroniczny korpus równoległy

- Relacyjna baza danych z możliwością zapisu w standardzie Unicode (np. MySQL)
- Zapis tekstów w języku XML (znakowanie rozdziałów, paragrafów, zdań; opis nagłówków)
- Dalsza anotacja również w XML (nieograniczone możliwości)

# Znakowanie semantyczne

- Umożliwia przedstawianie informacji na podstawie ontologii (WordNet, Laboratory of Applied Ontology), m. in. semantyczna anotacja czasu:

"Corriere della Sera" pisze w czwartek, że wciąż nie wiadomo, w których uroczystościach Wielkiego Tygodnia i Wielkanocy Jan Paweł II weźmie udział. (onet.pl, 10.III.2005)

<time date=2005-03-10 week=11 dayofweek=4 span=day>czwartek</time>

<time begin=2005-03-22 end=2005-03-28 week=17 span=week>Wielkiego Tygodnia</time>

# Implementacja

- Narzędzia edycyjne (Perl, Java, C, C++)
- Interfejs www korpusu i słownika (Perl, PHP)
- Słownik jako gotowy produkt (C, C++)