Językoznawcze studia doktoranckie IJP PAN 2015/2016

Leksykografia elektroniczna

Praktyki, metody, narzędzia

Krzysztof Nowak

Pracownia Łaciny Średniowiecznej

Instytut Języka Polskiego PAN

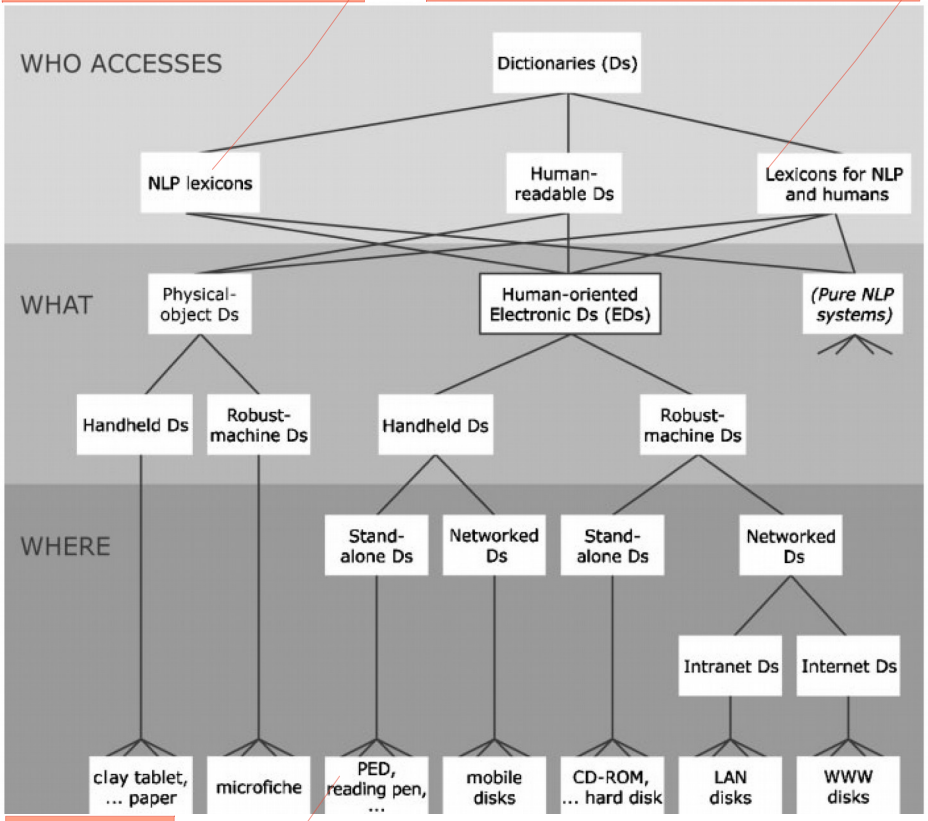
Plan wykładu

- **słowniki komputerowe**
  - definicje
  - typologia
  - cele
- **metody słownikowe w NLP**
  - rozpoznawanie mowy
  - spell-checkery
  - tagowanie morfo-syntaktyczne
- **WordNet**
  - Princeton WordNet
  - EuroWordNet
  - Słowosieć
  - Laboratorium: eksploracja Słowosieci
- **FrameNet**
- **słowniki walencyjne: Walenty**

Repetycja: słownik elektroniczny (de Schryver 2003) a słownik komputerowy

spellchecker

WordNet, Framenet



SMS

źródło: de Schryver 2003, 150.

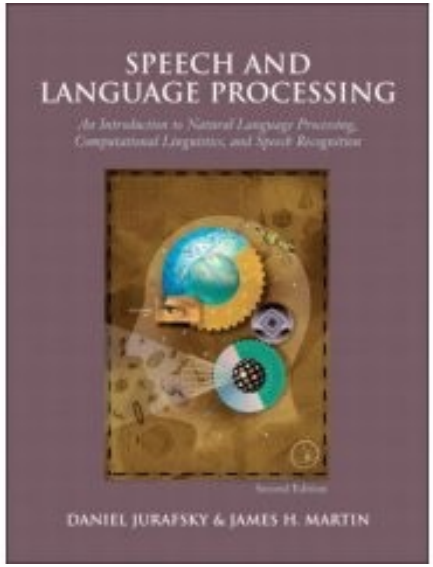
dostęp

- kto (człowiek / maszyna)
- do czego (medium)
- gdzie (przechowywanie)

definicje

- słownik elektroniczny: ED = *human-oriented electronic dictionary*
- słownik komputerowy, maszynowy: *NLP lexicon*, MRD (*machine-readable dictionary*)

Słowniki komputerowe w przetwarzaniu języka



Jurafsky D., Martin J. H., *Speech and Language Processing*, 2009

metody „słownikowe”
rzadko stosowane w izolacji, dziś częściej funkcjonują w obrębie większych infrastruktur wraz z metodami statystycznymi i np. uczenia maszynowego

przykłady użycia
- rozpoznawanie i synteza mowy
- korekta OCR, ortografii, gramatyki, stylu
- produkcja języka
- *programy do nauki słownictwa*
- tagowanie
- rozpoznawanie znaczenia
- *ekstrakcja informacji z tekstu*

Rozpoznawanie i synteza mowy

Automatic Speech Recognition
(Automatyczne Rozpoznawanie Mowy)

- co zawiera leksykon?
  - system otwarty czy zamknięty?
  - potrzeba szerokiej akwizycji (w tym: teksty specjalistyczne)
  - jednostki w rodzaju *hmmm* itp.

Text-To-Speech Conversion

(Zamiana Tekstu na Mowę)

- co zawiera leksykon?
  - informacje konieczne dla realizacji ciągu dźwiękowego
  - nie uwzględnia zróżnicowania geograficznego

- architektura leksykonu

Word	Pronunciation	Word	Pronunciation
cat	[kæt]	goose	[gus]
cats	[kæts]	geese	[gis]
pig	[pɪg]	hedgehog	[ˈhɛdʒ.hɔg]
pigs	[pɪgz]	hedgehogs	[ˈhɛdʒ.hɔgz]
fox	[fɑx]		
foxes	[ˈfɑk.sɪz]		

w bardziej zaawansowanych słownikach zapisywane w jednym z formalizmów

- problemy
  - brak w słowniku: propozycja w oparciu o model statystyczny
  - produktywność: nazwy własne, kompozycja, liczby

Spell-checker & co.

- **zadania**
  - korekta ortografii
  - sprawdzanie gramatyki
  - tezaurusy
  - dzielenie wyrazów
  - korekta stylu
- **właściwości leksykonu**
  - wyczerpujący
  - powinien zawierać nazwy własne (geografia, postaci itp.)
  - zawiera reguły analizy i syntezy (propozycje poprawek)
- **architektura**
  - rozbudowany słownik
  - lista lemmatów i afiksów
- **problemy**
  - kompozita w językach germańskich
  - reformy pisowni
  - neologizmy
- **inne informacje**
  - frekwencja
  - reforma
  - dialekt
  - rejestr
  - PoS, walencja, rodzaj, liczba

Laboratorium: budowa słownika dla Firefoksa

1. Otworzyć plik .xpi za pomocą programu archiwizacyjnego (https://goo.gl/iBm0px).

2. Otworzyć plik .dic

a. odnaleźć opracowywane przez siebie hasło

b. zapisać kod fleksyjny

3. Otworzyć plik .aff

a. odnaleźć reguły sufiksowania wedle kodu

Tagowanie części mowy (PoS tagging) 1

wejście

ciąg wyrazów w korpusie

wyjście

ciąg etykiet

Pawłow[1] udowodnił[2], że[3] ślinienie[4] …

[1]RZE_NOM_sg_NWŁ
[2]CZAS_PRZESZ_DOKON_3_sg
[3]SPÓJ

[4]RZE_NOM_sg ? RZE_ACC_sg ? RZE_VOC_sg ?

słownik

wątpliwość

reguły (*rule-based methods*)

metody stochastyczne (stochastic *methods*)

„zastrzeżenie” w kolokatorze PELCRA

Tagowanie części mowy (PoS tagging) 2

metody regułowe

ADVERBIAL-THAT RULE
Given input: "that"
if
(+1 A/ADV/QUANT); /* *if next word is adj, ad*
(+2 SENT-LIM); /* *and following which is*
(NOT -1 SVOC/A); /* *and the previous word is*
/* *'consider' which allows adjs as o*
then eliminate non-ADV tags
else eliminate ADV tag

reguły tagera ENGTWOL (~1995)

Word	POS	Additional POS features
smaller	ADJ	COMPARATIVE
entire	ADJ	ABSOLUTE ATTRIBUTIVE
fast	ADV	SUPERLATIVE
that	DET	CENTRAL DEMONSTRATIVE SG
all	DET	PREDETERMINER SG/PL QUANTIFIER
dog's	N	GENITIVE SG
furniture	N	NOMINATIVE SG NOINDEFDETERMINER
one-third	NUM	SG
she	PRON	PERSONAL FEMININE NOMINATIVE SG3
show	V	IMPERATIVE VFIN
show	V	PRESENT -SG3 VFIN
show	N	NOMINATIVE SG
shown	PCP2	SVOO SVO SV
occurred	PCP2	SV
occurred	V	PAST VFIN SV

Figure 8.8 Sample lexical entries from the ENGTWOL lexicon described in Voutilainen (1995) and Heikkilä (1995).

słownik tagera ENGTWOL (ok. 56 000 jednostek)

metody stochastyczne

wyboru spośród kandydatów zaproponowanych przez słownik dokonuje się na podstawie szacowania prawdopodobieństwa ich wystąpienia w danym kontekście

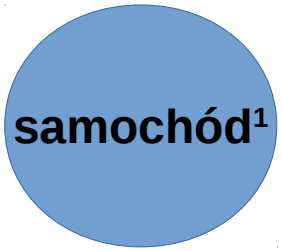
Princeton WordNet

Vossen 2007

semantyczna baza leksykalna

oparta o relacyjny model znaczenia

tradycyjny leksykon	WordNet
forma → znaczenia	pojęcia (≈ zbiory synonimów, tzw. *synsetów*) → formy
wóz	*napędzany silnikiem pojazd mechaniczny przeznaczony…*
1. samochód 2. w. konny (źródło: WSJP)	**wóz[1], auto[1], samochód[1]** (źródło: Słowosieć)



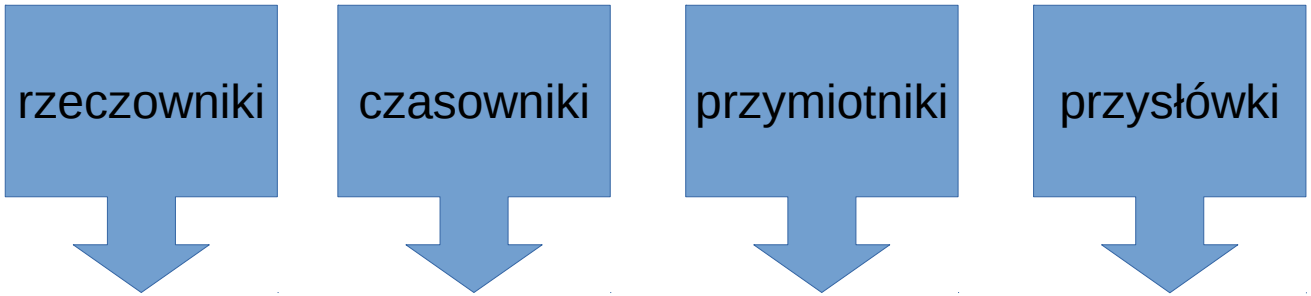
synonimy	hiperonimy	meronimy
auto, wóz, pojazd samochodowy	dwuślad, pojazd drogowy	silnik, deska rozdzielcza, kierownica, tłumik

(źródło: Słowosieć)

WordNet: struktura (1)

***synset* – zbiór synonimów powiązanych relacjami znaczeniowymi z innymi elementami sieci**

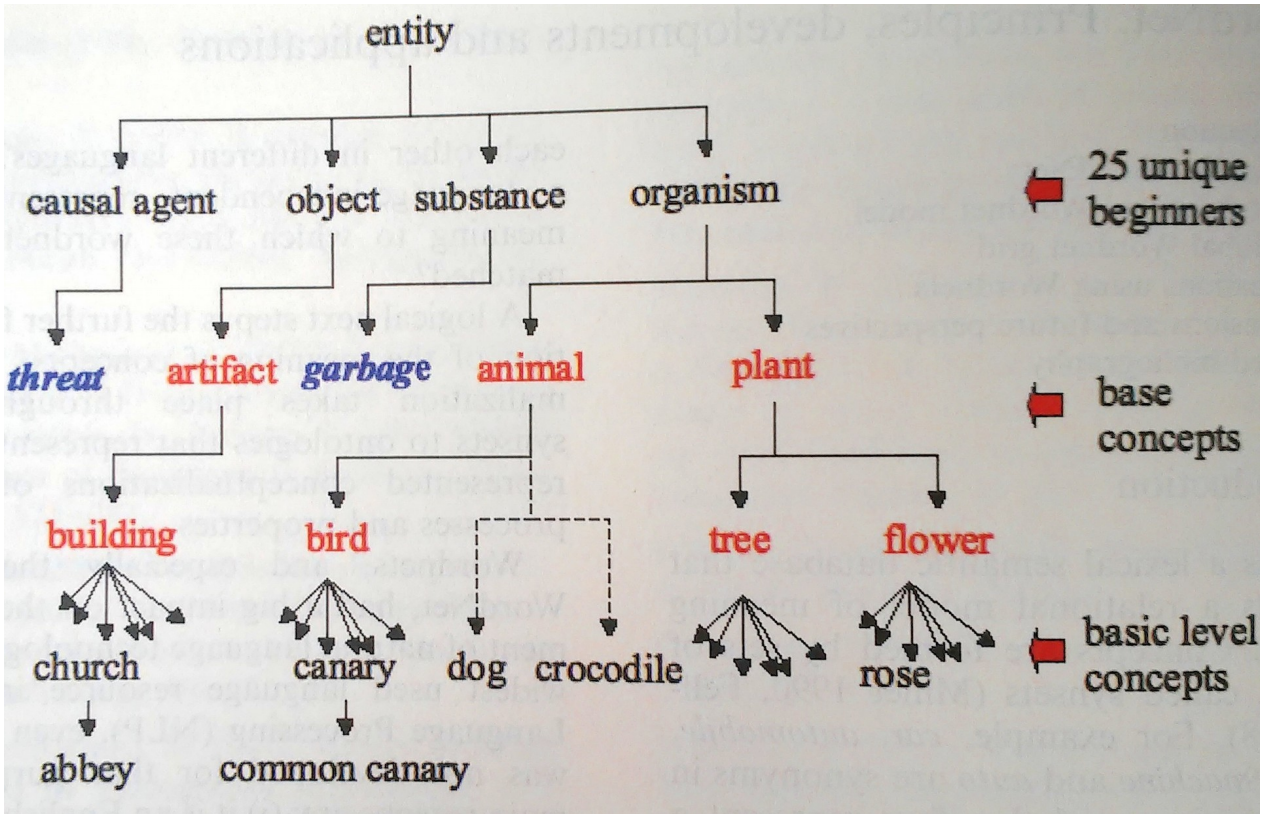
zasada substytucji – brak zmiany wartości prawdziwościowej



1. silna hierarchizacja
2. dominują relacje hiponimii i meronimii

WordNet: struktura sieci rzeczownikowej

(źródło: Vossen 2013)



relacje
1. hiponimia
2. meronimia

→ zróżnicowane poziomy abstrakcji

pojęcia izolowane – ograniczony wpływ na sieć; zwykle stanowią część jakiegoś procesu lub stanu

pojęcia bazowe (*Base Concepts*) – w dużej mierze definiują relacje w sieci

pojęcia poziomu bazowego (*Basic Level Concepts*) – *cf.* Rosch

WordNet: struktura (1)

***synset* – zbiór synonimów powiązany relacjami znaczeniowymi z innymi elementami sieci + glosa + przykład użycia**

zasada substytucji – brak zmiany wartości prawdziwościowej

rzeczowniki

1. silna hierarchizacja
2. dominują relacje hiponimii i meronimii

czasowniki

1. często płytka hierarchia, brak elementu nadrzędnego
2. dominują relacje troponimii, antonimii, wynikania, kauzacji

przymiotniki

1. struktura niehierarchiczna
2. relacja antonimii (np. *słodkie – kwaśne*) dla głównych synsetów i podobieństwa (tj. współdzielenia relacji)

przysłówki

nieliczne w WordNecie angielskim, gdyż w większość wywiedlne z przymiotników

relacje pomiędzy częściami mowy - derywacja

WordNet: po co?

Przetwarzanie języka: przypisanie znaczenia do wyrazów w tekście

- inwentarz znaczeń
- WSD: określanie znaczenia w kontekście *(chair* – mebel czy katedra?*)*
- tematyczna klasyfikacja tekstów
- rozwiązanie anafory („po zakupieniu auta, pojazd musi zostać...”)
- *sentiment analysis*

Aplikacje użytkowe: *information retrieval*

- rozszerzanie kwerend
- rozszerzanie indeksów
- indeksowanie synsetów zamiast wyrazów
- systemy QA
  - klasyfikacja poszukiwanej informacji
- tłumaczenie maszynowe

WordNet: uwagi Geeraertsa

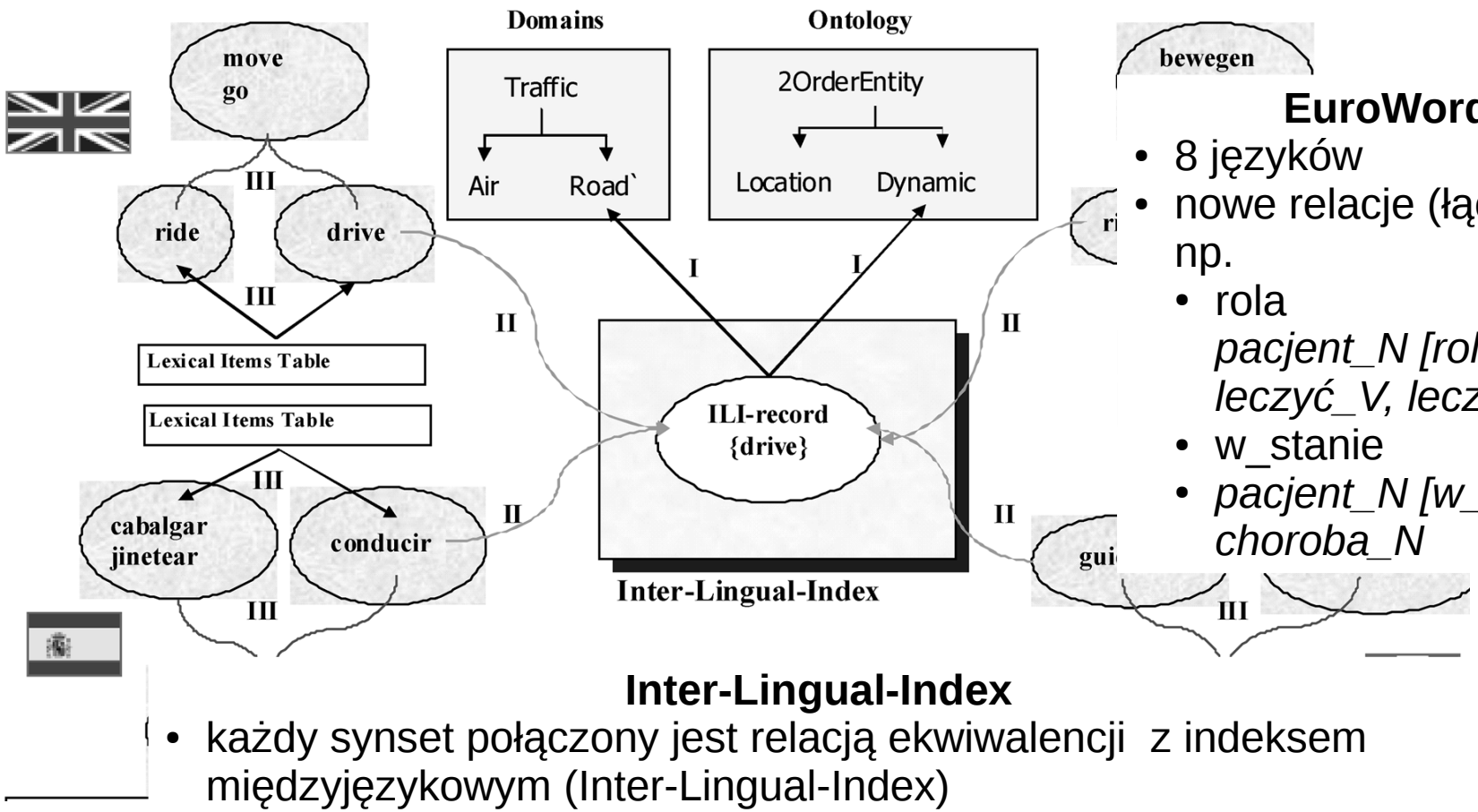
Geeraerts (2010) umieszcza WordNet w obrębie semantyki neostrukturalistycznej

dekompozycja znaczenia	relacje znaczeniowe
Wierzbicka Jackendoff Pustejovsky	WordNet Mielczuk ujęcia dystrybucyjne

- nie opisuje relacji pomiędzy elementami synsetu
- nie opisuje relacji syntagmatycznych
- niektóre z relacji domagają się uszczegółowienia (np.. antonimia
- relacje nie tłumaczą wszystkich aspektów znaczenia (stąd: definicje słownikowe)
- mimo wyjściowego celu (autorzy są psycholingwistami) nie jest modelem leksykonu mentalnego, a należy do leksykologii komputerowej

Poza Princeton WordNet: EuroWordNet

(źródło: Vossen 2004)



EuroWordNet

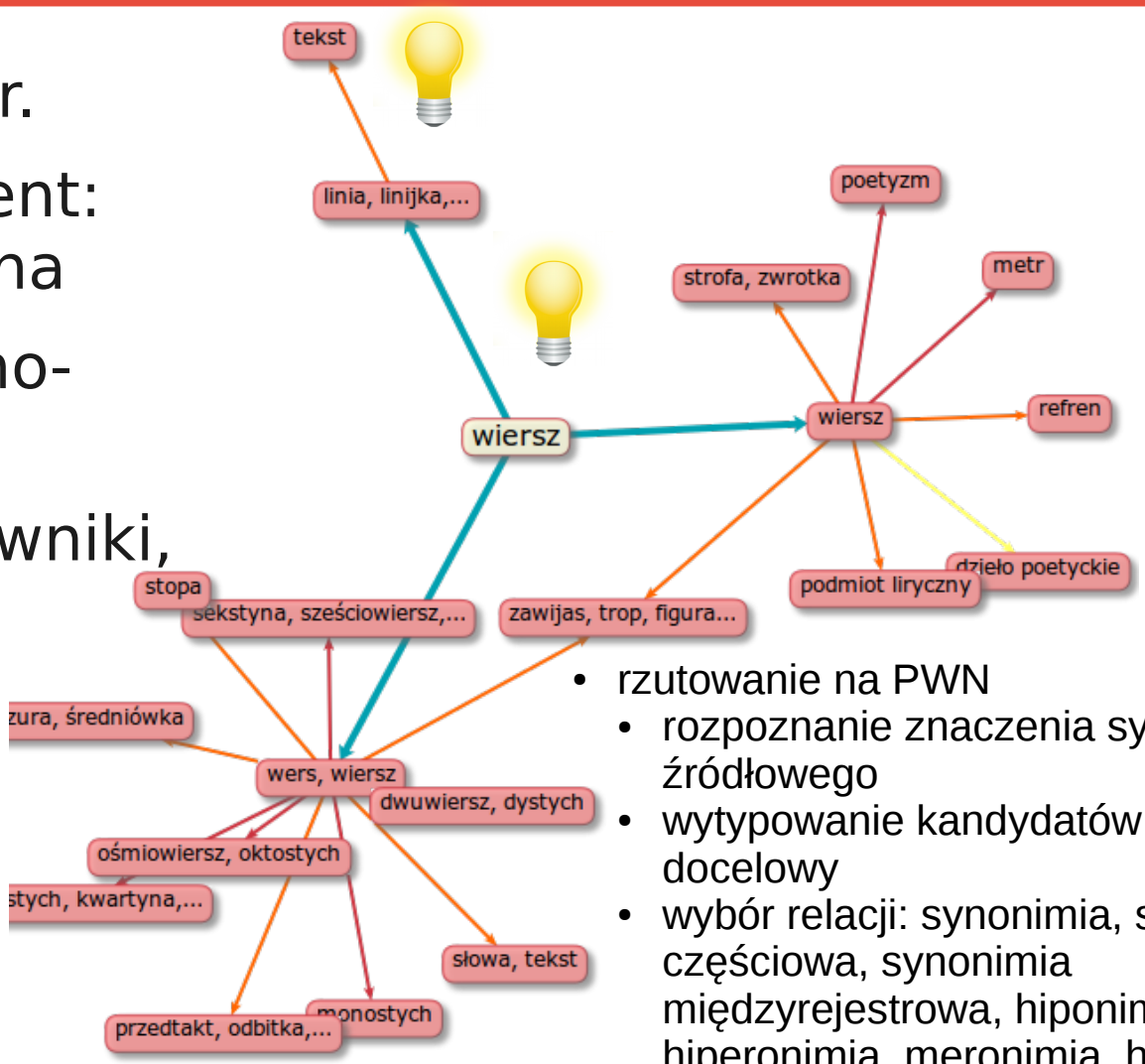
- 8 języków
- nowe relacje (łącznie: 41), np.
  - rola
    pacjent_N [rola_patiens] leczyć_V, leczenie_N
  - w_stanie
  - *pacjent_N [w_stanie] choroba_N*

Inter-Lingual-Index

- każdy synset połączony jest relacją ekwiwalencji z indeksem międzyjęzykowym (Inter-Lingual-Index)
- ILI początkowo został zasilony angielskim WordNetem
- 19 relacji z ILI mogą być złożone: od prostej ekwiwalencji do „podobne_do", "bliski_synonim" itp.

Laboratorium: Słowosieć 3.0

- tworzony od 2005 r.
- podstawowy element: jednostka leksykalna
- 41 relacji leksykalno-semantycznych
- rzeczowniki, czasowniki, przymiotniki
- kryterium przynależności do synsetu: relacje konstytutywne (nie: synonimia!)



(źródło: Słowosieć Viewer)

- rzutowanie na PWN
  - rozpoznanie znaczenia synsetu źródłowego
  - wytypowanie kandydatów na synset docelowy
  - wybór relacji: synonimia, synonimia częściowa, synonimia międzyrejestrowa, hiponimia, hiperonimia, meronimia, holonimia

Laboratorium: SłowoSieć 3.0

1. Wyszukaj opracowane przez siebie hasło w wyszukiwarce na stronie Słowosieci: http://plwordnet.pwr.wroc.pl/wordnet/

a) zidentyfikuj: jednostkę leksykalną, domenę znaczeniową, przykład użycia, elementy synsetu, glosę (definicję);

b) określ synonimy, hiperonimy, meronimy; czym różni się zestaw relacji dla rzeczownika i czasownika?

c) czy podane informacje są poprawne?

d) czy znaczenie określone jest poprawnie (porównaj z WSJP)? czy wnosi cokolwiek do charakterystyki opartej o dane korpusowe?

2. Wyszukaj słowo w przeglądarce na stronie http://nlp.pwr.wroc.pl/slowtool/

a) użyj funkcji wizulizacji;

b) czy potrafisz nazwać relacje semantyczne przedstawione na grafie?

3. Ważnym zastosowaniem WordNetu są algorytmy mierzące podobieństwo wyrazów

a) w parze zbuduj listę rankingową wyrazów *podobnych* do wybranego przez Ciebie (pies, samochód, lenistwo);

b) wykorzystaj funkcję podobieństwa znaczeniowego: czy intuicje kolegi / koleżanki się potwierdziły?

4. Słowosieć może pełnić funkcję repozytorium znaczeń w systemach WSD

a) Wybierz krótką próbkę dowolnego tekstu.

b) Przetestuj działanie „Narzędzie ujednoznaczniania znaczeń leksykalnych” (http://demo.clarin-pl.eu/demo/wsd.html). Dla tekstu niepolskiego wykorzystaj BabelFly (http://babelfy.org/)

c) Czy potrafisz zrozumieć plik wynikowy?

d) Jak możesz zastosować narzędzie we własnych badaniach?

5. Inne narzędzia zgromadzone w sieci CLARIN-PL

CLARIN-PL
Common Language Resources and Technology Infrastructure

FrameNet (1)

Berkeley FrameNet – baza leksykalna dla j. angielskiego oparta o korpus i semantykę ramową (*Frame Semantics*)

Rama semantyczna – schematyczne przedstawienie, skrypt opisujący zdarzenia, przedmioty itd. z ich uczestnikami i właściwościami

Elementy ramy – role semantyczne

*Matilde **fried** the catfish in a heavy iron skillet*

'Matylda usmażyła suma na ciężkiej żelaznej patelni'

jednostka leksykalna (*Lexical Unit*)

fry v. cook or be cooked in hot fat or oil'

aktywuje →

rama semantyczna (*Semantic Frame*)

Apply_heat

elementy ramy (*Frame Elements*)

Duration

A applies heat to , where the of the heat and Duration of application may be specified. A , generally indicated by a locative phrase, may also be expressed. Some cooking methods involve the use of a (e.g. milk or water) by which heat is transferred to the A less semantically prominent or is marked .

*Matilde **fried** the catfish in a heavy iron skillet*

'*Matylda* usmażyła suma na ciężkiej żelaznej patelni'

FrameNet (1)

Fillmore (1977-): język pozwala na wyrażenie konceptualizacji świata – widzimy świat poprzez modele pojęciowe, które wyrażamy, nadając im określoną perspektywę

analiza ramy

1. opis sytuacji
- identyfikacja kluczowych elementów

TRANSAKCJA HANDLOWA

1. jedna osoba obejmuje kontrolę lub własność nad cudzym dobrem na zasadzie umowy, w której pierwsza osoba daje drugiej jakąś sumę pieniędzy
- wiedza: rozumienie własności, wartości pieniądza, umów handlowych
- elementy: Kupiec, Sprzedawca, Dobro, Pieniądze

2. analiza wyrażeń i wzorców gramatycznych
- jak środki językowe podkreślają elementy sytuacji?

2. *kupować* i *sprzedawać* kodują określoną perspektywę, podkreślając wybrane elementy sceny

FrameNet (1)

Berkeley FrameNet – baza leksykalna dla j. angielskiego oparta o korpus i semantykę ramową (*Frame Semantics*)

Rama semantyczna – schematyczne przedstawienie, skrypt opisujący zdarzenia, przedmioty itd. z ich uczestnikami i właściwościami

Elementy ramy – role semantyczne

*Matilde **fried** the catfish in a heavy iron skillet*

'Matylda usmażyła suma na ciężkiej żelaznej patelni'

jednostka leksykalna (*Lexical Unit*)

fry v. cook or be cooked in hot fat or oil'

aktywuje →

rama semantyczna (*Semantic Frame*)

Apply_heat

elementy ramy (*Frame Elements*)

Duration

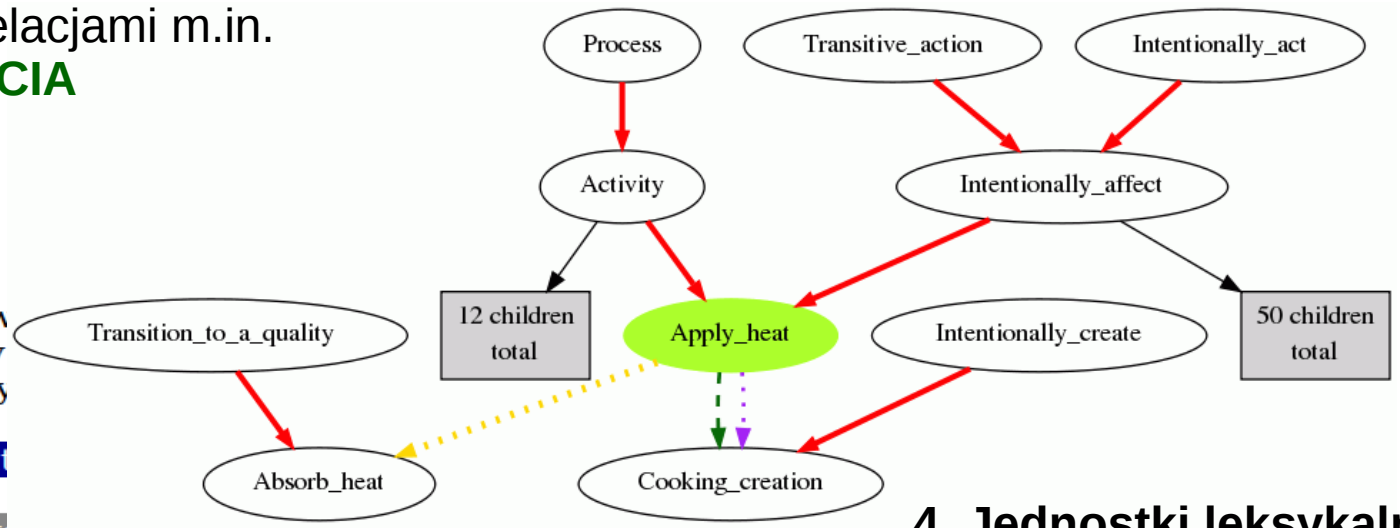
A applies heat to , where the of the heat and Duration of application may be specified. A , generally indicated by a locative phrase, may also be expressed. Some cooking methods involve the use of a (e.g. milk or water) by which heat is transferred to the A less semantically prominent or is marked .

*Matilde **fried** the catfish in a heavy iron skillet*

'*Matylda* usmażyła suma na ciężkiej żelaznej patelni'

FrameNet (2)

Ramy powiązane są relacjami m.in.
DZIEDZICZENIA UŻYCIA
KAUZACJI



(źródło: Framenet FrameGrapher)

1. Definicja

A Cook applies heat to Food, w
Heating_instrument, generally
Medium (e.g. milk or water) by
Co-participant.
Sally FRIED an egg in but

Sally FRIED an egg in a teflon p

2. Elements ramy

Container [Container]
Semantic Type: Container

The Contai
BOIL t
Things that

3. Relacje między ramami

Inherits from: Activity, Intentionally_affect
Is Inherited by:
Perspective on:
Is Perspectivized in:
Uses:
Is Used by: Cooking_creation
Subframe of:

4. Jednostki leksykalne

bake.v, baking.n, barbecue.v, blanch.v,
boiling.n, boil.v, braise.v, broiling.n, broil.v,
.v, char.v, coddle.v, cooking.n,
, deep fry.v, frying.n, fry.v, grilling.n,
melting.n, melt.v, microwave.v,
l.v, plank.v, poach.v, roasting.n,
, saute.v, scald.v, scorch.v, sear.v,
ering.n, simmer.v, singe.v,
ing.n, steam.v, steep.v, stewing.n,
, toasting.n, toast.v

FrameNet: zastosowania

- **inspiracja dla badań (lingwistyka kognitywna, semantyka leksykalna, pismo *Constructions and Frames*)**
- **leksykografia: rygorystyczna analiza konkordancji**
- **automatyczne etykietowanie ról semantycznych (*automatic semantic role labelling*)**
- ***gold standard* dla algorytmów klasteryzacji słownictwa angielskiego**

FrameNet a słowniki walencyjne: Walenty (1)

KIEROWAĆ 'rządzić, prowadzić'
kierować: _: imperf: subj{np(str)} + obj{np(inst)}

ARGUMENT
- dopełnienie
- fraza nominalna
- narzędnik

NEGATYWNOŚĆ (*negativity*)
- pozostaje w mocy dla wszystkich form

ASPEKT

ARGUMENT
- podmiot
- fraza nominalna
- strukturalny

kiepsko	sprawc
kierować	(F) spr
kierowca	sprawc

emat:	pewny [15441]	
kcja:	subj	obj
fraz:	np(str)	np(inst)
		ncp(inst,int)

02	Jarosław Pater zawsze, kiedy kierował służbowym autem, czuł niepewnie.	podkorpus zrównoważony NKJP (300M segmentów)
	to kierowane było przez	pełny NKJP

http://walenty.ipipan.waw.pl/

(źródło: Przepiórkowski et al. 2014)

Walenty (2)

- **kontrola składniowa**

OBIECAĆ
subj,controller{np(str)} + {np(dat)} + controllee{infp(_)}

≠

KAZAĆ
subj{np(str)} + controller{np(dat)} + controllee{infp(_)}

- **frazeologia**

ZBIĆ
subj{np(str)} + obj{np(str)} + {fixed('na kwaśne jabłko')}

PŁYNĄĆ
subj{lexnp(str,sg,'krew',atr)} + {preplexnp(w,loc,pl,'żyła',ratr)}

- **przyimki złożone**

ROZPACZAĆ
subj{np(str)} + {comprepnp(z powodu)}

automatyczna analiza składniowa zdań

- Uczyłem *ją* pisać ≠ Chciałem *ją* poznać (uczyłem → ją; poznać → ją)
- Czekałem *2 godziny* (okolicznik) ≠ Przeczekałem *2 godziny* (arg)

źródło: Przepiórkowski et al. 2013, 2014