

CLARIN-PL

– wielka infrastruktura badawcza technologii językowych dla nauk humanistycznych i społecznych



Grupa Naukowa G4.19 Technologii Językowej i Lingwistyki Informatycznej



Maciej Piasecki

Politechnika Wrocławska

Katedra Inteligencji Obliczeniowej

maciej.piasecki@pwr.edu.pl

2019-09-25

- CLARIN =
 - Common Language Resources and Technology Infrastructure
 - Wspólne zasoby językowe i infrastruktura technologiczna
- Rozproszona infrastruktura badawcza technologii językowych dla nauk humanistycznych i społecznych
 - 45 centrów technologicznych i wiedzy w 22 krajach
 - działających jako jeden wspólny system
- Zakres działania
 - zbiory danych i bazy danych opisujące język naturalny oraz jego użycie
 - programy komputerowe do analizy tekstu i mowy na różnych poziomach analizy języka naturalnego
 - aplikacje badawcze wspierające badania w obszarach nauk humanistycznych i społecznych

RPA



Podstawowe pojęcia



- Zasoby językowe
 - **zbiory danych** i **bazy danych** opisujące język naturalny oraz jego użycie
 - **sformalizowany opis** wybranych aspektów języka naturalnego
- Narzędzia językowe
 - **programy komputerowe** do przetwarzania tekstu i mowy na różnych poziomach analizy języka naturalnego
 - **automatyczna analiza** struktur językowych, np. analiza składniowa
 - **zastosowania użytkowe**, np. rozpoznawanie i klasyfikacja nazw własnych
- **Technologia językowa** = zasoby + narzędzia + infrastruktura
- **Infrastruktura językowa**
 - wspólna baza technologiczna zapewniająca połączenie zróżnicowanych narzędzi i zasobów językowych

Bariery w zastosowaniu technologii językowej



- **Fizyczna**
 - narzędzia i zasoby nie są dostępne w sieci
- **Informacyjna**
 - brak opisu narzędzi i zasobów
 - brak katalogów i możliwości łatwego odnalezienia
- **Technologiczna**
 - brak standardów, możliwości łączenia elementów technologii
 - brak wspólnej platformy – różnorodność rozwiązań technologicznych
 - brak sprzętu o określonych parametrach
- **Wiedzy**
 - wymagane umiejętności programistyczne
 - wymagana wiedza z zakresu inżynierii języka naturalnego
- **Prawna**
 - licencje ograniczające dostęp i wykorzystanie
 - szczególnie w odniesieniu do korpusów

Bariery w zastosowaniu technologii językowej



- **Fizyczna**
 - dostępność: www.clarin-pl.eu, www.clarin.eu
- **Informacyjna**
 - opis: metadane, dokumentacja
 - katalog: Virtual Language Observatory
- **Technologiczna**
 - standardy: CMDI, unifikacja standardów dla narzędzi
 - platforma – WebLicht, platforma polska w ramach CTJ CLARIN-PL
 - sprzęt: Centrum Technologii Językowych CLARIN-PL
- **Wiedzy**
 - aplikacje badawcze rozwijane wspólnie z użytkownikami
 - CTJ działające jako centrum typu K (szerzenia wiedzy)
- **Prawna**
 - otwartość! (ang. Open source, open access) otwarte rozwiązania!
 - trudne w odniesieniu do korpusów, ale model otwartości wyników przetwarzania
- **Finansowa**: otwartość kosztuje, a kto płaci? **MNiSW!** 😊 (dziękujemy!)

Podstawowe funkcje CLARIN



- Ułatwienie dostępu do zasobów językowych
 - federacja repozytoriów – [Virtual Language Observatory](#)
 - federacyjne przeszukiwanie korpusów – [Federated Content Search](#)
- Wsparcie dla automatycznej analizy tekstu i mowy
 - paleta gotowych do użycia narzędzi językowych
 - usługi sieciowe (Web Services) i aplikacje narzędziowe
 - dostęp poprzez repozytoria
 - typowe narzędzia – [Language Switchboard](#) (WebLicht)
- Aplikacje badawcze
 - budowane pod konkretne potrzeby, często we współpracy z użytkownikami, np. LEM, WebSty
 - oparte na technologii językowej, ale nie `narzucające' jej

Podstawowe funkcje CLARIN



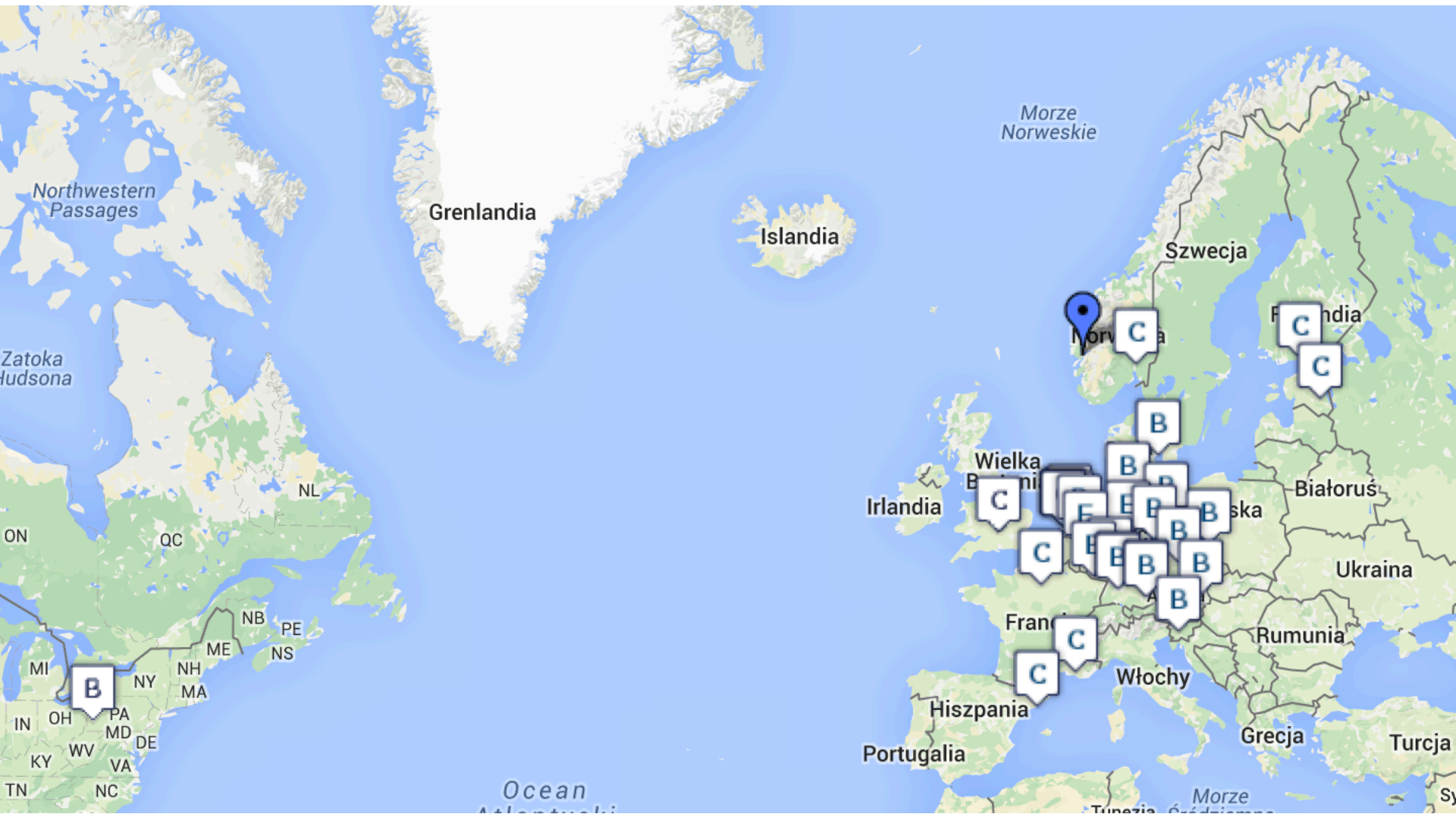
- Infrastruktura Szerzenia Wiedzy (*Knowledge Sharing Infrastructure*)
 - centra CLARIN typu K – oferujące dostęp do wiedzy eksperckiej i wsparcie; fizyczne i wirtualne
 - poradniki i punkty informacyjne, warsztaty, itd.
- Wymogi dla centrum typu K (wybrane)
 - strona WWW z jasno określonym zakresem usług
 - np. pomoc, wsparcie techniczne lub technologiczne, kursy, ...
 - zapewniają reaktywne usługi, np. odpowiedzi na zapytania użytkowników w ciągu 2 dni roboczych
 - aktywne, np. seminaria, warsztaty, szkolenia
 - dysponują odpowiednią kadrami

Przykłady centrów typu K



- CKLD: CLARIN Knowledge-Centre for linguistic diversity and language documentation
- CLARIN K-Centre DANSK - DANish helpdeSK (język duński)
- CLARIN Knowledge Centre for Language Learning Analysis (CLARIN-Learn) (TalkBank)
- CLARIN Knowledge Centre for Speech Analysis (CLARIN-SPEECH)
- CLARIN Knowledge Centre for Treebanking (Univ. Bergen i LINDAT, Prague)
- PolLinguaTec - CLARIN Knowledge Centre for Polish Language Technology
 - zastosowania technologii językowych
 - **kontakt: uzytkownicy@clarin-pl.eu**

CLARIN ERIC – Centra



Podstawowe funkcje CLARIN



CLARIN makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. CLARIN offers long-term solutions and technology services for deploying, connecting, analyzing and sustaining digital language data and tools. CLARIN supports scholars who want to engage in cutting edge data-driven research, contributing to a truly multilingual European Research Area. [Read more...](#)



Participating Organizations

The operations, services and centres of the CLARIN infrastructure are provided and funded by the national consortia in the countries that have joined CLARIN ERIC or by associated centres.

[Learn more](#)



Services, Tools and Data

CLARIN provides a wide variety of Services and Data sets

[Learn more](#)

Podstawowe funkcje CLARIN



Services



CLARIN portal

Get an example-based impression of what's currently available



Depositing services

Store language resources in a sustainable repository at a CLARIN centre



Virtual Language Observatory

Discover language resources using a faceted browser or a map



Easy access to protected resources

Get easy access to protected resources, with your institutional username and password.



Language Resource Switchboard

Explore and analyze language data with a wide variety of tools



Virtual Collections

Create your own digital bookmarks, ideal for citing data sets.



Language Resource Inventory

Submit and access information about language resources relevant to your research.



Content Search (prototype)

Search different corpora with a single search engine



Questions & Answers

Searching for a specific data set or application? Wondering how CLARIN can assist your research? Feel free to contact us!

Virtual Language Observatory



VLO / [Faceted search](#) / [Search results](#)



wordnet



Showing 1 to 10 of 88 results for wordnet

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language



Collection



Resource type



Modality



Type to search for more

writtenlanguage (5)

written (4)

multimodal (1)

other (1)

Format



WordNet

⊕ plWordNet is a lexico-semantic network which reflects the lexical system of the Polish language. There are at present ca. 144,000 nouns, verbs and adjectives in plWordNet, ca. 203,000 word senses and ca. 500,000 relations. It is already the second-largest wordnet in the world, and it keeps growing.

1

*

Finnish WordNet

⊕ The language resource is available in Kielipankki - the Language Bank of Finland at <http://urn.fi/urn:nbn:fi:lb-2016042205>; download: <http://urn.fi/urn:nbn:fi:lb-2014052713>. The Finnish WordNet is a lexical database for Finnish. It is a part of the FIN-CLARIN infrastructure project. FinnWordNet is lic...



Czech WordNet

⊕ Multilingual Lexicons; The Czech WordNet was developed by the Centre of Natural Language Processing at the Faculty of Informatics, Masaryk University, Czech Republic. The Czech WordNet captures nouns, verbs, adjectives, and partly adverbs, and contains 28,201 word senses (synsets). Every synset encodes the equivalenc...

1



VLO – wyszukiwanie fasetowe



- Fasetowe wyszukiwanie
- Atrybuty i ich zakresy wartości są odczytywane z meta-danych
- Wspólny standard: CMDI (Componet Metadata Infrastructure)

Use the categories below to limit the search results to those matching the selected value(s).

Language

Type to search for more

Dutch (442659)
Danish (148146)
English (111107)
German (77521)
Unspecified (54897)
French (19181)
Latin (17886)
Spanish; Castilian (15322)
Japanese (8442)
Deutsch (7878)
more...

Collection

Resource type

Modality

Format

Keyword

Centralne przeszukiwanie treści



Federated Content Search

Aggregator

Help

BETA

Search text

książka



Search for

Any Language ▾

Text Resources ▾

in

All available collections ▾

and show up to

10 ▾

hits

6 matching collections found

☐ Display as Key Word In Context

⬇ Download ▾

› pol_news_2011_1M – ASV Leipzig

👁 View

› Child Language Data Exchange System (CHILDES) – MPI for Psycholinguistics

👁 View

▼ ClarinPL Corpus 1.0 – CLARIN-PL Language Technology Centre

👁 View

że całość czyta się lekko , nie jest to książka do tramwaju , nie nadaje się na prezent

▼ Korpus PWr – CLARIN-PL Language Technology Centre

👁 View

Sukces CLARIN - Polska



- Centra
 - *Centrum Technologii Językowej* CLARIN-PL
 - repozytorium danych językowych (Certyfikat DSA)
 - *CLARIN Cloud* – prywatna chmura danych dla naukowców
 - usługi i aplikacje do analizy tekstów i mowy
 - międzynarodowy Certyfikat Centrum CLARIN typu B
 - *PolLinguaTec* - Centrum Wiedzy Technologii Językowych dla Języka Polskiego
 - międzynarodowy Certyfikat Centrum CLARIN typu K
- Współpraca międzynarodowa
 - usługi dla całego CLARIN ERIC,
 - bardzo duża liczba polskich narzędzi podłączona do systemów centralnych: *Language Switchboard*, *WebLicht*, *D4Science* (projekt PARTHENOS)

- Wytwórcy
 - Grupa Naukowa G4.19 Lingwistyki Informatycznej i Technologii Językowych, Katedra Inteligencji Obliczeniowej, Politechnika Wrocławska
 - Zespół Inżynierii Lingwistycznej, Instytut Podstaw Informatyki PAN,
 - Instytut Slawistyki, Polska Akademia Nauk
 - Polsko-Japońska Akademia Technik Komputerowych
 - Instytut Anglistyki, Uniwersytet Łódzki
 - Uniwersytet Wrocławski
- Beneficjenci:
 - **Wszystkie** jednostki naukowe i **Naukowcy** w Polsce, szczególnie z obszaru nauk humanistycznych i społecznych

- Formalizacja w opisie języka naturalnego
 - morfologia:
 - gramatyki formalne (regularne) poziomu znaków i morfemów
 - składnia:
 - gramatyki formalne: frazowe, unifikacyjne
 - semantyka:
 - opis znaczeń leksykalnych (słów) za pomocą grafu relacji
 - logika jako podstawa reprezentacji znaczenia
 - pragmatyka
 - modele oparte na logice i formalnej reprezentacji wiedzy
 - modele wieloagentowe

Kontekst - przykład



- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)

Kontekst - przykład



- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło **Laatzen** w pobliżu **Hanoweru** w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: **miejsowości**

Kontekst - przykład



- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas **autostrady A7** koło **Laatzen** w pobliżu **Hanoweru** w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: **miejsowości**, **drogi**

Kontekst - przykład



- „Ze względu na zagrożenie powodziowe policja zamknęła **boczny pas** autostrady A7 koło **Laatzen** w pobliżu **Hanoweru** w kierunku na **południe**.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: **miejsowości**, drogi, **części dróg**

Kontekst - przykład



- „Ze względu na zagrożenie powodziowe **policja** zamknęła **boczny pas** autostrady A7 koło **Laatzen** w pobliżu **Hanoweru** w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: **miejsowości**, drogi, **części dróg**, **służby**

- „Ze względu na **zagrożenie powodziowe** **policja** zamknęła **boczny pas** autostrady A7 koło **Laatzen** w pobliżu **Hanoweru** w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: **miejsowości**, drogi, **części dróg**, **służby**
- sytuacje: **zagrożenie powodziowe**

Kontekst - przykład



- „Ze względu na **zagrożenie powodziowe** policja **zamknęła** boczny pas autostrady A7 koło **Laatzen** w pobliżu **Hanoweru** w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: **miejsowości**, drogi, **części dróg**, **służby**
- sytuacje: **zagrożenie powodziowe**, **zamknięcie**

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas)

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas), przyczyna(Syt:zagrożenie, Syt:zamknięcie)

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas), przyczyna(Syt:zagrożenie, Syt:zamknięcie), bliskość(Pas, Laatzen)

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas), przyczyna(Syt:zagrożenie, Syt:zamknięcie), bliskość(Pas, Laatzen), bliskość(Pas, Hanower)

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas), przyczyna(Syt:zagrożenie, Syt:zamknięcie), bliskość(Pas, Laatzen), bliskość(Pas, Hanower)
- umiejscowienie w czasie: ??

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas), przyczyna(Syt:zagrożenie, Syt:zamknięcie), bliskość(Pas, Laatzen), bliskość(Pas, Hanower)
- umiejscowienie w czasie: ??
- wiedza nadawcy: ??

Kontekst - przykład



- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas), przyczyna(Syt:zagrożenie, Syt:zamknięcie), bliskość(Pas, Laatzen), bliskość(Pas, Hanower)
- umiejscowienie w czasie: ??
- wiedza i cele nadawcy: ??

Kontekst - przykład



- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas), przyczyna(Syt:zagrożenie, Syt:zamknięcie), bliskość(Pas, Laatzen), bliskość(Pas, Hanower)
- umiejscowienie w czasie: ??
- wiedza i cele nadawcy: ??
- powiązania z informacją w pozostałej części tekstu: ??

- „Ze względu na zagrożenie powodziowe policja zamknęła boczny pas autostrady A7 koło Laatzen w pobliżu Hanoweru w kierunku na południe.”
 - (Onet.pl za Deutsche Welle, 30 V 2013)
- obiekty: miejscowości, drogi, części dróg, służby
- sytuacje: zagrożenie powodziowe, zamknięcie
- relacje: przyczyna, przestrzenne
- elementy relacji: zamknięcie(Policja, Pas), przyczyna(Syt:zagrożenie, Syt:zamknięcie), bliskość(Pas, Laatzen), bliskość(Pas, Hanower)
- umiejscowienie w czasie: ??
- wiedza i cele nadawcy: ??
- powiązania z informacją w pozostałej części tekstu: ??
- ...

- Zasoby językowe
 - **zbiory danych** i **bazy danych** opisujące język naturalny oraz jego użycie
 - **sformalizowany opis** wybranych aspektów języka naturalnego
- Narzędzia językowe
 - **programy komputerowe** do przetwarzania tekstu i mowy na różnych poziomach analizy języka naturalnego
 - **automatyczna analiza** struktur językowych, np. analiza składniowa
 - **zastosowania użytkowe**, np. rozpoznawanie i klasyfikacja nazw własnych
- **Technologia językowa** = zasoby + narzędzia + infrastruktura
- **Infrastruktura językowa**
 - wspólna baza technologiczna zapewniająca połączenie zróżnicowanych narzędzi i zasobów językowych

1. Rozpoznanie struktury dokumentu i wydobycie tekstu
2. Segmentacja
na tokeny (poziom wyrazów), zdania oraz jednostki bardziej złożone
3. Analiza morfologiczna
4. Ujednoznacznianie morfo-syntaktyczne (tagowanie)
5. Ujednoznacznianie sensu słów (znaczeń leksykalnych)
6. Płytki parsing (*opcjonalnie*)
7. Rozpoznawanie wyrażeń wielowyrazowych
w tym jednostek identyfikujących, np. nazw własnych
8. Rozpoznawanie związków w tekście
np. anafory, koreferencji, relacji semantycznych, sytuacji
9. Głęboki parsing
10. Głęboka analiza semantyczna
11. Analiza pragmatyczna (w tym struktury dyskursu)

- Wydobywanie informacji (*Information Extraction*)
 1. *Analiza treści zadania
 2. Wydobycie tekstu
 3. Tokenizacja i segmentacja na zdania
 4. Tagowanie morfo-syntaktyczne: analiza i ujednoznacznianie
 5. Rozpoznawanie jednostek identyfikacyjnych ('bytów nazwanych')
 6. Płytki parsing
 7. Rozpoznawanie powiązań koreferencyjnych
 8. Rozpoznawanie relacji
 9. Łączenie instancji relacji w opisy sytuacji (*merging*)
 10. Konwersja opisów sytuacji do rekordów odpowiedzi
 11. Uzupełnianie i przygotowanie formy wynikowej
 12. *Przekształcenie sformalizowanej informacji w wiedzę oraz ewentualne wnioskowanie (*Text Mining*)

Podstawy inżynierii języka naturalnego: **segmentacja**



- Akapity (ang. *topic segmentation*)
 - kwestia granic
 - akapit jako element struktury dyskursu — semantyczne rozumienie
 - w WWW częste ‘graficzne’ użycie akapitu

- Zdania
 - kwestia granic
 - luźna interpunkcja i ‘plakatowość’ stron WWW

„Z radością informujemy o **30.** dziecku, które wybudziło się w Klinice Budzik.”
czytamy na stronie Fundacji Ewy Błaszczak **Akogo?**.

(<http://fakty.interia.pl/>, dostęp 12 X 2016)

- Haplologia kropki

„Przedawkowanie w antybiotykoterapii może prowadzić do zachwiania równowagi biologicznej i pojawienia się różnego rodzaju grzybic, drożdżyc **itp.**”

(Wiktionary, dostęp 12 X 2016)

- Specjalne konwencje, np. w bibliografii

Podstawy Inżynierii języka naturalnego (IJN): segmentacja



- Tokeny i wyrazy — problem definicji
 - tokenizacja — jednostki składowe tekstu na poziomie wyrazowych:, np. wyrazy, symbole, daty, itd.
- Problemy tokenizacji
 - dzielenie wyrazów — homografia dywizu, e.g. *small-town, co-operation*
 - partykuły, przedrostki i przyrostki, np. w modelu Narodowego Korpusu Języka Polskiego (NKJP):
 - partykuła *się*
 - *Boję **się** głośno roześmiać.*

Podstawy Inżynierii języka naturalnego (IJN): segmentacja



- Problemy tokenizacji
 - ruchome końcówki (NKJP)
 - *Dawno nie widziała**m** Janka.*
 - *Dawno**m** nie widziała Janka.*
 - *Długo szedł**m** i śpiewałem**m**.*
 - *Długo**m** szedł i śpiewał.*
 - formy skrócone i złączenia, np. *isn't*
 - wielowyrazowe, np. *biało-czerwony*
 - nazwy własne
 - akronimy i skróty

- Elementy nie-słowne poziomu zdania
 - znaki interpunkcyjne
 - użycia kropki w skrótach i wewnątrz skrótów, np. *b.r.*, *8. piętrze*, *ul.*
 - użycie kropki w symbolach, nieprawidłowo zapisanych ułamkach dziesiętnych
 - *haploglogia kropki*, np. *itd.*
 - parzyste znaki interpunkcyjne, np. nawiasy, cudzysłowy, ...
szczególny problem to obejmujące całe zdanie
 - liczby, daty, określenia czasu
 - typowe dla WWW: adresy IP i URL, oznaczenia protokołów,
 - różnorodne ‘ciała obce’, np.
 - fragmenty programu,
 - formuły matematyczne,
 - symbole
 - emotikony, itd.

Segmentacja polskich tekstów: przykład

MACA



- SRX (*Segmentation Rules eXchange*) (Pooley and Raya, 2008)
 - standard zapisu reguł segmentacji wyrażeń tekstowych, który z założenia ma być niezależny od producenta i implementacji
 - stworzony został z myślą o narzędziach przetwarzania języka naturalnego, w szczególności systemach tłumaczenia maszynowego
- Typy reguł*:
 - reguły podziału na zdania
 - oraz wyjątki
- Użycie
 - reguły są aplikowane sekwencyjnie i pierwsza pasująca reguła kończy przetwarzanie danego miejsca w tekście
- Budowa reguł
 - dwa wyrażenia regularne, określające warunki konieczne do odpalenia reguły
 - wyrażenie opisujące fragment tekstu, który występuje przed potencjalnym miejscem podziału (`beforebreak`)
 - oraz wyrażenie opisujące fragment tekstu, który występuje po nim (`afterbreak`).

* Opis SRX za fragmentem raportu NEKST nr TR-01-1-10, (Radziszewski i inni)

Segmentacja polskich tekstów: przykład

MACA



- Przykładowe reguły SRX dla języka polskiego (Miłkowski and Lipski, 2009)

```
<!--pok. 131, poz. 124-->
```

```
<rule break="no">
```

```
<beforebreak>\bpo[zk]\.\s</beforebreak>
```

```
<afterbreak></afterbreak>
```

```
</rule>
```

- Wyjątek: jeśli wystąpi ciąg *znak pustypoz.znak pusty* lub *znak pustypok.znak pusty*, to po takim ciągu **nie** wstawimy granicy zdań.

```
<rule break="yes">
```

```
<beforebreak>\s\p{L}[\.!?...]\s</beforebreak>
```

```
<afterbreak>\p{Lu}\p{Ll}</afterbreak>
```

```
</rule>
```

- Słowo rozpoczęte wielką literą następujące po słowie zakończonym znakiem interpunkcyjnym rozpoczyna nowe zdanie

Analiza morfologiczna: **forma** **przybliżona**



- Stemming
 - określenie symbolu reprezentującego różne formy wyrazowe różniące się wartościami kategorii gramatycznych (np. przypadek, liczba, rodzaj itd.), ale też czasami klas gramatycznych, np. bezokolicznik, imiesłów itd.
 - często realizowany bez uwzględnienia wiedzy lingwistycznej
- Rozwiązania
 - język angielski — stosunkowo proste algorytmy
 - języka polski
 - pełna analiza — homografy (!), np. *lecz*, *piec*,
 - analiza na bazie końcówek
 - indeks *a tergo* Tokarskiego
 - transduktor na podstawie indeksu,
np. stemer Weissa zastosowany w systemie *Carrot*
(obecnie część Morfologik: <http://morfologik.blogspot.com>)

Analiza morfologiczna: **forma przybliżona**



■ **Lematyzacja**

- określenie formy wyrazowej reprezentującej zbiór form wyrazowych różniące się wartościami kategorii gramatycznych dla wyrazu
- **lemat** – inaczej forma hasłowa (słownikowej) – forma wyrazowa wybrana jako podstawowa forma morfologiczna ze względu na arbitralnie ustalone wartości kategorii morfologicznych; reprezentuje cały zbiór form różniących się jedynie wartościami kategorii, np.
 - **być** dla *jest, jesteście, był, była*
 - **socha** dla *sochy, sosze, soch*
- lematyzacja – sprowadzenie wyrazów tekstowych do ich lematów
 - możliwa jedynie w kontekście, np. *mam, lata, grab*, itd.
 - problem niejednoznaczności lematyzacji
 - zwykle część tagowania

Analiza morfologiczna: **forma** **przybliżona**



- **Ujednoznacznianie morfo-syntaktyczne** (ang. POS tagging)
 - ustalenie dla każdego wyrazu:
 - **klasy gramatycznej** (lub części mowy)
 - wartości kategorii gramatycznych
 - rozwiązuje problem homografów, np. *piec* (V) vs *piec* (N)
 - ale nie rozwiązuje problemu polisemii, np. *automat* vs *automat*
 - por. znaczenia w Słownosieci: <http://plWordNet.pwr.edu.pl>

- Pojęcie **części mowy** (ang. *Part of Speech*)
 - kategorie / klasy na jakie dzielimy:
wyrazy — formy wyrazowe — leksemy (fleksemy), np.
 - rzeczownik, czasownik, przymiotnik, zaimek, imiesłów, przysłówki, itd.
 - kryteria podziału (Przepiórkowski):
 - dystrybucyjne (wzorce użycia, możliwości wymiany na określonej pozycji)
 - morfologiczne (odmiana, końcówki, **kategorie gramatyczne**)
 - semantyczne (liczebniki, zaimki)
 - pragmatyczne (wykrzykniki, wtrącenia)
- proste kategorie składniowe

- Kategorie gramatyczne
 - definicja?
 - np. NKJP: liczba, przypadek, rodzaj, osoba, stopień, aspekt, negacja, akomodacyjność, poprzymkowość, aglutacyjność, wokaliczność, wymaganie kropki
 - np. wartości kategorii rodzaj w Morfeuszu 2.0:
 - męski osobowy m1, męski zwierzęcy m2, męski rzeczowy m3, żeński f,
nijaki zbiorowy n1, nijaki zwykły n2, przymnogi osobowy p1,
przymnogi zwykły p2, przymnogi opisowy p3

Analiza morfologiczna: model (2)



- Znaczniki — ‘tagi’ (ang. *tags*)
 - reprezentujące informację o części mowy i wartościach kategorii gramatycznych
 - najczęściej POS są zbyt szerokie dla precyzyjnego opisu
 - POS dzielone są znacznikami na **klasy gramatyczne**
 - znaczniki:
 - **proste** — atomowe (np. VBG)
 - **strukturalne** — pozycyjne: wartości dla kategorii na określonych pozycjach
 - np. `<znacznik><lex>kilka</lex>num:pl:acc:f</znacznik>`
- Klasy gramatyczne (Przepiórkowski)
 - oparte na pojęciu **fleksemu** Bienia (1991) (za Woliński):
Fleksem jest zbiorem wyrazów morfologicznych różniących się tylko wartościami kategorii liczby, przypadku, rodzaju i osoby.

Analiza morfologiczna: klasy gramatyczne NKJP



Leksem	Fleksem	Symbol	Przykład
Rzeczownik	rzeczownik forma deprecjatywna	subst depr	<i>kot, profesorowie</i> <i>profesory</i>
Liczebnik	liczebnik główny liczebnik zbiorowy	num numcol	<i>sześć, dużo</i> <i>sześcioro, trojga</i>
Przymiotnik	przymiotnik przymiotnik przyprzymiotnikowy przymiotnik poprzyminkowy przymiotnik predykatywny	adj adja adjp adjc	<i>polski</i> <i>polsko</i> <i>polsku</i> <i>wesół</i>
Przysłówek		adv	<i>bardziej, kiedy</i>
Zaimek	nietrzecioosobowy trzecioosobowy SIEBIE	ppron12 ppron3 siebie	<i>ja, tobie</i> <i>on, jemu</i> <i>sobą</i>

NKJP (Przepiórkowski et al., 2012, str. 63)

Analiza morfologiczna: klasy gramatyczne NKJP



Czasownik	forma nieprzeszła forma przyszła czasownika BYĆ aglutynant czasownika BYĆ pseudoimiesłów rozkaźnik bezosobnik bezokolicznik imiesłów przysłówkowy współczesny imiesłów przysłówkowy uprzedni odśłownik imiesłów przymiotnikowy czynny imiesłów przymiotnikowy bierny	fin bedzie aglt praet impt imps inf pcon pant ger pact ppas	<i>jadam</i> <i>będę</i> <i>-śmy</i> <i>jadał</i> <i>jadaj</i> <i>jadano</i> <i>jadać</i> <i>jadając</i> <i>zjadłszy</i> <i>jadanie</i> <i>jadający</i> <i>jadany</i>
-----------	--	--	--

Analiza morfologiczna: klasy gramatyczne NKJP



CLARIN-PL

clarin-pl.eu

Czasownik typu WINIEN (forma teraźniejsza)		winien	<i>winna, powinni</i>
Predykatyw		pred	<i>trzeba, słychać</i>
Przyimek		prep	<i>pod, we</i>
spójnik	współrzędny	conj	<i>oraz, lub</i>
	podrzędny	comp	<i>że, aby</i>
Wykrzyknik		interj	<i>ach, psiakrew</i>
Burkinostka		burk	<i>omacku, trochu</i>
Kublik		qub	<i>nie, -ż, również</i>
Skrót		brev	<i>dr, np</i>
Ciało obce		xxx	<i>errare, humanum</i>
Interpunkcja		interp	<i>;, ., (,]</i>

NKJP (Przepiórkowski et al., 2012, str. 63)

Ujednoznacznianie morfo-syntaktyczne: **korpus**



- Narodowy Korpus Języka Polskiego (<http://nkjp.pl>)
 - zrównoważony: 250 mln wyrazów i ponad 300 mln tokenów
 - w tym aglutynaty oraz interpunkcja w tokenach
 - niezrównoważony: 1500 milionów słów i 1800 milionów tokenów
 - otwarty korpus ujednoznaczniiony ręcznie: ~1 mln. wyrazów (1 208 007)
 - Anotacja
 - morfosyntaktyczna
 - całość przetworzona tagerem Pantera o skuteczności około 93%
 - jednomilionowy: ujednoznaczniiony ręcznie w systemie 2+1
 - sensy słów: 100 słów o znaczeniach zgrubnych opisanych w słowniku
 - anotacja składniowa – wybrane wielowyrazowce i „grupy składniowe”
 - nazwy własne w anotacji hierarchicznej („jednostki nazewnicze”)

Ujednoznacznianie morfo-syntaktyczne: **korpus**



- Informacja morfosyntaktyczna
 - znaczniki strukturalne — pozycyjne
 - 35 klasy gramatyczne, 12 kategorii gramatycznych
 - 1021 różnych znaczników w korpusie jednomilionowym
- Znaczniki w Korpusie IPI PAN:
 - 4179 potencjalnych
 - 1642 w korpusie frekwencyjnym
 - 945 we wstępnym (przed ujednoznacznieniem 1149)

- Przykład:

Wiele mam nie przyszło.

<tok>

<orth>Wiele</orth>

<lex><base>wiele</base><ctag>num:pl:nom:m2</ctag></lex>

<lex><base>wiele</base><ctag>num:pl:nom:m3</ctag></lex>

<lex><base>wiele</base><ctag>num:pl:nom:f</ctag></lex>

<lex><base>wiele</base><ctag>num:pl:nom:n</ctag></lex>

<lex><base>wiele</base><ctag>num:pl:acc:m2</ctag></lex>

<lex><base>wiele</base><ctag>num:pl:acc:m3</ctag></lex>

<lex><base>wiele</base><ctag>num:pl:acc:f</ctag></lex>

<lex><base>wiele</base><ctag>num:pl:acc:n</ctag></lex>

</tok>

■ Przykład:

*Wiele **mam** nie przyszło.*

„Wiele” (8 analiz)

<tok>

<orth>mam</orth>

<lex><base>mama</base><ctag>subst:pl:gen:f</ctag></lex>

<lex><base>mamić</base><ctag>impt:sg:sec:imperf</ctag></lex>

<lex><base>mieć</base><ctag>fin:sg:pri:imperf</ctag></lex>

</tok>

„nie” (11 analiz)

„przyszło”

(1: <lex><base>przyjść</base><ctag>praet:sg:n:perf</ctag></lex>)

„.” (1: <lex><base>.</base><ctag>interp</ctag></lex>)

Wiele *mam* nie przyszło.

„Wiele” (<lex disamb="1"><base>wiele</base>
<ctag>num:pl:nom:f</ctag></lex>)

<tok>

<orth>mam</orth>

<lex disamb="1"><base>mama</base>

<ctag>subst:pl:gen:f</ctag></lex>

<lex><base>mamić</base><ctag>impt:sg:sec:imperf</ctag></lex>

<lex><base>mieć</base><ctag>fin:sg:pri:imperf</ctag></lex>

</tok>

„nie” (<lex disamb="1"><base>nie</base><ctag>qub</ctag></lex>)

„przyszło”

(<lex><base>przyjść</base><ctag>praet:sg:n:perf</ctag></lex>)

„.” (<lex><base>.</base><ctag>interp</ctag></lex>)

Ujednoznacznianie morfo-syntaktyczne: **podejścia**



- Metody lingwistyczne (regułowe)
 - ręcznie napisane, odwołujące się do różnorodnych cech
- Metody statystyczne:
 - oparte na ukrytych modelach Markowa,
 - oparte na modelowaniu maksimum entropii
 - warunkowych polach losowych Conditional Random Fields (CRF),
 - np. WCRFT1, WCRFT2, Concraft
- Metody regułowe (nieoparte jawnie na modelach probabilistycznych):
 - oparte na pamięci (ang. *Memory Based Learning*, *Case Based Learning*),
 - WMBT (oparty na TiMBL)
 - oparte na transformacjach, np. Pantera
- Maszynowe uczenie
 - sieci neuronowe – perceptron wielowarstwowy, np. MorphoDiTa-pl
 - głębokie sieci neuronowe (Deep Neural Networks),
 - głównie sekwencyjne, dwukierunkowe: bLSTM (bidirectional LSTM),
 - np. Toyger, CMC Tagger

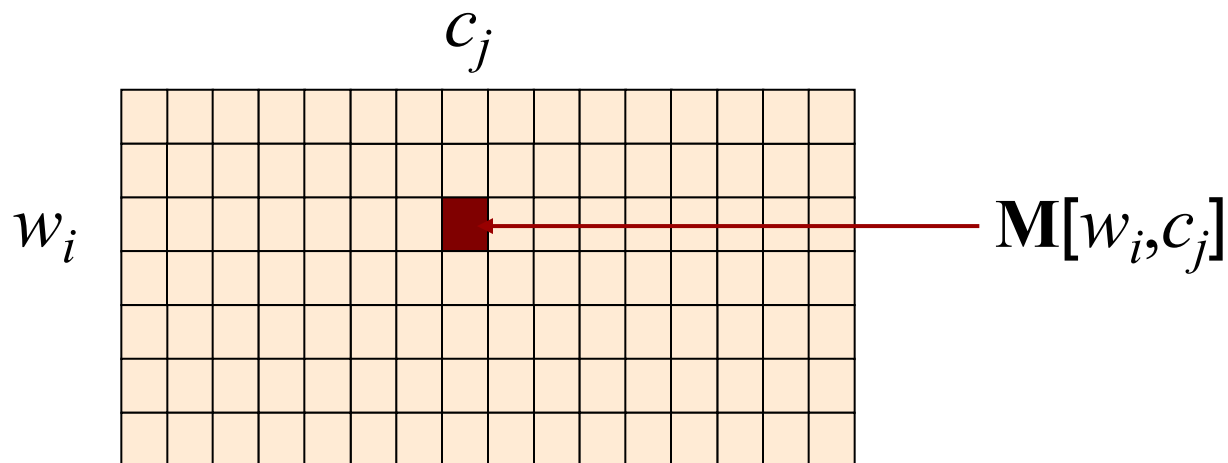
Reprezentacja znaczeń leksykalnych



- Postulaty znaczeniowe — aksjomaty logiczne
- Reprezentacja dekompozycyjna
 - wyrażenia odwołujące się do znaczeń atomowych
 - zalety: precyzja, duża siła ekspresji
 - wady: pracochłonna, problem spójności
- **Leksykalna sieć semantyczna**
 - opisująca **leksykalne relacje znaczeniowe** pomiędzy poszczególnymi **jednostkami leksykalnymi**
 - jednostka leksykalna — jedno lub wielowyrazowy leksem, reprezentowany jako para $\langle \textit{lema}, \textit{sens} \rangle$, np. *zamek 1*,
 - zalety: efektywność budowania, generalizacja
 - wady: **opis nie wprost**, ograniczona siła ekspresji
- Reprezentacja wektorowa
 - znaczenie jako punkt w wielowymiarowej przestrzeni
 - tematyczne przestrzenie $w_{m1} = \langle t_1, \dots, t_n \rangle$, t_i - temat/dziedzina
 - dystrybucyjne przestrzenie $w_{m1} = \langle k_1, \dots, k_m \rangle$, k_i - kontekst użycia

Reprezentacja wektorowa

- Wzorzec dystrybucji → konteksty użycia
- Macierz (współwystępowania) koincydencji



gdzie:

- w_i - lemat
- c_j - cecha charakteryzująca kontekst
- $\mathbf{M}[w_i, c_j]$ – częstość współwystąpienia z określoną cechą

Model dystrybucyjny: przykład



- Przykładowy korpus

(Landauer, Foltz i Laham, 1998)

- tytuły krótkich artykułów technicznych z dziedziny interakcji człowiek-komputer i teorii grafów
 - podkreślone są słowa analizowane dalej

c1: *Human machine interface* for ABC computer applications

c2: A survey of user opinion of computer system response time

c3: The *EPS* user interface management system

c4: System and human system engineering testing of *EPS*

c5: Relation of user perceived response time to error measurement

m1: The generation of random, binary, ordered trees

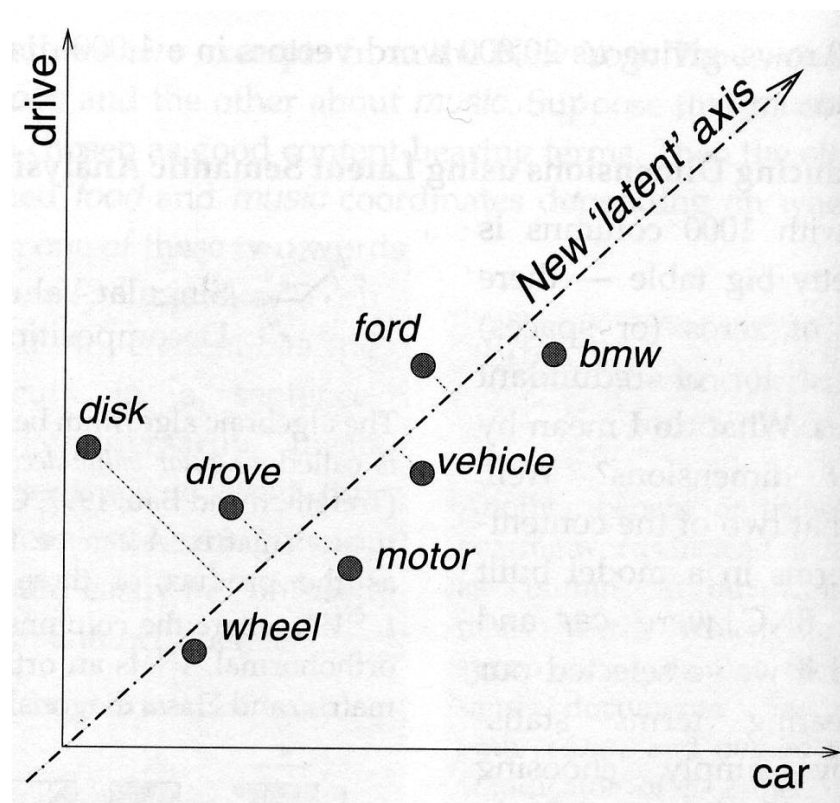
m2: The intersection graph of paths in trees

m3: Graph minors IV: Widths of trees and well-quasi-ordering

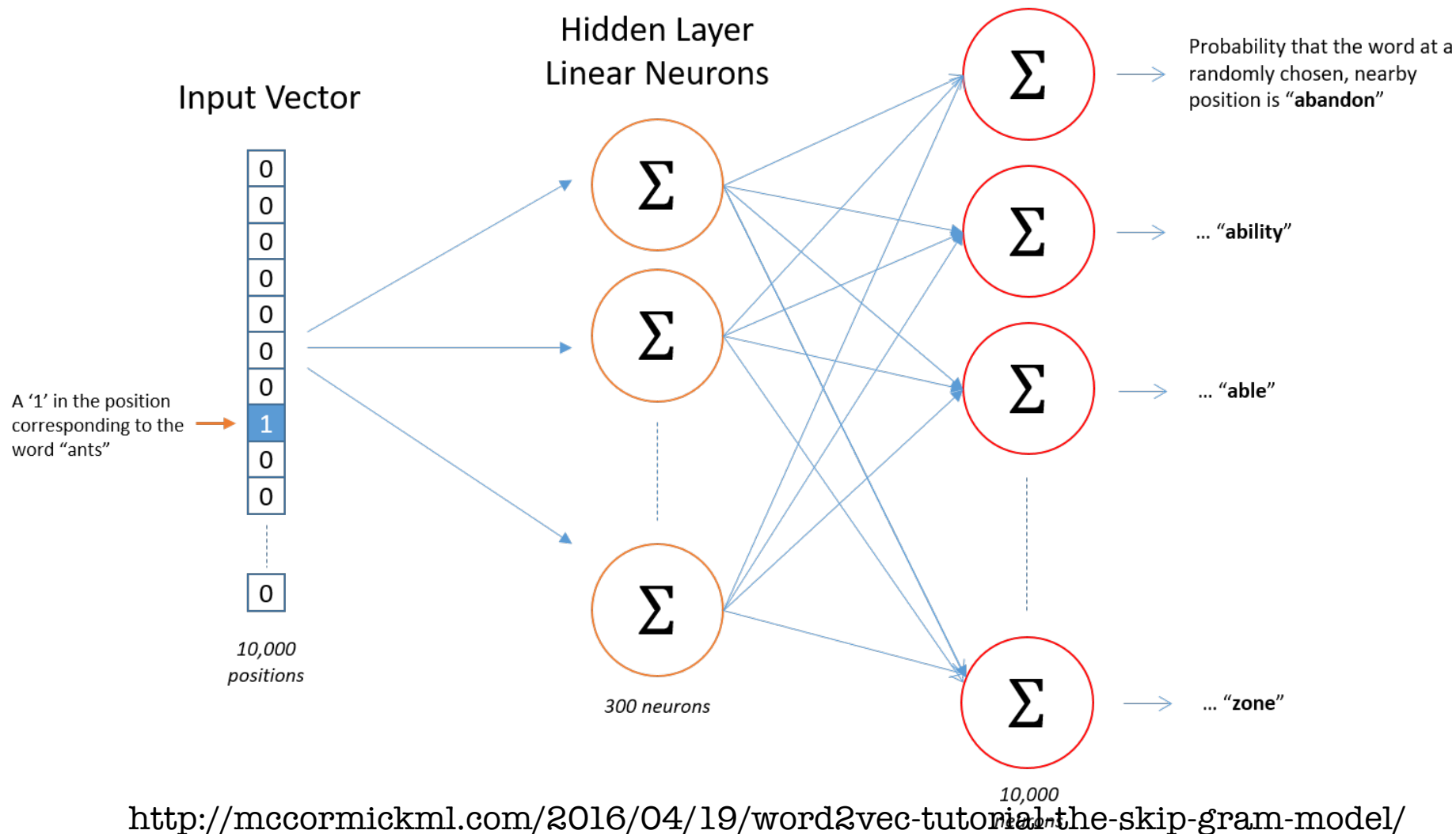
m4: Graph minors: A survey

Model dystrybucyjny: podobieństwo

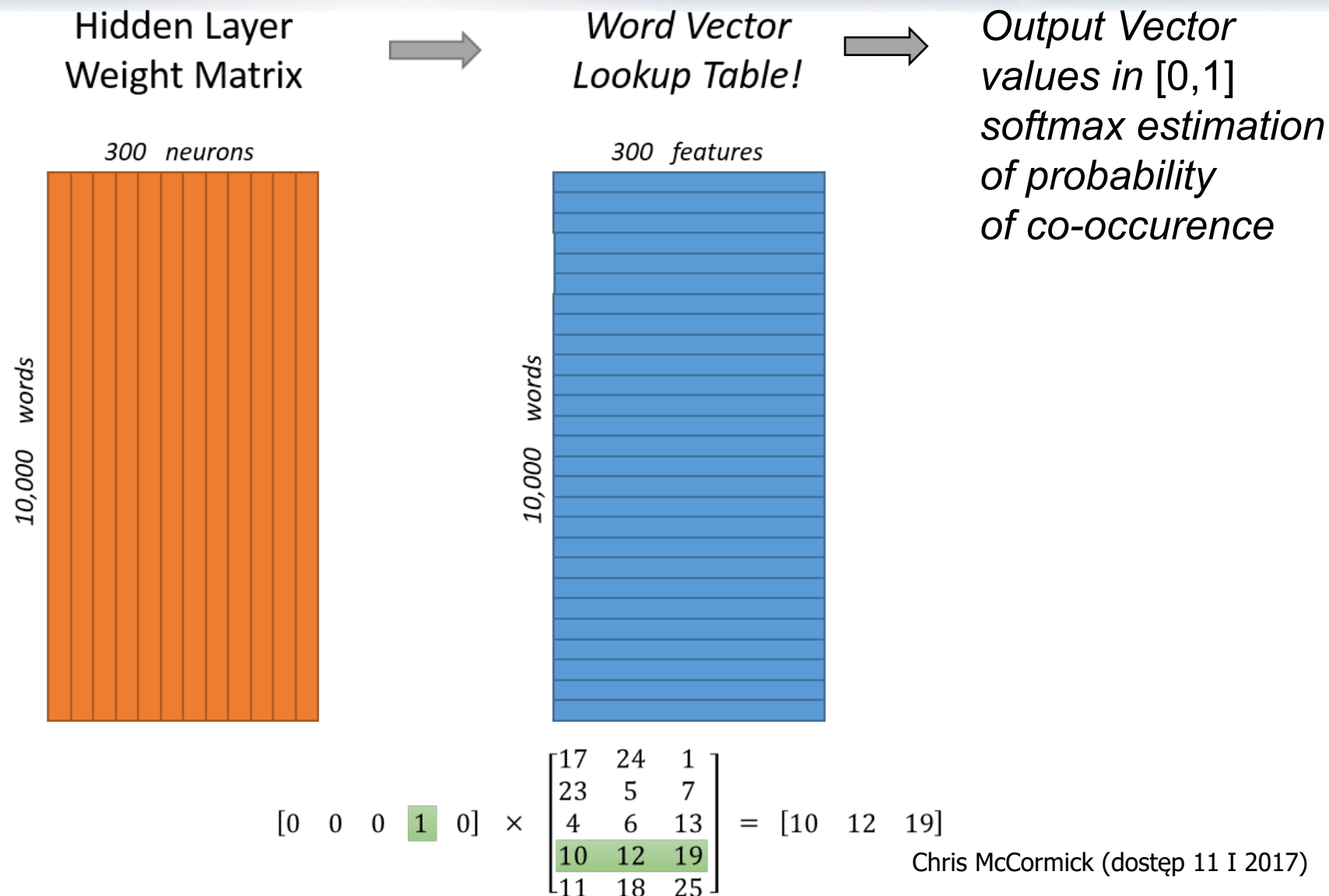
- Efekt uboczny:
 - następuje uogólnienie podobieństw między słowami, np.
 - $p(\text{human}, \text{user}) = -0,38$
 $p(\text{human}, \text{minors}) = -0.29$
 - po redukcji do 2 wymiarów:
 $p(\text{human}, \text{user}) = 0,94$
 $p(\text{human}, \text{minors}) = -0.83$ (!)
- Przykład: wzmacnianie podobieństwa między słowami po wprowadzeniu nowej osi (Widdows, 2004)



Word2Vec: SkipGram na lekko



SkipGram – działanie



Chris McCormick (dostęp 11 I 2017)

pies *dog*

piesek *pooch* (*deminutive*) ←

pieski *canine/wretched* ←

czworonóg *tetrapod* ←

szczeniak *puppy* ←

psiak *puppy/pooch* (dem.) ←

czworonogi *tetrapods* ←

doberman *doberman* ←

kundel *mongrel* ←

psisko *pooch* (*augmentative*) ←

← co-hyponym

← holonym

Modele wektorów osadzeń: klasyfikacja



- Klasyfikacja (za BERT, orientacyjna)
- Bezkontekstowe (modele poziomu wyrazowego)
 - ten sam wektor w różnych kontekstach
 - wyrazów lub lematów, np. word2vec, GloVec
 - znaczeń, np. AutoExtend
 - modele subwyrazowe, np. fastText
- Kontekstowe
 - strukturalne
 - oparte na uśrednianiu przez konwolucję: USE(DAN)
 - jednokierunkowe (*jeżeli takie są*)
 - dwukierunkowe
 - sekwencyjnie połączone: Flair, ELMo, USE(Transformer)
 - w pełni połączone: BERT (Bidirectional Encoder Representations from Transformers)
- Kros-językowe
 - oparte na rzutowaniu jednej przestrzeni na drugą
 - oparte na schemacie: koder – dekodek, np. Laser (na BPE)

Ujednoznacznianie znaczeń leksykalnych: wprowadzenie



- Przykład wprowadzający:

Pokazał Zosi zamek.

Pokazał Zosi zamek na wysokiej górze.

*Pokazał Zosi zamek i wyjaśnił, że jak będzie
ciągnąć to tylko bardziej się zablokuje.*

*Pokazał Zosi zamek i podkreślił, że dopóki jest
zamknięty, to nie ma po co szarpać za klamkę,
bo drzwi i tak się nie otworzą.*

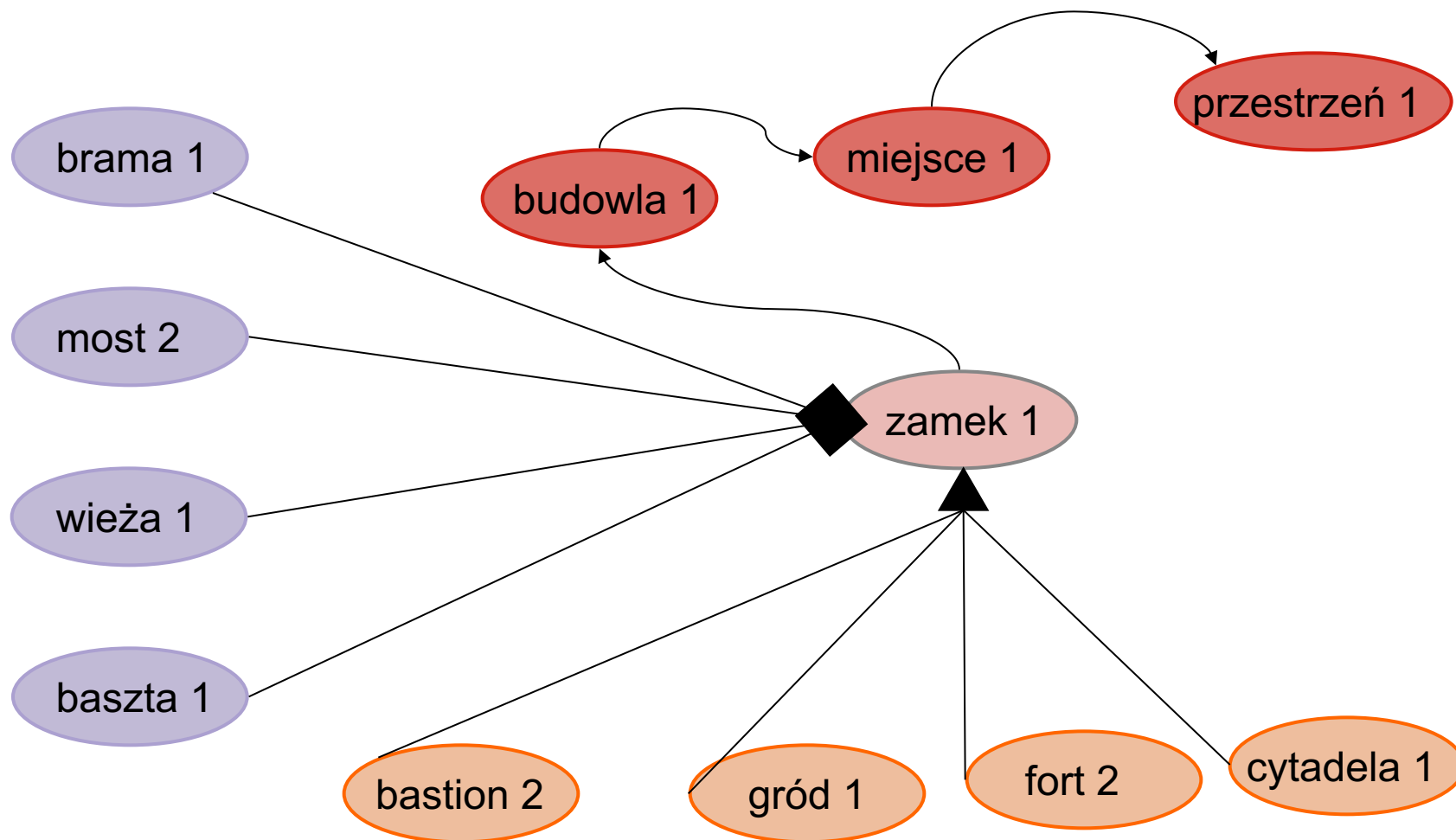
*Pokazał Zosi zamek, bo wcześniej nie rozumiała,
dlaczego łatwiej wtedy strzelić gola.*

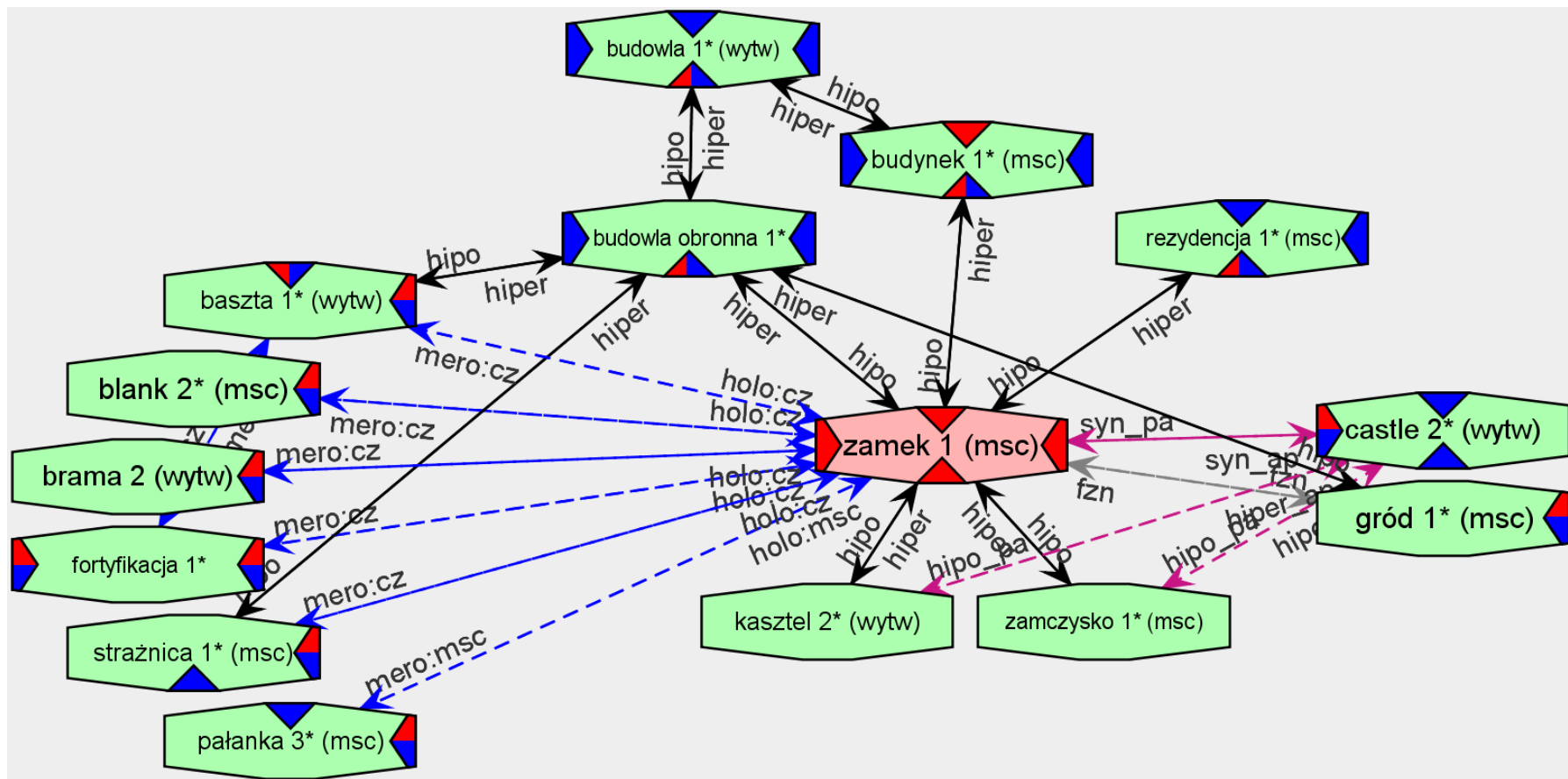
Ujednoznacznianie znaczeń leksykalnych: [wprowadzenie](#)



- Słowa o różnym znaczeniu w różnym kontekście:
 - homonimy – ten sam lemat, ale zasadniczo odmienne pochodzenie różnych znaczeń, np. *zamek*
 - lematy polisemiczne, np. *linia* – znaczenia wyodrębnione na skutek różnych procesów
 - wieloznaczność jest częstym zjawiskiem, np. *agent*, *akcja*, *automat*, , *jądro*, *komora*, *ognisko*, *punkt*, *sprawa*, itd.
- Ujednoznacznienie znaczenia w kontekście jest niezwykle ważnym etapem przetwarzania

Sieć leksykalnych relacji semantycznych

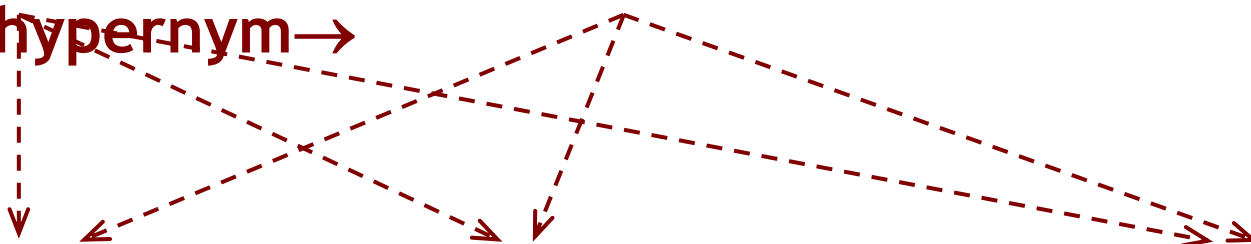




- Example

{wzór 1 `paragon', wzorzec 2 `pattern', ...}

—**hypernym**→



{idol 1 `idol', bożyszcze 1 `~idol', gwiazdora 1 ~ `star'}

- Synset as a notational convention

- for a group of lexical units sharing certain **constitutive relations**
- What are wordnet **constitutive relations**?
- Are relations enough to define synsets?

- **Relacje synsetów**

- Hiperonimia/hiponimia

- Holonomia/meronimia

- części
- porcji
- miejsca
- elementu
- materiału
- taksonomiczna

- Typ/instancja

- Bliskoznaczność

- Synonimia międzyrejestrów

- Fuzzynimia

- **Relacje jednostek leksykalnych**

- Antonimia

- Właściwa
- Komplementarna

- Konwersja

- Synonimia międzyparadygmatyczna

- Rola

- Np. agens / subiekt, narzędzie, wytwór / rezultat, miejsce, pacjens / obiekt, czas, podtyp nieokreślony,

- Żeńskość

- Nacechowanie

- Istota młoda
- Deminutywność
- Ekspresywność i augmentatywność

Słownosieć: relacje łączące synsety



■ Meronimia

- podtypy: część, element kolekcji, materiał, porcja,
{iskra 1, skierka 1, skra 1, iskierka 1} —m.cz.→ {ogień 1}
{kula 4, nabój 2} —m.e.k.→ {amunicja 1}
{drewno 1, drzewo 1} —m.mat.→ {stolarka 2}
{termin 1, data 1} —m.msc.→ {czas 1}
{blacha 1} —m.p.→ {metal 2}

■ Część testów dla potypów

- meronimia CZĘŚCI: X jest CZĘŚCIĄ Y-a,
- meronimia PORCJI: X jest PORCJĄ Y-a,
- meronimia MIEJSCA: X jest MIEJSCEM w (na) Y-u,
- meronimia ELEMENTU: X jest ELEMENTEM zbioru (grupy) Y,
- meronimia MATERIAŁU: X jest MATERIAŁEM, z którego zrobione jest Y.

Słownosieć: relacje łączące synsety



■ Holonimia

- podtypy: identyczne jak dla meronimii
 - $\{kula\ 4, nabój\ 2\} \rightarrow \text{h.cz.} \rightarrow \{łuska\ 1, gilza\ 1\}$
 - $\{bydło\ 2\} \rightarrow \text{h.e.k.} \rightarrow \{krowa\ 1\}$
 - $\{deszczowiec\ 1\} \rightarrow \text{h.m.} \rightarrow \{tkanina\ 1\}$
- Nie są relacjami automatycznie odwrotnymi
 - np. *szprycha* —m.cz.—> *koło*
 - np. *rodzynka* jest elementem *bakalii* i *bakalie* zawsze składają się z *rodzynek* (relacja odwrotna), ale już *kora* jest częścią *drzewa*, choć nie wszystkie *drzewa* mają *korę*.

■ Bliskoznaczność

- np. $\{adorator\ 1\} \rightarrow \text{—blsk.—} \rightarrow \{załotnik\ 1\}$
- $\rightarrow \text{—blsk.—} \rightarrow \{absztyfikant\ 1\}$
- $\{apollo\ 2\} \rightarrow \text{—blsk.—} \rightarrow \{przystojniak\ 1\}$

■ Page $P^{(new)} = c \cdot M \cdot P^{(old)} + (1 - c) \cdot v$

■ G : graph with N nodes $n_1, \dots, n_N$,

■ d_i : the degree of the node n_i ,

■ M : matrix with $N \times N$ dimension:

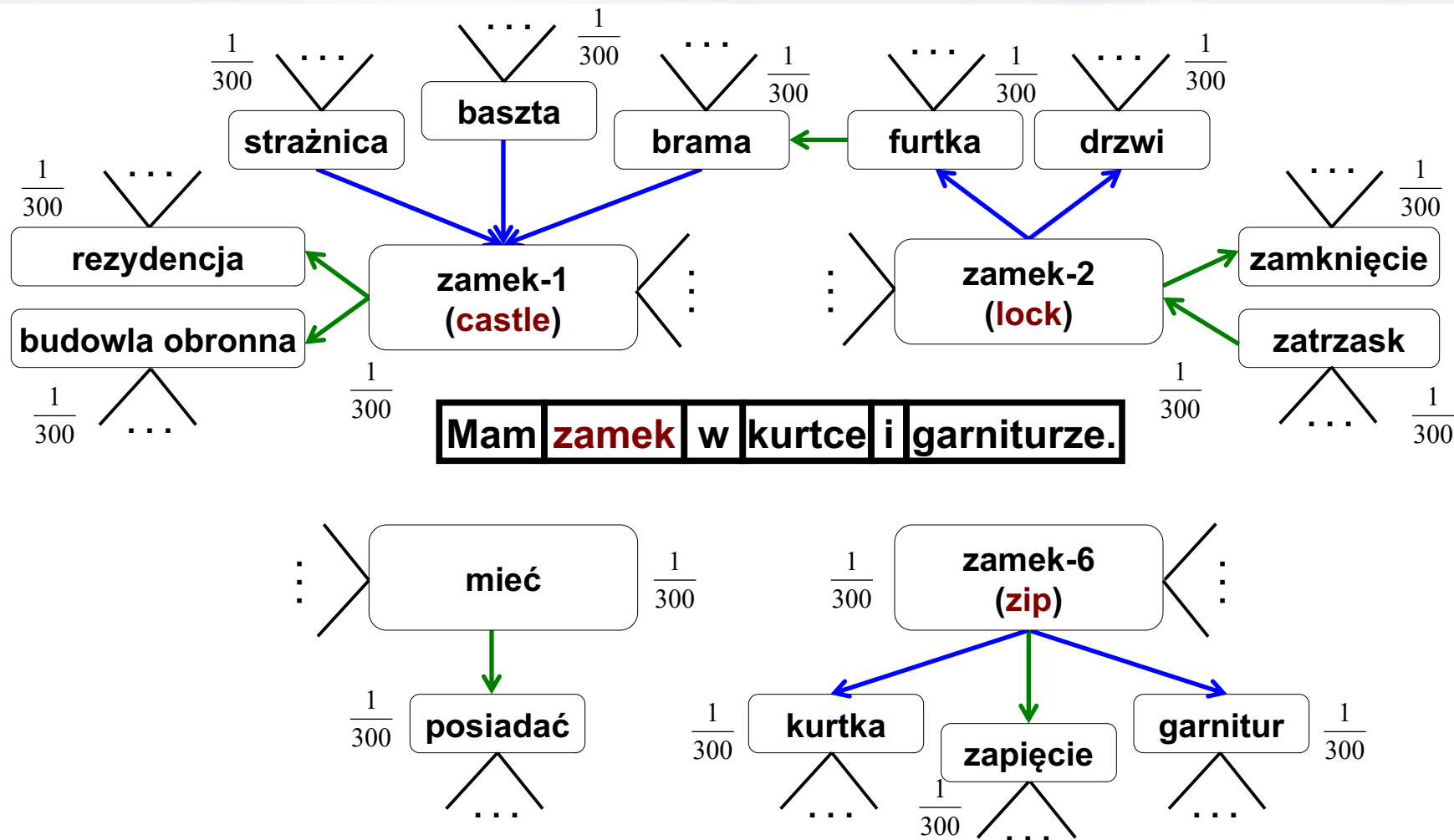
$$M_{ji} = \begin{cases} \frac{1}{d_i} & \text{- if exist an edge between } n_i \text{ and } n_j \text{ nodes} \\ 0 & \text{- otherwise} \end{cases}$$

■ c : damping factor $\langle 0, 1 \rangle$

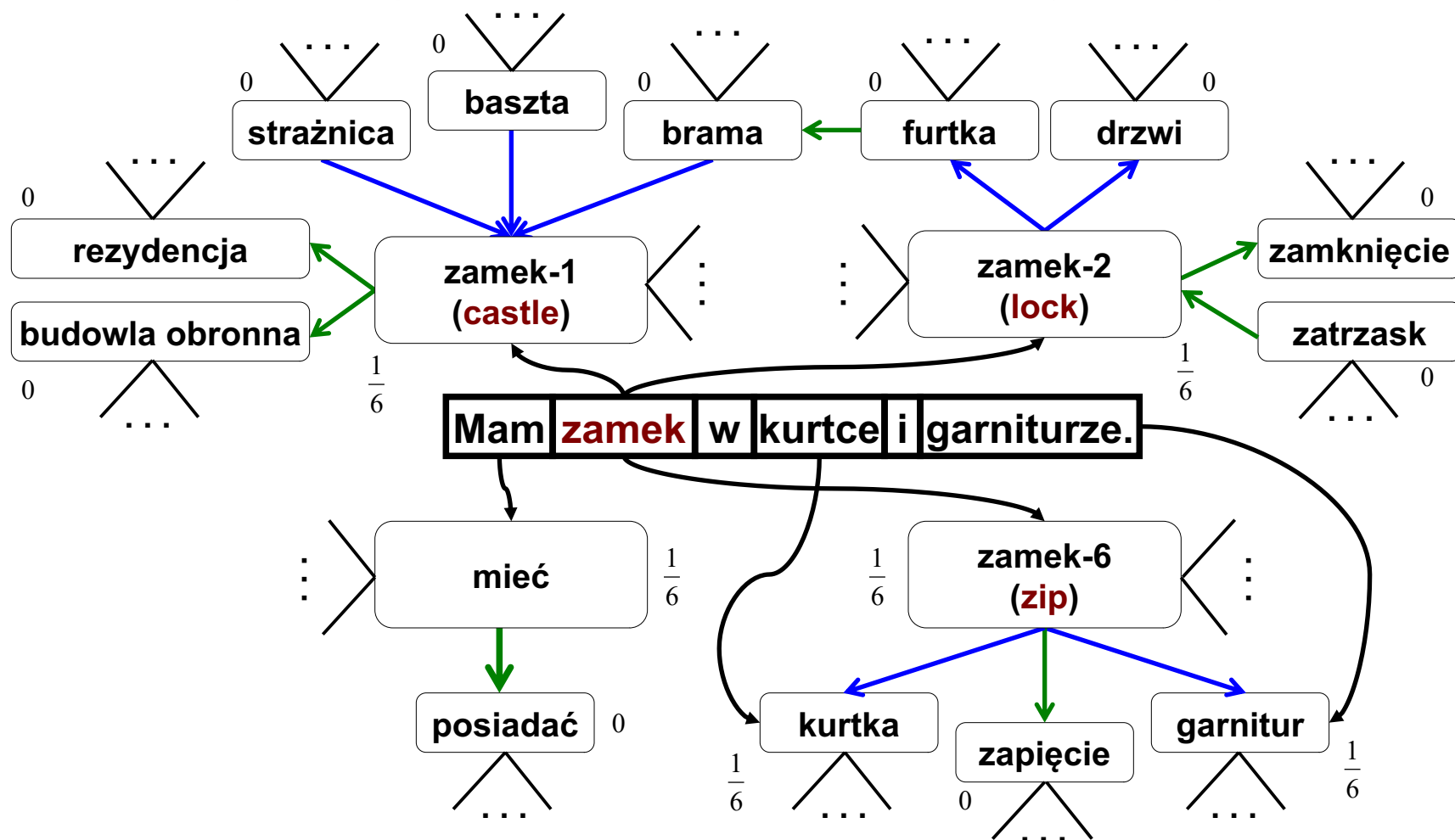
■ v : vector $N \times 1$

$$v_i = \frac{1}{N}$$

Statyczny PageRank w WSD: przykład



Spersonalizowany PageRank w WSD: przykład



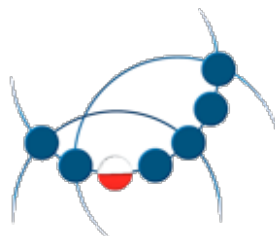
CLARIN

Common Language Resources and Technology Infrastructure



Dziękuję bardzo za uwagę

CLARIN-PL
Common Language Resources and Technology Infrastructure



CENTRUM TECHNOLOGII JEZYKOWYCH
CLARIN-PL



www.clarin-pl.eu

www.clarin.eu