Elementy metod grafowych i kombinatorycznych w badaniach frazeologicznych

Piotr Pęzik[1] Bartosz Zieliński[2]

[1]Pracownia Językoznawstwa Korpusowego i Komputerowego,
Instytut Anglistyki,
Uniwersytet Łódzki

[2]Katedra Informatyki,
Wydział Fizyki i Informatyki Stosowanej,
Uniwersytet Łódzki

ZIL @ IPI PAN 2014/05/12

Jaka frazeologia?

- „Nie ma jednej frazeologii” (Lewicki, 1976).
- Frazeologia dystrybucyjna, neofirthiańska, *strong tea* vs. *powerful car*, tendencja idiomatyczna (Sinclair, 1996), *lexicality*
- (Korpusowa) “frazeologia nadawcy” (Lewicki, 1976), frazematyka (Chlebda, 2003)
- Reprodukcja podstawowym kryterium formuliczności
- Por. niekompozycyjność w tradycyjnym ujęciu (czysta i figuratywna idiomatyczność)

Formuliczność i jej psycholingwistyczne kryteria identyfikacji

- Czas produkcji, płynność, tempo, dystrybucja pauz w języku mówionym (Pawley & Syder, 1983)
- Akwizycja, utrata języka (Wray, 2002)
- Śledzenie ruchu gałek ocznych (Siyanova-Chanturia, Conklin, & Schmitt, 2011)
- Obrazowanie mózgu (Tremblay & Bayeen, 2010)

HASK, czyli słowniki kombinacji wyrazowych

- Dla NKJP: 106 723 "haseł" zawierających łącznie 5 636 857 kombinacji.
- Niektóre reguły są symetryczne -> kombinacje częściowo się powtarzają.
- Szczegóły: (Pęzik, 2013).

Node	F	Type	Positions	Colls	Coll	F	T	JD
mocny	13018	adj-noun	[-1|1|-2|2]	829	strona	891	27.12	0.92
mocny	13018	adj-noun	[-1|1|-2|2]	829	pozycja	401	19.25	0.89
mocny	13018	adj-noun	[-1|1|-2|2]	829	słowo	325	15.79	0.93
mocny	13018	adj-noun	[-1|1|-2|2]	829	wrażenie	274	15.76	0.91
mocny	13018	adj-noun	[-1|1|-2|2]	829	uderzenie	199	13.88	0.9
mocny	13018	adj-noun	[-1|1|-2|2]	829	kawa	189	13.4	0.89
mocny	13018	adj-noun	[-1|1|-2|2]	829	trunek	159	12.55	0.89
mocny	13018	adj-noun	[-1|1|-2|2]	829	argument	146	11.46	0.9
mocny	13018	adj-noun	[-1|1|-2|2]	829	herbata	127	10.93	0.86
mocny	13018	adj-noun	[-1|1|-2|2]	829	coś	210	10	0.89

Table: Zob. pelcra.pl/hask_pl, pelcra.pl/hask_en.

Poziomy aktywacji w rozpoznawaniu słów

- Pilotażowe eksperymenty psycholingwistyczne

przymiotnik	wiatr
	pozycja
	wpływ
	skład
	wola
	emocja
	ból
	konkurencja
	więź
	strzał

przymiotnik	uderzenie
	akcent
	strzał
	uścisk
	postanowienie
	głos
	argument
	punkt
	cios
	makijaż

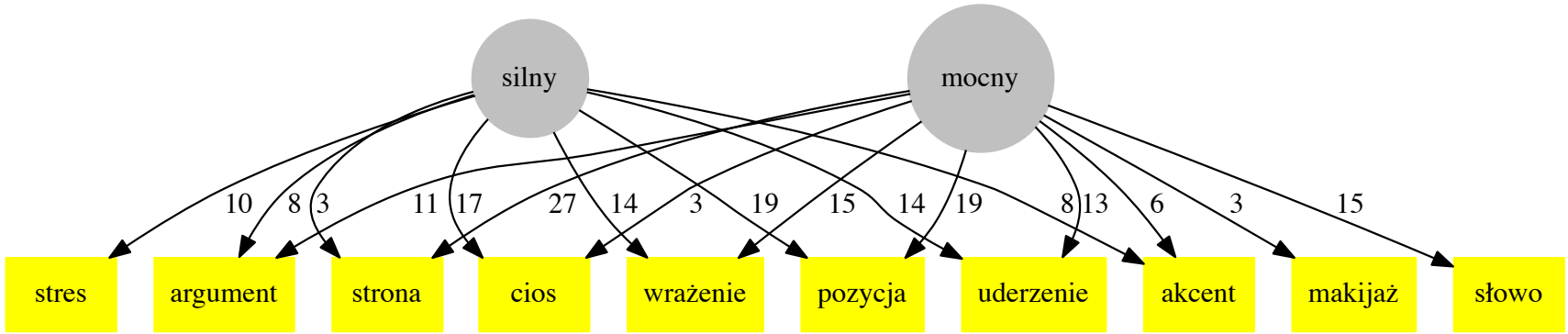
Poziomy aktywacji w rozpoznawaniu słów

- 38/39 badanych wymienia **silny** dla 1. zbioru
- 39/39 badanych wymienia **mocny** dla 2. zbioru

siln(y/a/e)	mocn(y/a/e)
wiatr	uderzenie
pozycja	akcent
wpływ	strzał
skład	uścisk
wola	postanowienie
emocja	głos
ból	argument
konkurencja	punkt
więź	cios
strzał	makijaż

Silny vs. Mocny

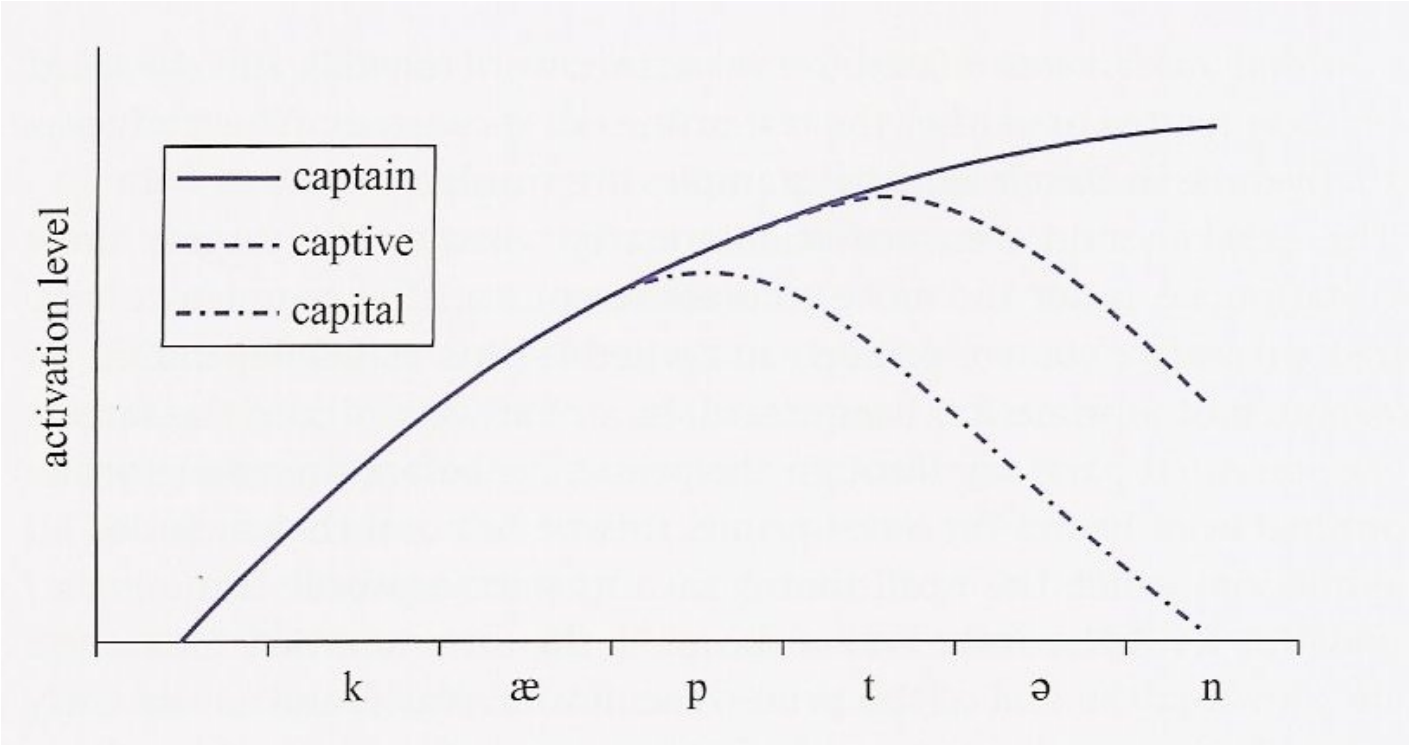
- Wyniki są interesujące, zważywszy na stopniowalność rozróżnienia między frazemem a syntagmą.
- *Mocne słowa* : *silne słowa* 284:5 w zrównoważonej części NKJP
- *Silny stres* : *mocny stres* 101:1, *mocny makijaż* : *silny makijaż* 39:1



Frazeologiczne punkty unikalności

- Niektóre podzbiory kolokatów są unikalne (globalnie w całym grafie słownika) na przyjętym poziomie utrwalenia dla danych ośrodków kolokacji.
- Przykładem szczególnych przypadków unikalnych, jednoelementowych zbiorów są składniki tzw. "cranberry collocations", czyli frazemy z **wyrazami uwięzionymi** (por. frazeologia.pl, Piotr Müldner-Nieckowski), np. *put the **kibosh** on*, *zbić kogoś z **pantałyku***
- Czy i w jaki sposób różne kombinacje (docelowo permutacje) kolokatów (primes) odpowiadają różnym poziomom aktywacji w procesie rozpoznawania ośrodków kolokacji (targets)?

Poziomy aktywacji w rozpoznawaniu słów



(Warren, 2013)

- Teoria kohorty
- Punkty unikalności (jedyne pasujące słowo), izolacji (wstępne rozpoznanie słowa), rozpoznania (80% pewność), całkowitej akceptacji (100% pewność) (Field, 2004).

Torowanie leksykalne, składniowe i frazeologiczne

- Badania efektu torowania, również z wykorzystaniem korpusowych kryteriów dystrybucyjnych.
- Również w NLP – warsztat CogALex-IV@COLING 2014

$$\text{whiskey} \begin{cases} gin \\ drink \\ bottle \\ soda \end{cases}$$

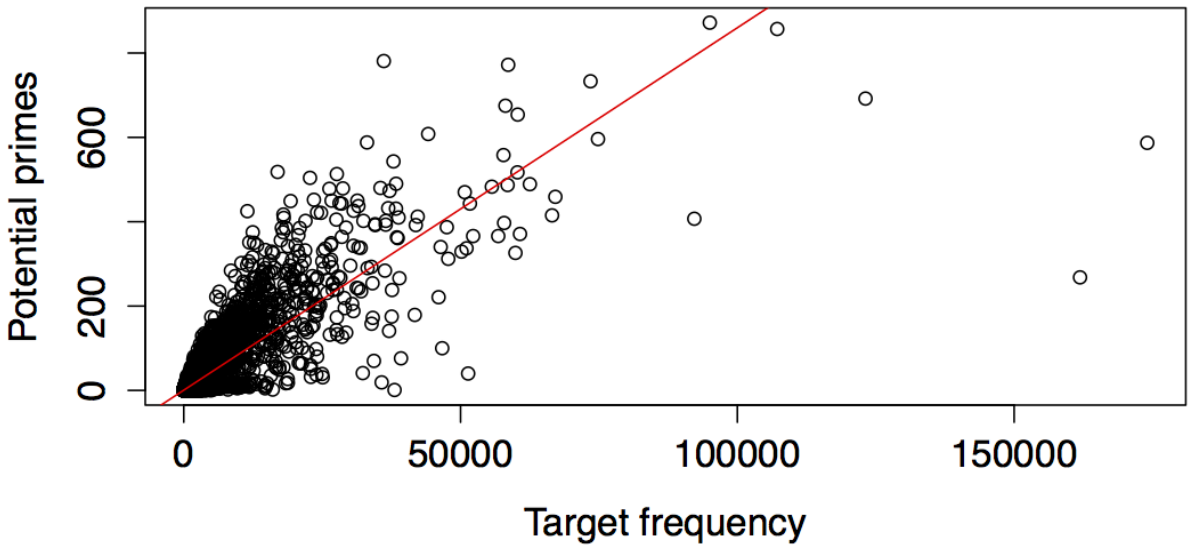
- Zawężenie badań do zaświadczonych w korpusach kontekstów frazeologicznych.

$$\text{whiskey} \begin{cases} Irish \\ strong \\ Scotch \\ old \end{cases}$$

Możliwe wyniki badań psycholingwistycznych

- Badania podstawowe: kompozycyjność/formuliczność języków.
- W jakim stopniu punkty unikalności określane na podstawie korpusów referencyjnych znajdują potwierdzenie w eksperymentach psycholingwistycznych?
- Weryfikacja korpusowych metod identyfikacji jednostek frazeologicznych. Które miary, lub które kombinacje miar skutkują wyłonieniem wyrazów najsilniej naprowadzających?
- Co decyduje o aktywacji? Częstość i rozproszenie czy uwięzienie?

Problem kombinatoryczno-grafowy



- Od czego zacząć? Słownik zawiera miliony kombinacji...
- Minimalne zbiory skutkujące punktami unikalności wydają się najprostsze do analizy (np. wyrazy uwięzione).
- Przewidywana konieczność generowania permutacji na późniejszym etapie.
- Problem: jak skutecznie znaleźć możliwie niewielkie zbiory?

Elementy metod grafowych w badaniach frazeologicznych

- Inne zastosowania modeli grafowych w problemach frazeologicznych.
- Sieci kolokacyjne: (Mehler, 2007)
- Inne problemy: weryfikacja poziomu synonimii metodami grafowymi
- Spójność synsetów w tezaurusach i słowosieciach

Dwa problemy NP-trudne

Problem "Hitting Set"

Niech $\{S_i\}_{i \in I}$ będzie skończoną rodziną zbiorów. Należy znaleźć najmniejszy zbiór C taki że $C \cap S_i \neq \emptyset$ dla każdego $i \in I$.

Problem pokrycia zbioru

Niech $\{S_i\}_{i \in I}$ będzie skończoną rodziną zbiorów. Należy znaleźć najmniejszy podzbiór $J \subseteq I$ taki że $\bigcup_{i \in I} S_i = \bigcup_{j \in J} S_j$.

Szczególne przypadki problemu grafowego

Niech $(U, V, E \subseteq U \times V)$ będzie grafem dwudzielnym. Znaleźć minimalny podzbiór $C \subseteq U$ taki że dla każdego $v \in V$ istnieje $u \in C$ takie że $(u, v) \in E$.

Problem "Hitting Set"

- $U := \bigcup_{i \in I} S_i$, $V := I$,
- $(s, i) \in E$ wtedy i tylko wtedy gdy $s \in S_i$.

Set cover

- $U := I$, $V := \bigcup_{i \in I} S_i$,
- $(i, s) \in E$ wtedy i tylko wtedy gdy $s \in S_i$.

Dodatkowa Notacja

Niech $(U, V, E \subseteq U \times V)$ będzie grafem dwudzielnym.

Wprowadzamy następującą notację:

- $E(u) := \{v \in V \mid (u,v) \in E\}$ dla każdego $u \in U$.
  Dla $A \subseteq U$ definiujemy $E(A) := \bigcup_{u \in A} E(u)$.
- $E^{-1}(v) := \{u \in U \mid (u,v) \in E\}$ dla każdego $v \in V$.
  Dla $B \subseteq V$ definiujemy $E^{-1}(B) := \bigcup_{v \in B} E^{-1}(v)$.

Rozwiązania Przybliżone: Algorytm Zachłanny

Problem przeformułowany

Znaleźć dla grafu $(U, V, E \subseteq U \times V)$ minimalny $C \subseteq U$ t.ż. $E(C) = V$.

Algorytm zachłanny

- Zaczynamy od $C := \emptyset$.
- Dopóki $E(C) \neq V$ zwiększamy $C := C \cup \{u\}$ (u maksymalizuje $|E(u) \setminus E(C)|$).

Gwarancje algorytmu zachłannego

Niech C^* będzie rozwiązaniem optymalnym, a C znalezionym przez algorytm zachłanny.

- $|C| \leq H_k |C^*|$, gdzie $H_k = \sum_{i=1}^{k} 1/i \simeq \ln k$, $k := \max\{|E(u)| \mid u \in U\}$.
- $|C|/|C^*| \leq \ln n - \ln\ln n + O(1)$, gdzie $n := |V|$.

Granice Przybliżania

Istnieją lepsze algorytmy przybliżone niż zachłanny, ale ...

Twierdzenie

Minimalnego pokrycia zbioru nie można przybliżać w czasie wielomianowym ze stosunkiem $|C^*|/|C|$ lepszym niż $(1 - o(1)) \ln n$.

Źródło

Feige, Uriel. *"A threshold of ln n for approximating set cover."* JACM 45(4) (1998), 634-652.

Problem korpusowych punktów minimalności dla kolokacji leksykalnych

Graf kolokacji

- U — zbiór ośrodków kolokacji.
- V — zbiór kolokatów.
- $(u, v) \in E$ — słowo u pojawia się w korpusowym słowniku kolokacji (z częstością powyżej granicznej) wraz z kolokatem v.

Problem

Dla każdego ośrodka kolokacji $u \in U$ znaleźć (o ile istnieje) najmniejszy zbiór kolokatów $C_u \subseteq V$ jednoznacznie wyznaczający u, tzn. taki że

$$\forall u' \in U \,.\, C_u \subseteq E(u') \Rightarrow u = u'.$$

Związek z "Set cover problem"

Niech $u \in U$. Zdefiniujmy

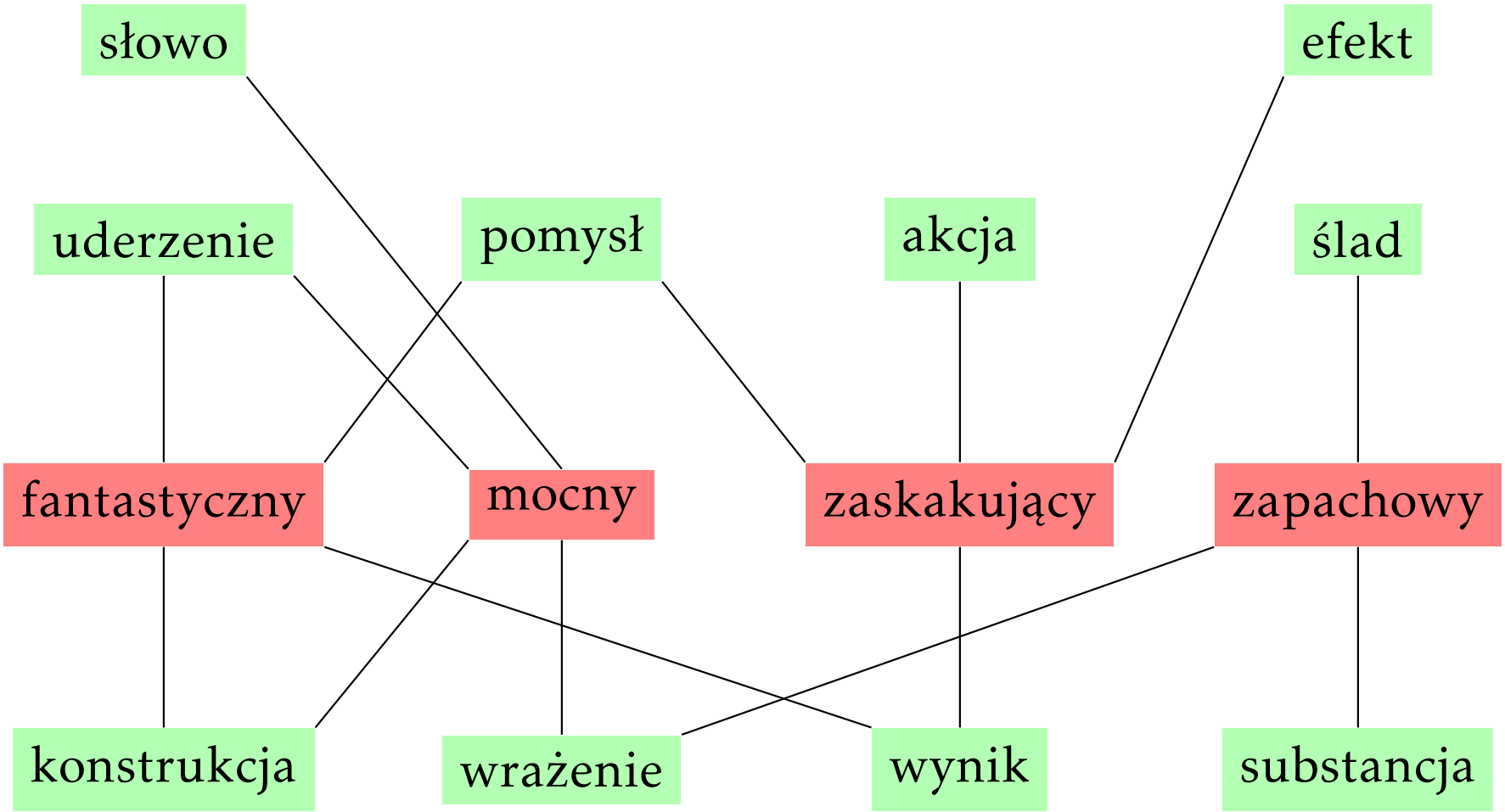
Graf dwudzielny $\mathcal{G}_u' := (U_u, V_u, E_u)$, gdzie

- $U_u := E(u)$, $V_u := E^{-1}(E(u)) \setminus \{u\}$,
- $E_u := \{(v,u) \in U_u \times V_u \mid (u,v) \notin E\}$.

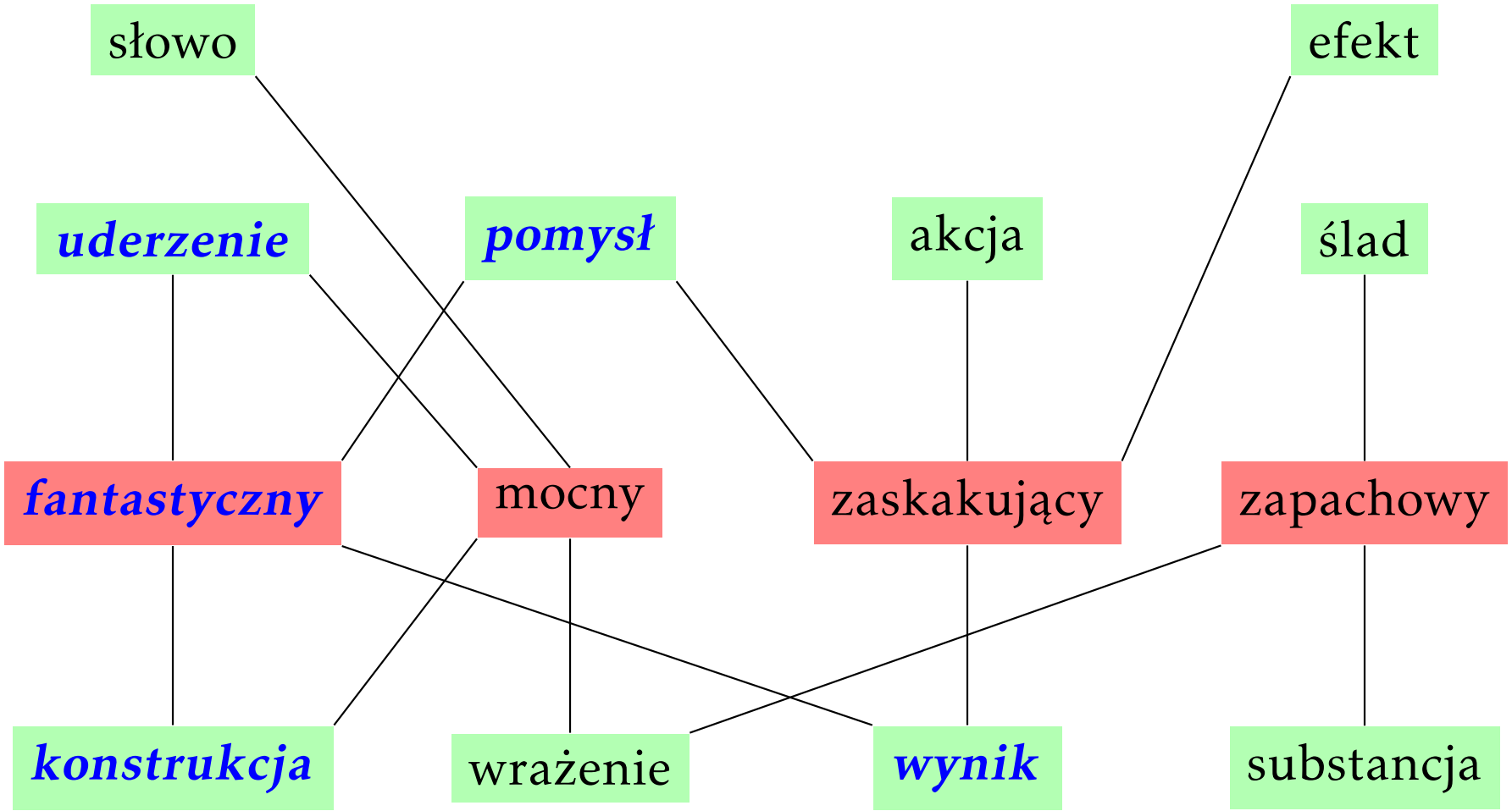
Obserwacja

Dla ośrodka kolokacji $u \in U$ znalezienie najmniejszego unikalnego zbioru kolokatów jest tożsame z rozwiązaniem grafowego uogólnienia problemu "Set cover" dla grafu $\mathcal{G}_u'$.

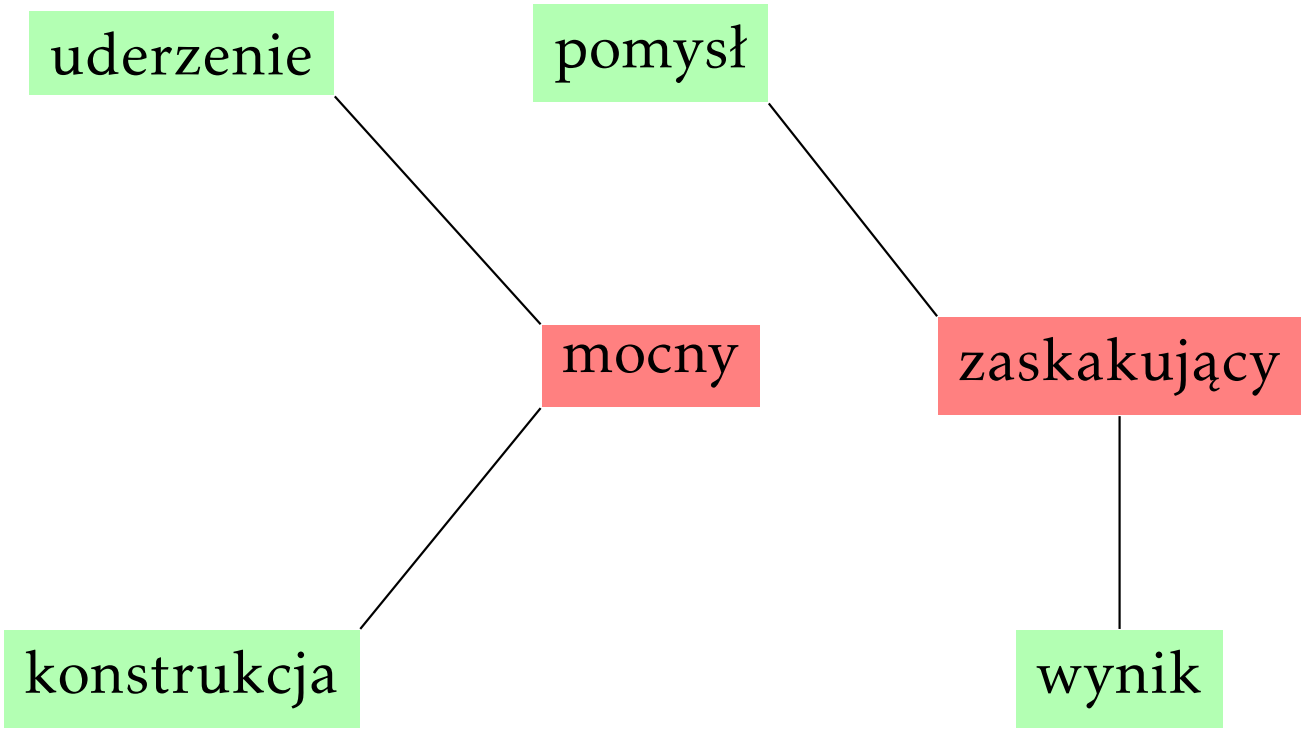
Przykład grafu kolokacji



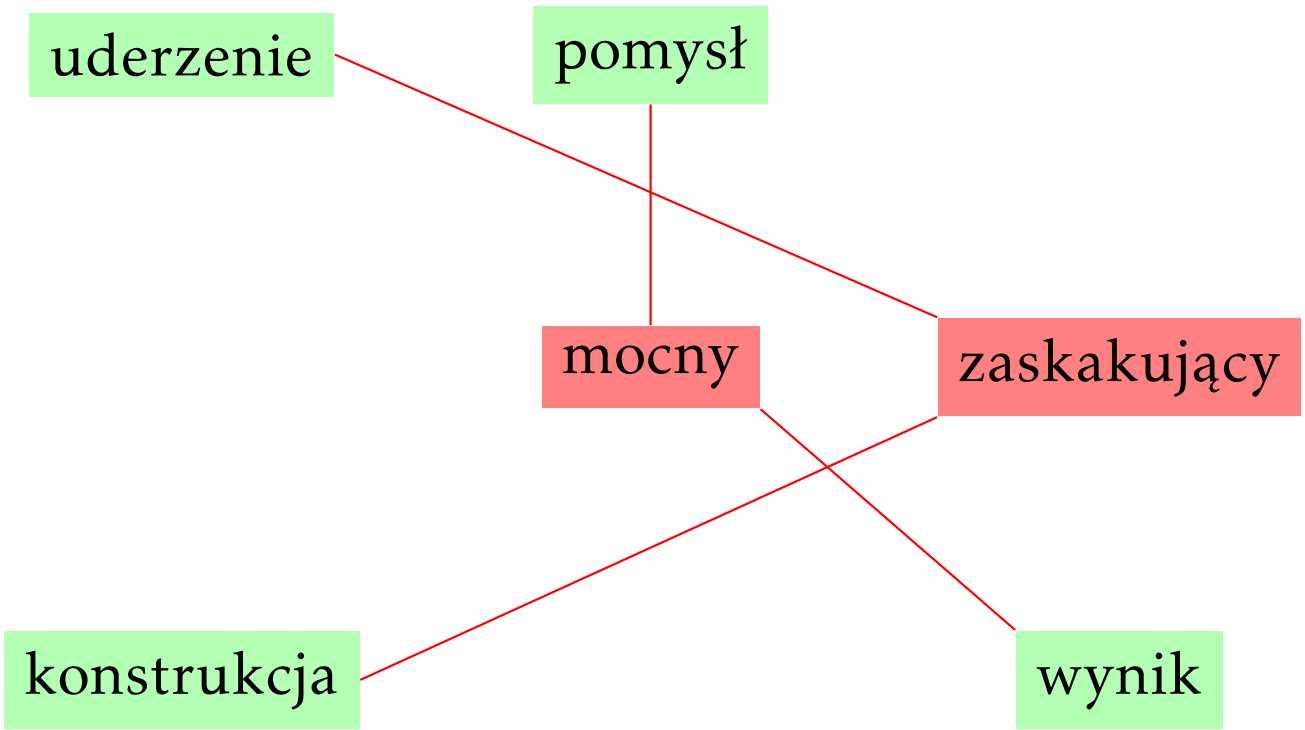
Przykład grafu kolokacji



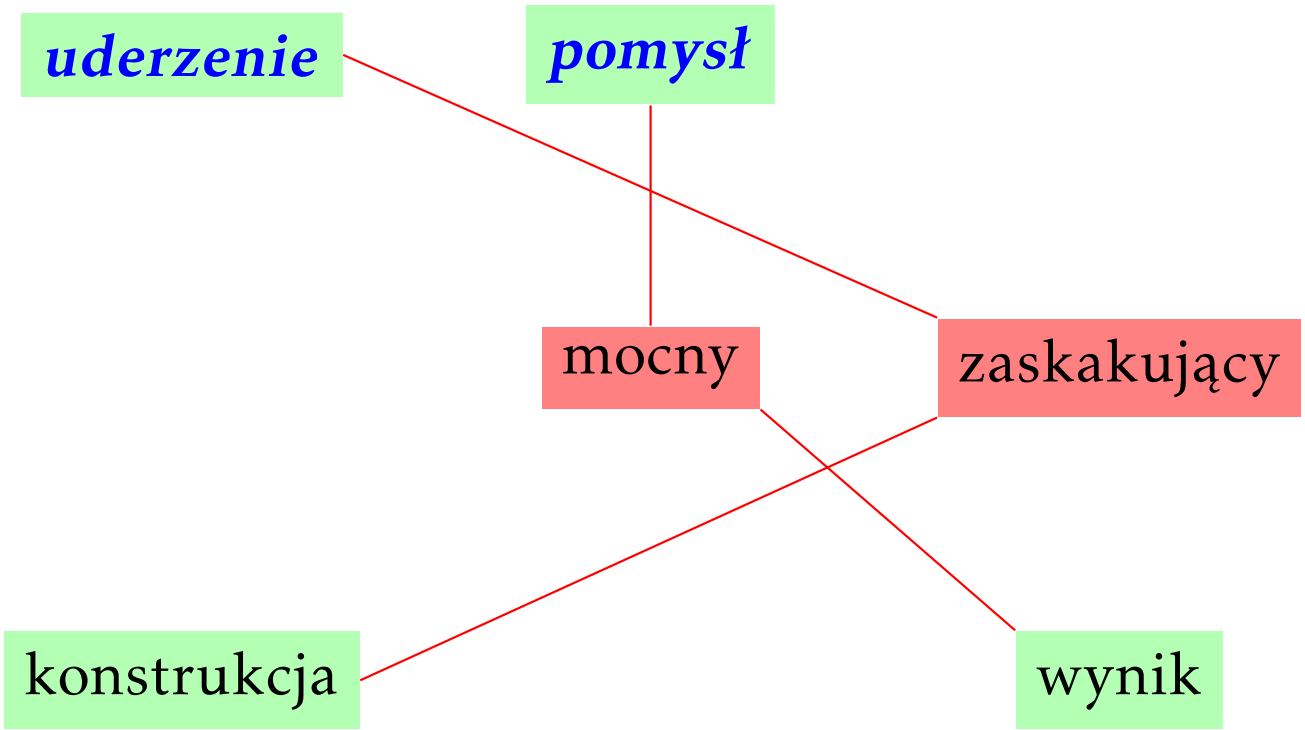
Podgraf dla u = fantastyczny



Graf $\mathcal{G}'$ fantastyczny



Graf $\mathcal{G}'_{\text{fantastyczny}}$, np. $C_{\text{fantastyczny}} = \{\text{uderzenie}, \text{pomysł}\}$



Szkielet Algorytmu Poszukiwania Korpusowych Punktów Minimalności

U — ośrodki kolokacji, V — kolokaty, $E \subseteq U \times V$ — kolokacje.

for all $u \in U$ **do**
  if IsUnique($E(u), u$) **then**
    Print(Min. unik. podzb. kolokatów dla u: MinCol(u))
  else
    Print(Brak unikalnych podzbiorów kolokatów dla u)
  end if
end for

Funkcja IsUnique

```
function IsUnique(C, u)
    for all u' ∈ U \ {u} do
        if C ⊆ E(u') then
            return false
        end if
    end for
    return true
end function
```

Funkcja Zwracająca Minimalne Unikalne Zbiory Kolokatów

Require: u ma unikalny zbiór kolokatów
  function MinCol(u)
    $n \leftarrow |E(u)|, \mathcal{C} \leftarrow \{E(u)\}$
    for all $C \in 2^{E(u)} \setminus \{E(u)\}$ **do**
      if IsUnique(C, u) **then**
        if $|C| < n$ **then**
          $n \leftarrow |C|, \mathcal{C} \leftarrow \{C\}$
        else if $|C| = n$ **then**
          $\mathcal{C} \leftarrow \mathcal{C} \cup \{C\}$
        end if
      end if
    end for
    return $\mathcal{C}$
  end function

Szanse Faktycznego Wykonania Obliczeń

Pętla wykonywana $2^{|E(u)|} - 1$ razy (to bardzo źle), ale:

- W korpusowym słowniku kolokacji zbiory kolokatów dla większości ośrodków kolokacji są niewielkie (< 10).
- Można te rozmiary regulować modyfikując wartości progowe częstości.
- Można też przyjąć że interesują nas tylko niewielkie unikalne zbiory kolokatów i tylko takie rozważać w algorytmie.
- W ostateczności można użyć algorytmów przybliżonych.

Ulepszenia Algorytmu: Reprezentacja Zbiorów Kolokatów

- IsUnique(C,u), gdzie dla każdego $u' \in U \setminus \{u\}$ sprawdzamy czy $C \subseteq E(u')$, jest wąskim gardłem.
- Kluczowa jest reprezentacja zbiorów $E(u)$ zapewniająca:
  - efektywne wykonywanie operacji na zbiorach
  - i efektywną serializację/deserializację (zbiory danych mogą nie mieścić się w pamięci).

Pomysł: Kompresowane Bitmapy

- Reprezentujemy $E(u)$ jako skompresowaną bitmapę $(B_{uv})_{v \in V}$, gdzie $B_{uv} := \begin{cases} 1 & \text{if } (u,v) \in E \\ 0 & \text{if } (u,v) \notin E. \end{cases}$
- Istnieją homomorficzne algorytmy kompresji.
- Istnieją implementacje dla Javy: JavaEWAH (Lemire et al., Data & Knowledge Engineering 69(1), 2010)

Ulepszenia Algorytmu: Generowanie Podzbiorów

- Podzbiory należy efektywnie generować w kolejności rozmiaru (1-elementowe, 2-elementowe, ...).
- Można zatrzymać się na podzbiorach K-elementowych dla arbitralnie ustalonego K.
- Dla ustalonego k, k-elementowe podzbiory powinny różnić się minimalnie.

Revolving door algorithm

Knuth, *The Art of Computer Programming*, Vol 4.

Ulepszone Obliczanie Unikalnego Minimalnego Zbioru Kolokatów

function MinCol(u)
  for $n \in 1..\min\{K, |E(u)|\}$ **do** ▷ K — wybrana wartość graniczna
    $\mathcal{C} \leftarrow \emptyset$
    for $C \in$ RevDoor$(n, E(u))$ **do**
      if IsUnique(C, u) **then**
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{C\}$
      end if
    end for
    if $\mathcal{C} \neq \emptyset$ **then**
      return $\mathcal{C}$
    end if
  end for
  return $\emptyset$
end function

Inne Problemy: Analiza Spójności Grup Ośrodków Kolokacji

- Problem: Zmierzyć jak spójne są synsety — stopień w jakim synonimy mają również podobne zbiory kolokatów.
- Pomysł: Użyć miar spójności w ważonych grafach dwudzielnych.
- Korzystamy z miar lokalnych — miara dla każdego synonimu.
- Jeśli potrzebna jest miara globalna synsetu, miary lokalne można uśrednić

Oznaczenia: Graf dwudzielny ważony (U, V, a)

- U — synset, V — kolokaty synsetu,
- $a : U \times V \to \mathbb{R}_+$ — miara współwystępowania ośrodka kolokacji z kolokatem w korpusie językowym.

Prosta Miara Spójności w Grafie Dwudzielnym Ważonym

Waga wierzchołka $u \in U$:

$$d_u := \sum_{v \in V} a_{uv}$$

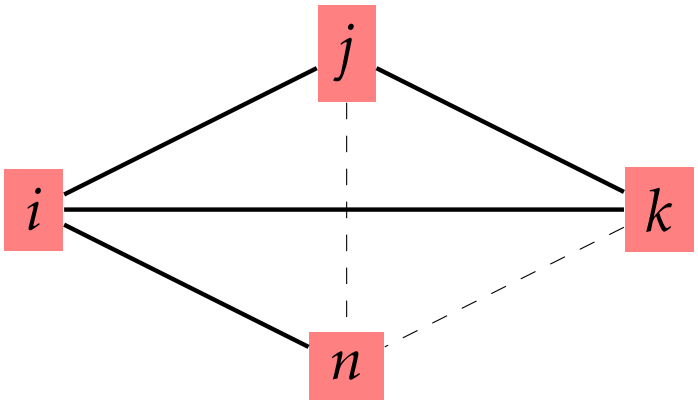
Prosta lokalna miara spójności dla $u \in U$:

$$S_u := \frac{1}{d_i(|U|-1)} \sum_{v \in V} \sum_{u' \in U \mid u' \neq u} a_{uv} a_{u'v}.$$

Miary Oparte Na Trójkątach

Watts, Duncan J., and Steven H. Strogatz.
Collective dynamics of small-world networks.
Nature 393.6684 (1998), 440–442.

- Stosunek ilości trójkątów z wierzchołkiem w danym wierzchłku do ilości możliwych.
- Istnieją niezliczone wersje dla grafów ważonych.
- W grafie dwudzielnym trójkątów nie ma.
- Należy zatem przerobić graf dwudzielny na jednodzielny.



Trójkątów dla i: 1

Możliwych trójkątów dla i: 3

Przykład Lokalnej Miary Opartej Na Trójkątach

Przerabiamy (U, V, a) na (U, w) gdzie

$$w : U \times U \to \mathbb{R}_+, \quad w_{uu'} := \frac{1}{\sqrt{d_u d_{u'}}} \sum_{v \in V} a_{uv} a_{u'v}.$$

Lokalna miara spójności dla $i \in U$

$$C_i := \frac{\sum_{j,k \in U \,|\, j \neq k \neq i} w_{ij} w_{ik} w_{kj}}{\sum_{j,k \in U \,|\, j \neq k \neq i} w_{ij} w_{ik}}.$$

Bibliografia I

Chlebda, W. (2003). *Elementy frazematyki, wprowadzenie do frazeologii nadawcy* (2nd ed.). Leksem.

Field, J. (2004). *Psycholinguistics: the key concepts*. London ; New York: Routledge.

Lewicki, A. M. (1976). *Wprowadzenie do frazeologii syntaktycznej: teoria zwrotu frazeologicznego*. Uniwersytet Śląski.

Mehler, A. (2007). Large text networks as an object of corpus linguistic studies. *Corpus Linguistics. An International Handbook of the Science of Language and Society. de Gruyter, Berlin/New York*.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and communication, 191*, 225.

Bibliografia II

Pęzik, P. (2013). Paradygmat dystrybucyjny w badaniach frazeologicznych. powtarzalność, reprodukcja i idiomatyzacja. In P. Stalmaszczyk (Ed.), *Metodologie językoznawstwa. ewolucja języka, ewolucja teorii językoznawczych*. Wydawnictwo Uniwersytetu Łódzkiego.

Sinclair, J. (1996). The search for units of meaning. *Textus*, *9*(1), 75–106.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011, April). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*(2), 251–272. Available from http://slr.sagepub.com/cgi/doi/10.1177/0267658310382068

Bibliografia III

Tremblay, A., & Bayeen, H. (2010). Holistic processing of regular four-word sequences: A behavioural and ERP study of the effects of structure, frequency and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: acquisition and communication*. London ; New York: Continuum.

Warren, P. (2013). *Introducing psycholinguistics*. Cambridge: Cambridge University Press.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge (GB); New York: Cambridge university press.