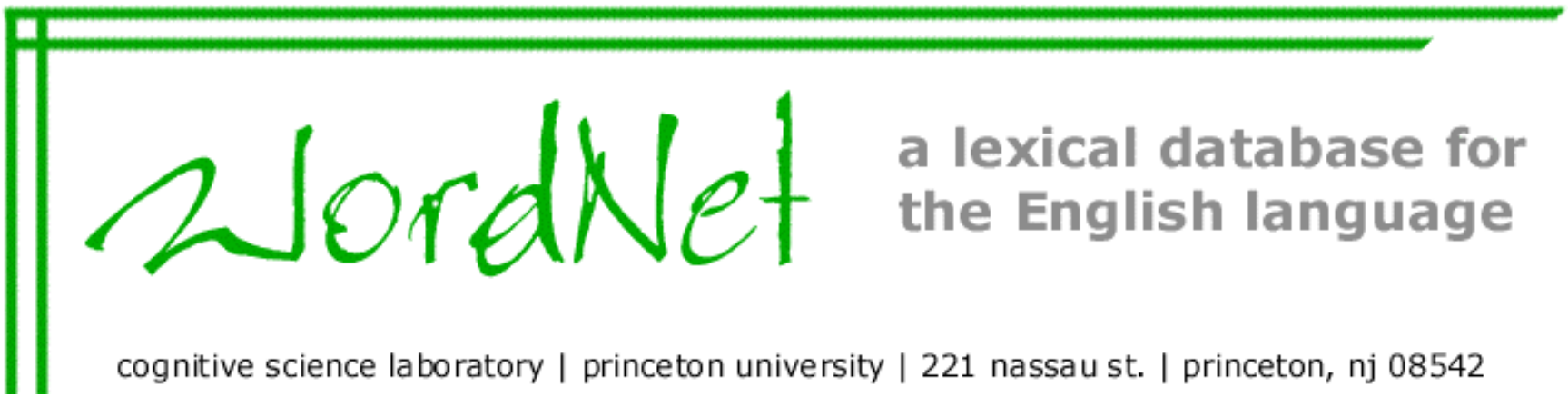


Mikołaj Gierulski

Wstęp

- Psycholingwistyka
  - Zajmuje się psychologiczną bazą funkcjonowania języka, tzn. tym jak język jest przetwarzany przez człowieka (źródło - http://pl.wikipedia.org/wiki/Psycholingwistyka)
- Poniższa prezentacja opisuje koncepcje stosowane w systemie WordNet, ale również badania i analizy z zakresu psycholingwistyki.
- Ponieważ system WordNet zbudowany jest w języku angielskim, przykłady nie będą tłumaczone na polski.

WordNet

Co to jest WordNet

- Prace nad projektem WordNet rozpoczęto 1985 roku w Princeton. Celem powstającego projektu było dostarczenie narzędzia ułatwiającego koncepcyjne (a nie alfabetyczne) przeszukiwanie słownika. (http://www.cogsci.princeton.edu/~wn/)
- W miarę rozwoju projektu cel wyewoluował do znacznie bardziej wyrafinowanego – stworzenie językowego systemu bazodanowego opartego na teoriach psycholingwistycznych dotyczących ludzkiej „pamięci leksykalnej". Inaczej: jest to obszerny zbiór semantycznych zależności pomiędzy słowami w języku angielskim.

WordNet

Czym WordNet nie jest

- WordNet nie jest słownikiem języka angielskiego, jego celem nie jest szczegółowe objaśnienie znaczenia danego słowa.
- WordNet nie jest próbą opisu rzeczywistości za pomocą języka angielskiego.
- WordNet jest projektem, w którym celem nie jest zamodelowanie wiedzy zawartej w języku.

Co to jest WordNet

- WordNet ma na celu zamodelowanie pamięci leksykalnej.
- WordNet może przypominać wyrafinowany słownik wyrazów bliskoznacznych, jest to jednak tylko część jego pełnej funkcjonalności.
- WordNet, w przeciwieństwie do typowego słownika, organizuje słowa pod względem znaczenia, a nie budowy.

WordNet

Semantyka leksykalna

- Każde słowo jest powiązaniem znaczeniowego pojęcia i wyrażenia pełniącego rolę syntaktyczną.
- Niech semantyka leksykalna będzie mapowaniem pomiędzy „formą słowa” a „znaczeniem słowa”.

Znaczenia słowa	Formy słowa				
	F1	F2	F3	...	Fn
Z1	E 1,1	E 1,2			
Z2		E 2,2			
Z3			E 3,3		
...				...	
Zm					E m,n

- F1 i F2 są synonimami, a F2 jest wieloznaczne.
- Widać, że mapowanie jest postaci wiele do wiele.

WordNet

Reprezentacja znaczeń w WordNecie

Jak reprezentować znaczenia słów? Za pomocą definicji.

Czym są definicje?

- Teoria konstruktywna – definicja zawiera wystarczające informacje do jednoznacznego zidentyfikowania (skonstruowania) pojęcia
  - wg. badaczy bardzo trudne do zrealizowania w praktyce, np. definicje słownikowe nie spełniają wymagań tej teorii.
- Teoria różnicowa (differential) – definicja służy do jednoznacznego rozróżnienia pojęć (np. poprzez symbole)
  - wystarczająca, jeśli dane znaczenie jest znane użytkownikowi

WordNet

Reprezentacja znaczeń w WordNecie

Znaczenia są reprezentowane w drugi z wymienionych sposobów. Każde znaczenie jest jednoznacznie reprezentowane przez zbiór słów (synonimów), z którymi jest związane

Np. dwa ze znaczeń słowa *board* reprezentowane są przez zbiory:

- {*board*, *plank*}
- {*board*, *committee*}

Takie zbiory synonimów będziemy nazywać **synsetami**. Są to podstawowe jednostki semantyczne w WordNecie.

WordNet

Synsety

- Synsety nie objaśniają, co znaczy dane pojęcie. Pokazują jedynie, że dane pojęcie istnieje.
- Czasem nie istnieją synonimy dla słowa w danym znaczeniu. Wtedy dołącza się do niego krótkie objaśnienie konkretyzujące dane pojęcie
  - {*board*, (a person's meals, provided regularly for money)}
- Objaśnienie nie ma na celu stworzenia nowego pojęcia u osoby, która go nie zna, a jedynie wyjaśnienie, które ze znaczeń dany synset reprezentuje.

WordNet

Struktura WordNet

- WordNet składa się ze zbioru synsetów oraz związków między nimi.
- Związki pomiędzy synsetami to relacje leksykalne i semantyczne. Są to m.in.:
  - Jednoznaczność (Synonymy)
  - Przeciwieństwo (Antonymy)
  - Hiponimia (Hyponymy)
  - Przynależność (Meronymy)
  - Relacje morfologiczne

Jednoznaczność

- Podobieństwo znaczeniowe jest jedną z podstawowych relacji w WordNet.
- Istnieje definicja mówiąca, że dwa wyrażenia są synonimami, jeśli podstawienie jednego za pomocą drugiego nigdy nie zmieni prawdziwości danego zdania.
- Takie synonimy występują nader rzadko. Ta definicja jest dla nas zbyt mocna, dlatego skorzystamy z innej: wyrażenia są synonimami w kontekście C, jeśli podstawienie jednego za drugie kontekście C nie zmieni prawdziwości zdania.
- Np.: podstawienia *board* za *plank* praktycznie zawsze można dokonać w kontekście stolarstwa, ale istnieją konteksty, w których nie miałoby ono sensu.

WordNet

Jednoznaczność

- Taka definicja jednoznaczności wymusza podział zbioru słów na rzeczowniki, czasowniki, przymiotniki i przysłówki.
- Podział taki jest naturalny dla człowieka.
- Badania psycholingwistyczne dowodzą, że taki podział występuje w leksykalnej pamięci ludzkiej.

Przeciwieństwo (Antonymy)

- Relacja leksykalna, zachodzi pomiędzy formami leksykalnymi, a nie pomiędzy pojęciami
- Relacja bardzo trudna do zdefiniowania.
- Przykład:
  - Zarówno {*rise, ascend*}, jak i {*fall, descend*} mają przeciwne znaczenie, ale nie są antonimami, natomiast [*rise/fall*] oraz [*ascend/descend*] to już antonimy.
- Relacja pomiędzy antonimami stanowi podstawową konstrukcję sieci zależności przymiotników i przysłówków w WordNecie i jest odróżniona od zwykłego przeciwieństwa znaczeniowego.

WordNet

Hiponimia (hyponymy/ISA)

- Semantyczna relacja pomiędzy pojęciami
- Oznacza bycie rodzajem czegoś, np.
  - drzewo jest rodzajem rośliny
  - klon jest rodzajem drzewa
- Relacja przechodnia i antysymetryczna
- Wyznacza podstawową strukturę organizacji rzeczowników w WordNet.

Przynależność (meronymy)

- Oznacza bycie częścią czegoś
  (*A y has an x (as a part)./An x is a part of y)*
- Jest przechodnia (z pewnymi zastrzeżeniami) i antysymetryczna

Relacje morfologiczne

- Stanowią znaczną część relacji językowych
- Początkowo w ogóle nie brane pod uwagę w projekcie
- Wymuszone ze względów praktycznych (patrz przykład)
- Zaimplementowane na poziomie interfejsu do bazy.
- Przykład:
  - zapytanie o *trees* powinno zwrócić wszystkie odpowiedzi dotyczące *tree*, ale bez uwzględnionych relacji morfologicznych nic nie zwróci.

Rzeczowniki

WordNet

Rzeczowniki – system dziedziczenia

- Rzeczowniki (synsety) zorganizowane są w struktury (prawie) drzewiaste.
- Struktury drzew wyznacza relacja ISA (*x is a (kind of) y*)
- Oznacza się ją przez @→
  - *oak* @→ *tree* @→ *plant* @→ *organism*
- ISA jest symetryczna. Relację specjalizacji oznacza się przez ~→
- Struktura zależności rzeczowników tworzy LIS (Lexical Inheritance System)

WordNet

Podział rzeczowników

- Wszystkie synsety podzielone są na kategorie.
- Każda kategoria wywodzi się o jednego, najogólniejszego synsetu – korzenia.
- Zdecydowano nie wyprowadzać wszystkich kategorii z jednego nienaturalnego metakorzenia.
- Drzewa są dość płytkie, zwykle nie przekraczają głębokości 10 poziomów.
- Istnieją (rzadkie) krawędzie pomiędzy drzewami.

WordNet

Korzenie

- Wyodrębnione zostały następujące korzenie:

{*act, action, activity*}
{*animal, fauna*}
{*artifact*}
{*attribute, property*}
{*body, corpus*}
{*cognition, knowledge*}
{*communication*}
{*event, happening*}
{*feeling, emotion*}
{*food*}
{*group, collection*}
{*location, place*}
{*motive*}
{*natural object*}
{*natural phenomenon*}
{*person, human being*}
{*plant, flora*}
{*possession*}
{*process*}
{*quantity, amount*}
{*relation*}
{*shape*}
{*state, condition*}
{*substance*}
{*time*}

WordNet

Cechy szczególne

- Synsety są rozróżnialne także poprzez ich cechy szczególne.
- Wyróżnione cechy szczególne to:
  - Atrybuty
  - Części
  - Funkcje
- Atrybuty odwołują się do przymiotników.
- Części odwołują się do rzeczowników.
- Funkcje odwołują się do czasowników.
- Odnośniki do czasowników i rzeczowników zostały zaimplementowane, ale nie wprowadzone do bazy danych.

Objaśnienia

- W późniejszej fazie projektu do synsetów dołączono także krótkie objaśnienia w stylu słownikowym, jako elementy zbioru synsetów
- Przykład: hierarchia dla {*artifact*} zawierająca *case*

{*carton, case0, box*,@ (a box made of cardboard; opens by flaps on the top)}
{*case1, bag*,@ (a portable bag for carrying small objects)}
{*case2, pillowcase, pillowslip, slip2, bed linen*,@ (a removable and washable cover for a pillow)}
{*bag1, case3, grip, suitcase, traveling bag*,@ (a portable rectangular traveling bag for carrying clothes)}
{*cabinet, case4, console, cupboard*,@ (a cupboard with doors and shelves)}
{*case5, container*,@ (a small portable metal container)}
{*shell, shell plating, case6, casing1, outside surface*,@ (the outer covering or housing of something)}
{*casing, case7, framework*,@ (the enclosing frame around a door or window opening)}

WordNet

Przynależność (meronymy)

- Relacja całość-część.
- Jest przedstawiana poprzez cechę *parts* synsetów.
- Jest dziedziczona w dół drzewa ISA.
- Jest przechodnia (ale tylko do pewnego stopnia).

Przechodniość relacji przynależności

- Przechodniość wydaje się być w wielu przypadkach mocno ograniczona, np.:
  - klamka jest częścią drzwi, a drzwi częścią domu, jednak zdanie „Dom ma klamkę” jest nienaturalne.
  - gałąź jest częścią drzewa, a drzewo częścią lasu, ale „Gałąź jest częścią lasu” jest nienaturalne.
- Zaproponowano podejście, wg. którego relacje gałąź/drzewo i drzewo/las, nie są tymi samymi relacjami.

Relacje przynależności

- Wyróżniono sześć różnych typów relacji przynależności:
  - componnent-object (*branch/tree*)
  - member-collection (*tree/forest*)
  - portion-mass (*slice/cake*)
  - stuff-object (*aluminum/airplane*)
  - feature-activity (*paying/shopping*)
  - place-area (*Princeton/New Jersey*)
- Trzy typy zostały zamieszczone w WordNecie:
  - *Wm* #p→ *Wh* component(*Wm*)-object(*Wh*)
  - *Wm* #m→ *Wh* member(*Wm*)-collection(*Wh*)
  - *Wm* #s→ *Wh* stuff(*Wm*)-object(*Wh*)

WordNet

Funkcje

- Opisują czynności lub funkcje, jakie dane pojęcie wykonuje lub spełnia
- Zwykle dobrze charakteryzują znaczenie danego pojęcia.
- Są jedynymi cechami pojęć abstrakcyjnych, np.:

  synset {*ornament, decoration*} może być nieokreślonego rozmiaru lub kształtu, nie posiada atrybutów, jest jedynie definiowany przez funkcję, jaką ma spełniać.

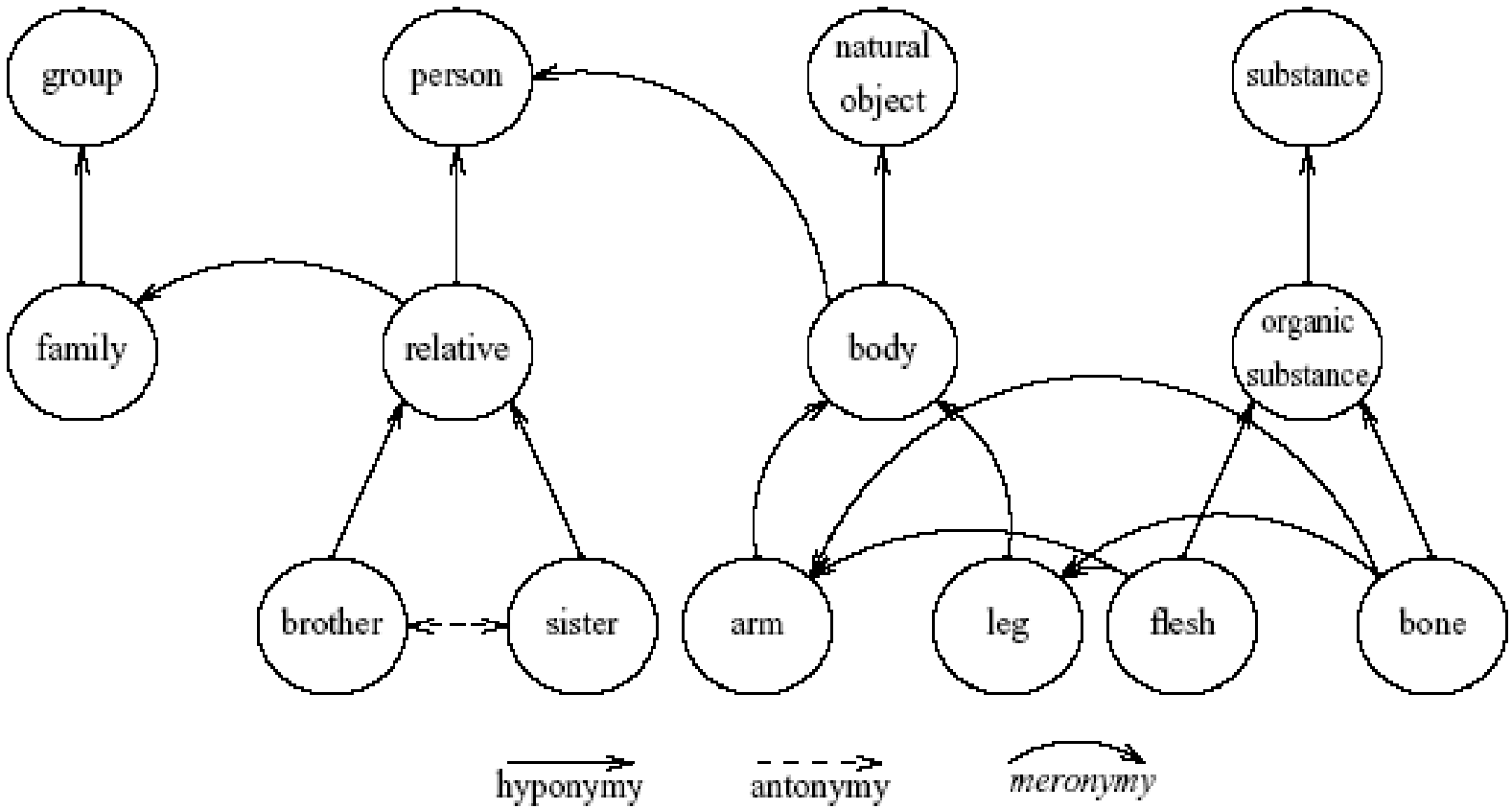
Antonimy

- Relacja najistotniejsza pod względem psychologicznym
- Nie jest kluczowa dla organizacji elementów WordNeta, ale w nim występuje.
- Przykład:

  { [*man, woman*,!], *person*,@. . . (a male person) }

  { [*woman, man*,!], *person*,@. . . (a female person) }

Przykład – reprezentacja trzech semantycznych relacji.



WordNet

Dane liczbowe

- W roku 1993 WordNet zawierał ok. 57 000 rzeczowników pogrupowanych w ok. 48 000 synsetów.

Czasowniki

WordNet

Czasowniki

- Charakteryzują się znacznie większą wieloznacznością, niż rzeczowniki
- *Collins English Dictionary* zawiera 43,636 rzeczowników, a tylko 14,190 czasowników
- Na rzeczownik przypada ok. 1,74 znaczeń, a na czasownik 2,11.

WordNet

Czasowniki w WordNet

- Są przechowywane w synsetach, charakteryzujących ich znaczenie, np.:
  - {*beat, strike, hit*}
  - {*beat, flatten*}
  - {*beat, throb, pulse*}
  - {*beat, defeat*}
  - {*beat, flog, punish*}
  - {*beat, circumvent (the system)*}
  - {*beat, shape, do metalwork*}
  - {*beat, baffle*}
  - {*beat, stir, whisk*}
  - {*beat, mark*}

WordNet

Synsety czasowników

- Aby zwiększyć czytelność i jednoznaczność synsetów, powinny zawierać referencje do rzeczowników, do których czasowniki z danego synsetu się odnoszą – opcja obecnie niezaimplementowana.

Czasowniki w WordNet

- WordNet zawiera ok. 21 000 form czasownikowych, pogrupowanych w ok. 8400 synsetów (w tym tzw. phrasal verbs, jak *look up* czy *fall back)*
- Synsety pogrupowane są w 15 plikach.
- 14 z nich dotyczy konkretnych grup semantycznych.
- 1 plik zawiera synsety dotyczące stanu (np. *belong*, *resemble*). Synsety z tego pliku nie tworzą grupy semantycznej.

WordNet

Synsety czasownikowe

- Występuje bardzo mało prawdziwych synonimów czasownikowych (np. *shut*, *close*). Istnieje natomiast duża grupa słów o podobnym znaczeniu, ale nie wymiennych (np. *begin-commence*, *end-terminate*, *rise-ascend*, *blink-nictate*, *behead-decapitate*, *spit-expectorate*)
- W większości w skład synsetów wchodzą peryfrazy, np. {*swim,travel through water*}, {*mumble, talk indistinctly*}, {*saute, fry briefly*}
- Obrazują one często proces słowotwórczy, np. {*whiten, become white*}, {*enrich, make rich*}

WordNet

Dekompozycja, a model relacyjny

- Dekompozycja polega na badaniu czasowników poprzez rozbiór ich na części składowe, np. zbiór bazowych czasowników i przymiotniki.
- Podejście relacyjne bada język, jako zbiór istniejących „rekordów” powiązanych relacjami, nadającymi sens poszczególnym elementom.
- Struktura WordNetu jest znacznie bliższa drugiemu podejściu, choć zawiera pewne elementy pierwszego (pogrupowanie czasowników w semantyczne domeny wywodzące się od pojęć bazowych)

Relacje między czasownikami

- Podstawową relacją narzucającą strukturę synsetów czasownikowych w WordNet jest relacja „pociągania”/implikacji (*Lexical Entailment*)

Lexical Entailment

- Odpowiednik logicznej implikacji
- Zachodzi pomiędzy V1 i V2, jeśli stwierdzenie
  - *Someone V1*

  pociąga za sobą prawdziwość stwierdzenia
  - *Someone V2*

  mówimy wtedy, ze V1 pociąga za sobą V2

  Przykład (*snore/sleep*)
    - *He is snoring*

    Pociąga za sobą
    - *He is sleeping*

WordNet

Lexical Entailment

- Jest to relacja antysymetryczna (za wyjątkiem synonimów)
- Pomiędzy elementami może zachodzić tymczasowe zawieranie
  - Kierunek zawierania nie jest zawsze ten sam, np.:

    snore pociąga za sobą *sleep* i się w nim zawiera, natomiast

    buy pociąga za sobą *pay* i jednocześnie je zawiera
  - Jeśli V1 pociąga za sobą V2 oraz zachodzi między nimi częściowe zawieranie, to zachodzi pomiędzy nimi związek całość-część.

WordNet

Troponymy

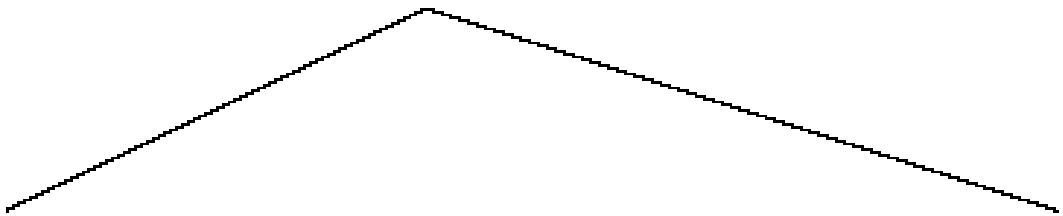
- Jest odpowiednikiem ISA dla czasowników
- Zachodzi pomiędzy V1 i V2 jeśli

  To V1 is to V2 in some particular manner

  np. *to limp* is *to walk in a certain manner*
- Jest szczególnym podzbiorem relacji pociągania (entilement)
- Pary w tej relacji mają zawsze wspólny „zasięg”

WordNet

Entailment - podział



+Troponymy

(Co-extensiveness)

limp-walk

lisp-talk

-Troponymy

(Proper Inclusion)

snore-sleep

buy-pay

WordNet

Klasyfikacja czasowników

- Struktura grafów czasowników jest znacznie bardziej „krzaczasta” niż rzeczowników
- Drzewa zwykle mają nie więcej niż 4 poziomy
- W każdej hierarchii wyróżnia się pewien poziom najsilniejszego rozbudowania

Przeciwstawność

- Relacja bardzo ważna psychologicznie także w przypadku czasowników
- Niekoniecznie wyrazy o przeciwnym znaczeniu są antonimami (*rise/fall*, *ascend/descend*, *rise/descend*)
- Niekoniecznie antonimy mają przeciwne znaczenie (*walk/run*)
- Wiele antonimów współdzieli wyraz w relacji pociągania, np. *hit/miss*, obydwa pociągają *aim*
- Współdzielony wyraz w tej relacji oznacza zwykle sekwencję czynności (brak tymczasowego zawierania)

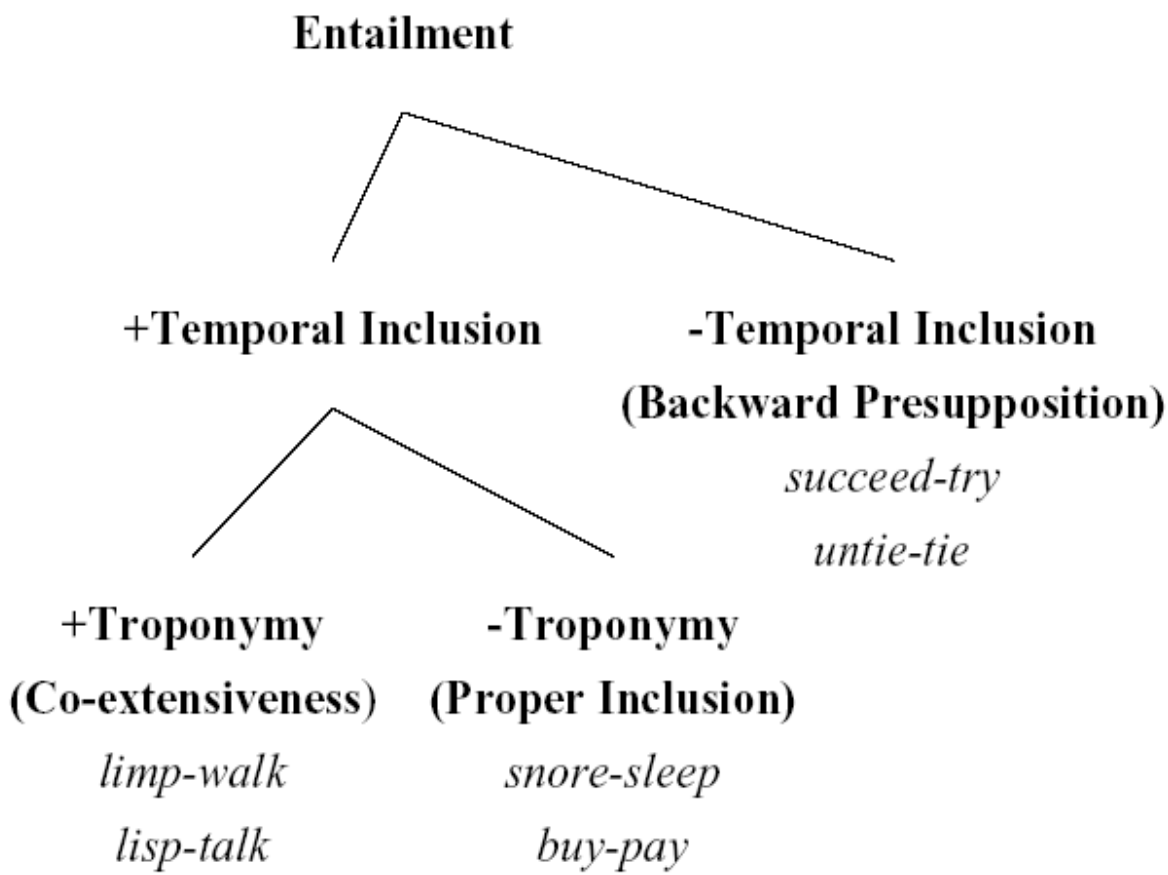
  (*try* jest przed *fail/succeed*, *play* przed *win/lose*)

WordNet

Wsteczna presupozycja

- Informacja zawarta nie bezpośrednio w wypowiedzi, ale milcząco w niej założona
- Jest również istotnym podzbiorem relacji pociągania
- Wyklucza relację całość-część

Entailment - podział



WordNet

Relacja przyczyny

- WordNet zawiera informacje o parach relacji przyczynowo-skutkowej. (np. {*teach*, *instruct*, *educate*} i {*learn*, *acquire knowledge*})
- W przeciwieństwie do relacji pociągania, nie jest ona dziedziczona przez troponimy.
- Czasowniki przyczynowe należą do grupy *cause to be/become/happen/have* lub *cause to do*.
- Przeprowadzają one czasowniki do grupy stanów lub akcji

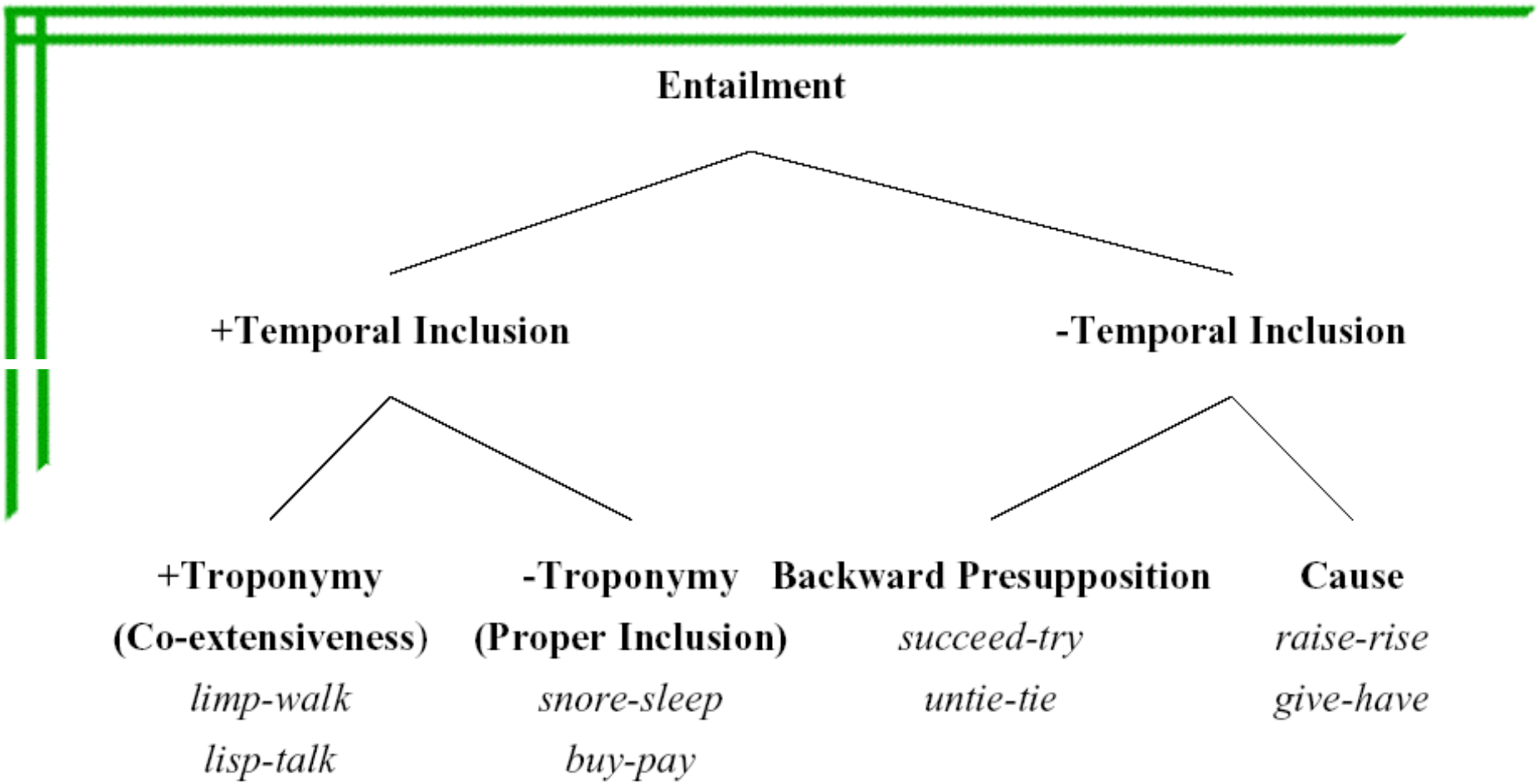
  np. *give*, *teach* prowadzą do stanów *have*, *know* a *feed* do czynności *eat*

WordNet

Przyczyna i pociąganie

- Jeśli V1 jest przyczyną V2, to V2 pociąga za sobą V1.
- Związek ten charakteryzuje się brakiem tymczasowego zawierania (*expel/leave*)

Pełny model relacyjny dla czasowników



WordNet

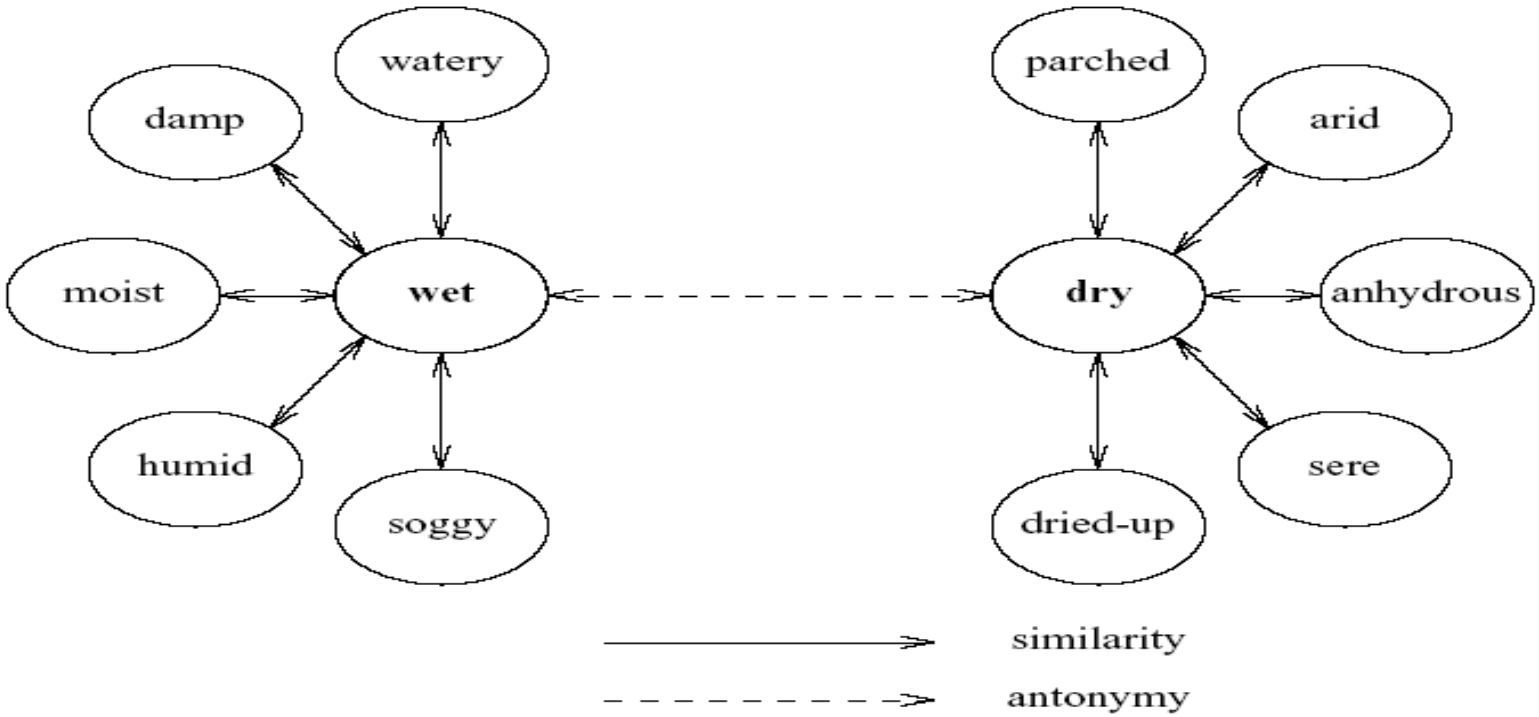
Przymiotniki

- WordNet zawiera ok. 19 500 przymiotników pogrupowanych w ok. 10 000 synsetów
- Przymiotniki podzielone są na grupy
  - Opisowe (*heavy*, *high*)
  - Zmieniające punkt odniesienia (reference modifying) (*former*, *present*, *alleged*, *likely*)
  - Kolory (*black*, *yellow*)
  - Związane z pojęciem (*dental hygiene*, *atomic bomb*)

WordNet

Przymiotniki opisujące

- Powiązane znaczeniowo (synsety)
- Powiązane antonimy
- Powiązanie bipolarne



WordNet

Źródła

- http://www.cogsci.princeton.edu/~wn/

Dokumenty znane jako „5 papers on wordnet”:

- **Introduction to WordNet: An On-line Lexical Database** George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller
- **Nouns in WordNet: A Lexical Inheritance System** George A. Miller
- **Adjectives in WordNet** Christiane Fellbaum, Derek Gross, Katherine Miller
- **English Verbs as a Semantic Net** Christiane Fellbaum
- **Design and Implementation of the WordNet Lexical Database and Searching Software†** Richard Beckwith, George A. Miller, Randee Tengi

WordNet

Koniec

Koniec

- Jakieś pytania?

WordNet