Uniwersalistyczna anotacja jednostek wielowyrazowych i wielojęzyczne studia korpusowe w projekcie PARSEME

Agata Savary

Université Paris-Saclay, Francja

Komisja Językoznawstwa Komputerowego, PAN, 11.12.2023

PARSEME

Sieć naukowa

- Akcja COST Action nt. parsowania i jednostek wielowyrazowych (JW, ang. *multiword expressions*, *MWE*), finansowana przez Komisję Europejską w **2013-2017**, nadal **aktywna**
- 31 krajów, 30 języków, 6 dialektów z 10 rodzin (*genera*, cf. WALS)
- Publikacje, zasoby, materiały szkoleniowe, metodologie, seria książkowa (PMWE)

Korpus (https://gitlab.com/parseme/corpora/-/wikis/)

- Wynik współpracy 26 zespołów, 35 liderów, 200 annotatorów
- **Podręcznik** anotatora dla jednostek wielowyrazowych **czasownikowych**, **ujednolicony** dla 26 języków, w duchu **uniwersalizmu**
- Korpus zaanotowany **ręcznie**: **26 języków**, na otwartych licencjach
- **Ciągłe usprawnianie** podręcznika i korpusu (w. 1.0, 1.1, 1.2, 1.3)

Badania uniwersalistyczne

- Zainteresowanie "uniwersalizmami statystycznymi" [Evans and Levinson(2009)], a nie "absolutnymi" [Greenberg(1996), Chomsky(1975)]
- Ich przydatność w **przetwarzaniu języka naturalnego**
- Modelowanie **wielu języków** jednocześnie
- Nazywanie i modelowanie zjawisk **podobnych** w ten sam sposób
- Dokumentowanie i podkreślanie zjawisk prawdziwie **specyficznych** dla danego języka
- Universal Dependencies [de Marneffe *et al.*(2021)], UniMorph [Kirov *et al.*(2018)], PARSEME [Savary *et al.*(2018)], Universal Anaphora [Poesio *et al.*(2023)], CorefUD [Nedoluzhko *et al.*(2022)]

Jednostki wielowyrazowe

Tak się składa, że obecnie czołówka Ministerstwa Sprawiedliwości dołączyła do opinii publicznej i naturalnie czerpie z tego korzyści propagandowe: nareszcie mamy ministra i prokuratora generalnego, który nie będzie się z przestępcami cackał!

Jednostki wielowyrazowe

Tak się składa, że obecnie czołówka Ministerstwa Sprawiedliwości dołączyła do opinii publicznej i naturalnie czerpie z tego korzyści propagandowe: nareszcie mamy ministra i prokuratora generalnego, który nie będzie się z przestępcami cackał!

Definicja [Baldwin and Kim(2010)]

Kombinacja co najmniej **dwóch słów**, posiadająca cechy **idiosynkratyczne** na poziomie leksykalnym, morfologicznym, składniowym i/lub semantycznym.

Idiosynkratyzm

Zachowanie lub własność specyficzna dla niewielkiej liczby osobników. Cecha **nietypowa**.

Najwyrazistszy idiosynkratyzm jednostek wielowyrazowych

Niekompozycyjność znaczeniowa

Znaczenie jednostki nie może być uzyskane w sposób regularny na podstawie znaczeń jej słów składniowych i jej struktury składniowej.

*sami woleli **zejść** ludziom **z oczu***

*a to **jest na rękę** PiS-owi*

*tu **wchodzą w grę** zbyt duże pieniądze*

Najwyrazistszy idiosynkratyzm jednostek wielowyrazowych

Niekompozycyjność znaczeniowa

Znaczenie jednostki nie może być uzyskane w sposób regularny na podstawie znaczeń jej słów składniowych i jej struktury składniowej.

*sami woleli **zejść** ludziom **z oczu***

*a to **jest na rękę** PiS-owi*

*tu **wchodzą w grę** zbyt duże pieniądze*

Problematyka

Testowanie niekompozycyjności znaczeniowej **wprost** jest zbyt trudne.

Przybliżanie niekompozycyjności znaczeniowej poprzez ograniczoną wariantywność

Hipoteza

Jednostki wielowyrazowe są **mniej** wariantywne niż konstrukcje regularne o tej samej strukturze składniowej [Gross(1988)].

Konstrukcja regularna	**Jednostka wielowyrazowa**	**Ograniczona wariantywność**
biały kot $\approx^{1}$ *jasny kot* $\approx$ *biały pies*	***biały kruk*** vs. *#jasny kruk* vs. *#biała wrona*	leksykalna
wchodzić w sytuację $\approx$ *wchodzić w sytuacje*	***wchodzić w grę*** vs. *#wchodzić w gry*	morfologiczna
myślałem, że nas woda zaleje $\approx$ *myślałem, że będziemy zalani wodą*	*myślałem, że mnie* ***krew zaleje*** vs. *#myślałem, że będę zalany krwią*	składniowa

[1] '$\approx$' zmianę znaczenia można wydedukować ze zmiany formalnej

Czasownikowe jednostki wielowyrazowe (CzJW) - cechy szczególne

- Nieciągłość w tekście:

  *Władymir Michalow **posadził** naczelnika od ciepłownictwa za lekceważenie sądu na trzy dni **do aresztu**.*

- Niejednoznaczność:

  *siostra ojca postraszyła Annę, że matka wkrótce **wyciągnie nogi***
  Siedziała wyciągnąwszy nogi w sandałach

- Pokrywanie się:

  *wielkoduszne przebaczenie **[[przyniosło]**$_{1,2}$ **więcej [szkody]**$_1$ **niż [pożytku]**$_2$**]**$_3$

- (Częściowa) wariantywność: leksykalna, morfologiczna, składniowa

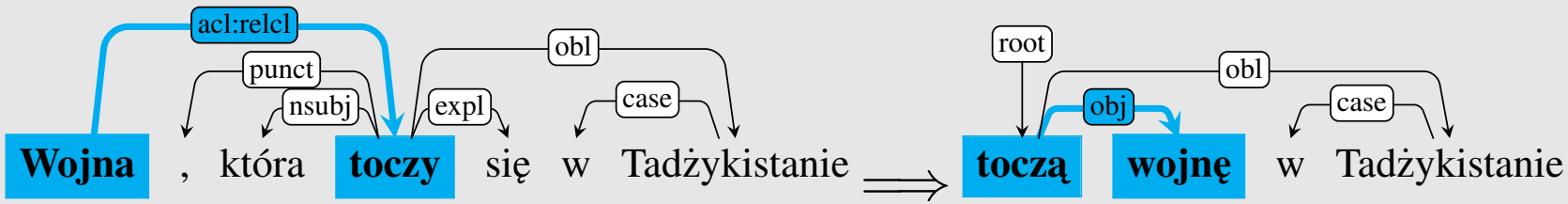
  ***nosić imię** Jana Pawła II* vs. ***nosi** jego **imię**, **noszących imię** Bronisława Malinowskiego*, ***noszą imiona** przywódców*

Neutralizacja wariantywności

Forma kanoniczna

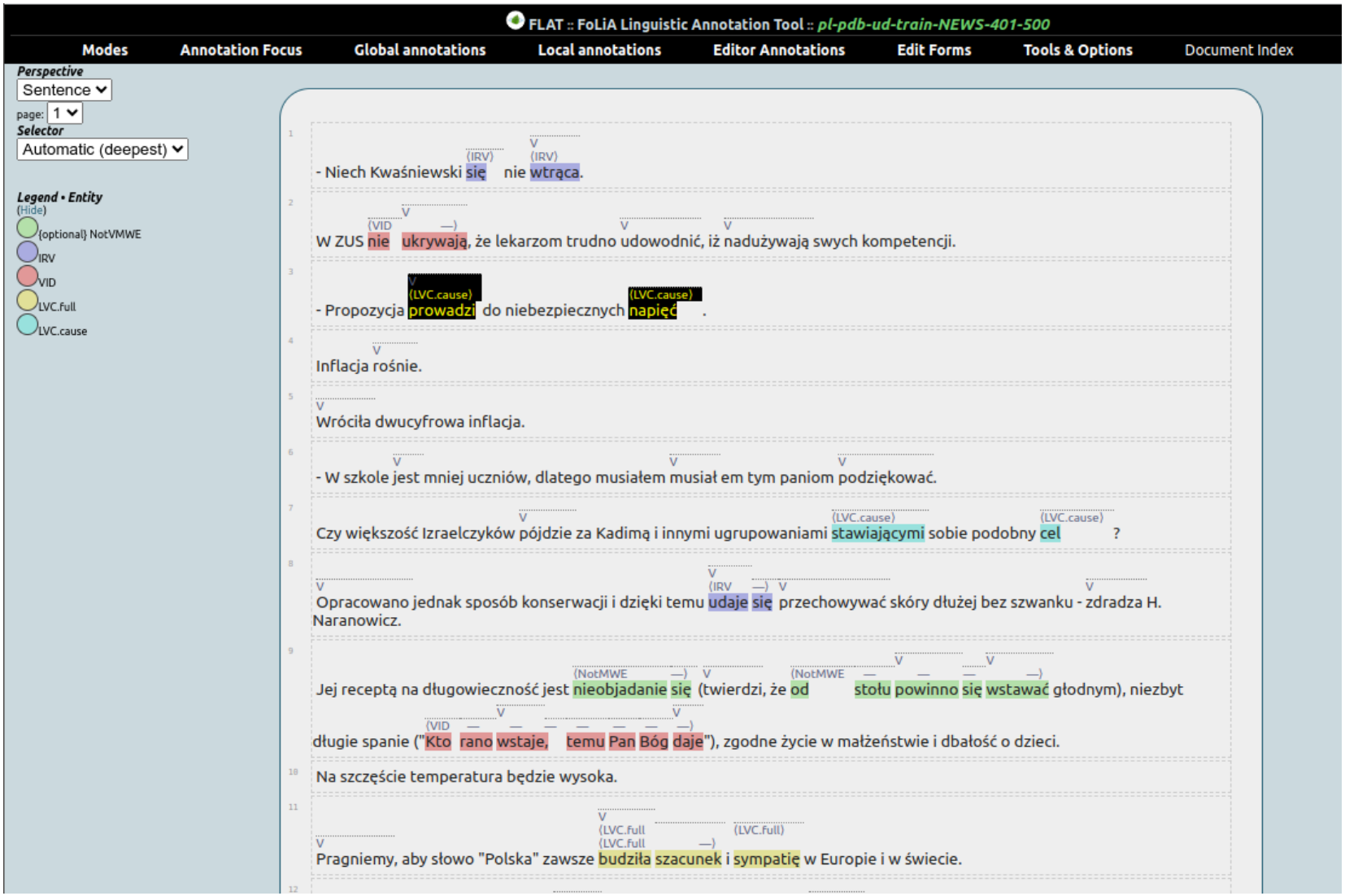
Najmniej składniowo nacechowany wariant, który zachowuje znaczenie idiomatyczne.

forma finitywna $<_m$ bezokolicznik/imiesłów; strona czynna $<_m$ str. bierna; forma bez negacji $<_m$ f. zanegowana; forma bez ekstrakcji $<_m$ f. z ekstrakcją, ... ($<_m$ = mniej nacechowana składniowo)



Formy kanoniczne służą do formalizacji własności jednostek wielowyrazowych. Stosowane są w **podręczniku anotatora**.

Annotacja czasownikowych jednostek wielowyrazowych



Reguły anotacji (https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3)

Cele naukowe

- Formalizacja idiomatyczności w duchu **uniwersalizmu**: unifikacja rzeczywistych **podobieństw** strukturalnych, poszanowanie i podkreślenie rzeczywistych **różnic**.
- Formalizacja **stosowalna** w automatycznym przetwarzaniu języka:
  - Potrzeba jasno zdefiniowanych **granic** między kategoriami,
  - Niekompozycyjność jest kwestią **skali**, ale decyzje mają być **binarne**.
- **Powtarzalność** decyzji podejmowanych w procesie anotacji.

Klasyfikacja CzJW (w. 1.3)

- Kategorie "**uniwersalne**" (istniejące we wszystkich językach dotąd studiowanych) :
  - **LVCs**: konstrukcje z rzeczownikiem predykatywnym (ang. *light verb constructions*)
    - **LVC.full**: ***czerpać korzyści***
    - **LVC.cause**: ***zwracać*** *czyjąś* ***uwagę*** *na coś*
  - **VIDs** idiomy czasownikowe (ang. *verbal idioms*)

    zejść *komuś* ***z oczu***
- Kategorie **quasi-uniwersalne** (istniejące w wielu językach):
  - **IRVs**: czasowniki zwrotne właściwe (*inherently reflexive verbs*)

    cacakać się, ***liczy się*** *z kimś*
  - **VPCs**: konstrukcje czasownikowe z partykułą (*verb-particle constructions*)
    - **VPC.full** EN *to* ***do in*** 'zabić'
    - **VPC.semi** EN *to* ***eat up*** 'wyjeść'
  - **MVCs**: konstrukcje wieloczasownikowe (*multiverb constructions*)

    HI ***kar le-na*** (lit. '*zrobić wziąć*') 'zrobić coś dla własnej korzyści'
- Kategoria **eksperymentalna**
  - **IAVs**: czasowniki inherentnie adpozycyjne (*inherently adpositional verbs*)

    nie ***doszło do*** *żadnych ustaleń*

Podręcznik w formie diagramów decyzyjnych

If you are annotating **Italian** or **Hindi**, go to the Italian-specific decision tree or Hindi-specific decision tree. I

- Apply test S.1 - [**1HEAD**: Unique verb as functional syntactic head of the whole?]
  - **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
    - **YES** ⇒ Annotate as a VMWE of category **VID**
    - **NO** ⇒ It is not a VMWE, **exit**
  - **YES** ⇒ Apply test S.2 - [**1DEP**: *Verb v has exactly one lexicalized dependent d?*]
    - **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
      - **YES** ⇒ Annotate as a VMWE of category **VID**
      - **NO** ⇒ It is not a VMWE, **exit**
    - **YES** ⇒ Apply test S.3 - [**LEX-SUBJ**: *Lexicalized subject?*]
      - **YES** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
        - **YES** ⇒ Annotate as a VMWE of category **VID**
        - **NO** ⇒ It is not a VMWE, **exit**
      - **NO** ⇒ Apply test S.4 - [**CATEG**: *What is the morphosyntactic category of d?*]
        - **Reflexive clitic** ⇒ Apply IRV-specific tests ⇒ *IRV tests positive?*
          - **YES** ⇒ Annotate as a VMWE of category **IRV**
          - **NO** ⇒ It is not a VMWE, **exit**
        - **Particle** ⇒ Apply VPC-specific tests ⇒ *VPC tests positive?*
          - **YES** ⇒ Annotate as a VMWE of category **VPC.full** or **VPC.semi**
          - **NO** ⇒ It is not a VMWE, **exit**
        - **Verb with no lexicalized dependent** ⇒ Apply MVC-specific tests ⇒ *MVC tests positive?*
          - **YES** ⇒ Annotate as a VMWE of category **MVC**
          - **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
            - **YES** ⇒ Annotate as a VMWE of category **ID**
            - **NO** ⇒ It is not a VMWE, **exit**
        - **Extended NP** ⇒ Apply LVC-specific decision tree ⇒ *LVC tests positive?*
          - **YES** ⇒ Annotate as a VMWE of category **LVC**
          - **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
            - **YES** ⇒ Annotate as a VMWE of category **VID**

Diagram decyzyjny dla IRV

IRV-specific decision tree

- ↳ Apply test IRV.1 - [**INHERENT**]
  - ↳ **YES ⇒** Annotate as **IRV**
  - ↳ **NO ⇒** Apply test IRV.2 - [**DIFF-SENSE**]
    - ↳ **YES ⇒** Annotate as **IRV**
    - ↳ **NO or UNSURE ⇒** Apply test IRV.3 - [**DIFF-SUBCAT**]
      - ↳ **YES ⇒** Annotate as **IRV**
      - ↳ **NO ⇒**
        - ↳ verb has no subject ⇒ Apply test IRV.4 - [**IMPERS**]
          - ↳ **YES ⇒** It is not a VMWE, **exit**
          - ↳ **NO ⇒** Annotate as **IRV**
        - ↳ verb has a subject ⇒ Apply test IRV.5 - [**MIDDLE-INCHO**]
          - ↳ **YES ⇒** It is not a VMWE, **exit**
          - ↳ **NO ⇒** Apply test IRV.6 - [**REFL**]
            - ↳ **YES ⇒** It is not a VMWE, **exit**
            - ↳ **NO or UNSURE ⇒**
              - ↳ subject is SINGULAR ⇒ Apply test IRV.7 - [**REFL-MUTUAL**]
                - ↳ **YES ⇒** It is not a VMWE, **exit**
                - ↳ **NO ⇒** Annotate as **IRV**
              - ↳ subject is PLURAL ⇒ Apply test IRV.8 - [**RECIPRO**]
                - ↳ **YES ⇒** It is not a VMWE, **exit**
                - ↳ **NO ⇒** Annotate as **IRV**

Testy na ograniczoną wariantywność

Test IRV.5 - [MIDDLE-INCHO] - Middle or Inchoative

When you move the subject to the object position, remove the RCLI and add a generic subject (people, somebody), thus building a transitive version, does it imply the REFLV version? In other words, *people/somebody V [to] X ⇒ X REFLV*?

↳ **YES ⇒** do **NOT** annotate as verbal MWE

- *(BG) някой отваря вратата ⇒ вратата се отваря*
- *(DE) man kann die Häuser gut verkaufen ⇒ die Häuser verkaufen sich gut*
  jemand öffnet die Tür ⇒ die Tür öffnet sich
- *(ES) la gente cuenta historias ⇒ se cuentan historias*
  alguien abrió la puerta ⇒ la puerta se abrió
- *(FR) on vend bien ce produit ⇒ ce produit se vend bien*
  quelqu'un ouvre la porte ⇒ la porte s'ouvre,
- *(PL) ktoś sprzedaje te domy ⇒ te domy się sprzedają*
  ktoś otwiera drzwi ⇒ drzwi się otwierają
  ktoś nasila skargi ⇒ s
  ktoś rozgrywa mecz ⇒ mecz rozgrywa się
- *(RO) cineva spune glume ⇒ se spun glume*
  cineva a deschis uşa ⇒ uşa s-a deschis
- *(SL) nekdo pripoveduje šale ⇒ šale se pripovedujejo*
  nekdo je odprl vrata ⇒ vrata so se odprla
- *(SR) неко је отварао врата ⇒ врата се отварају [NEKO JE OTVARAO VRATA ⇒ VRATA SE OTVARAJU],*
  неко шири гласине ⇒ гласине се шире [NEKO ŠIRI GLASINE ⇒ GLASINE SE ŠIRE]

lit. *somebody plays a game ⇒ the game plays*

↳ **NO ⇒** next test

Korpus PARSEME (w. 1.3) – wyniki [Savary *et al.*(2023)]

Annotacje

Zdania	Tokeny	CzJW	VID	IRV	LVC.full	LVC.cause	VPC.full	VPC.semi	IAV	MVC
455,629	9,264,811	127,498	26,214	29,062	**40,933**	3,238	9,164	6,443	7,375	5,032

Znaczące fakty

- **Różnorodność**: 26 języków z 10 rodzin
  - AR, BG, CS, DE, EL, EN, ES, EU, FA, FR, GA, HE, HI, HR, HU, IT, LT, MT, PL, PT, RO, SL, SV, SR, TR, ZH
  - bałtycka, baskijska, celtycka, chińska, germańska, grecka, indyjska, irańska, romańska, semicka, słowiańska, turkijska, ugryjska
- Publikacja na **swobodnych licencjach** (warianty Creative Commons & GPL)
- "**Uniwersalność**" kategorii **LVC** i **VID**
- Ujawnienie jakościowego i ilościowego znaczenia kategorii **IRV**
- Pokrycia i zagniedżenia są rzadkie w korpusie

Język polski

Język polski w PARSEME

Zdania	Tokeny	CzJW	VID	IRV	LVC.full	LVC.cause
23 547	396 140	7 313	833	**3 688**	2 478	314

Korpus średniej wielkości:

- 7. (il. zdań), 10. (il. tokenów), 6. (il. CzJW), 3. (il. IRV), 8. (il. LVC.full)

Automatyczna identyfikacja CzJW w tekście [Constant *et al.*(2017a)]

Kampanie ewaluacyjne PARSEME
[Savary *et al.*(2017), Ramisch *et al.*(2018), Ramisch *et al.*(2020)]

- **Zadanie**: Automatyczna identyfikacja wystąpień CzJW w zadanym tekście.
- **3 edycje** w oparciu o korpus PARSEME
- Udział kilkunastu systemów **wielojęzycznych**
- Zadanie pozostaje **trudne** ($F_1 = 70\%$), nawet dla modeli opartych na uczeniu głębokim.
- CzJW **niewidziane** (niewystępujące w korpusie treningowym) są szczególnie trudne do identyfikacji

Niejednoznaczność CzJW

[Savary *et al.*(2019)]: Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, Voula Giouli (2019): *Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir*, The Prague Bulletin of Mathematical Linguistics, 112, pp. 5-54.

Niejednoznaczność JW

siostra ojca postraszyła Annę, że matka wkrótce ***wyciągnie nogi***

Siedziała wyciągnąwszy nogi w sandałach

Znaczenia dosłowne JW - stan wiedzy

Filozofia języka: obliczanie interpretacji idiomów w mózgu

- [Grice(1989)]: interpretacja **dwu-stopniowa** (dosłowna → idiomatyczna)
- [Recanati(1995)]: **bezpośredni** dostęp do interpretacji idiomatycznej

Lingwistyka

- [Popiel and McRae(1988), Cacciari and Corradini(2015), Geeraert *et al.*(2018)]: interpretacje przenośne są częstsze i bardziej **preferowane** niż dosłowne
- [Abeillé and Schabes(1989), Lichte and Kallmeyer(2016)]: **kodowanie interpretacji** dosłownych i przenośnych w **gramatykach formalnych** (LTAG)
- [Sheinfux *et al.*(2019), Pausé(2017)]: związki między interpretacjami dosłownymi i przenośnymi tłumaczą **wariantywność morfoskładniową**

Znaczenia dosłowne JW - stan wiedzy w przetwarzaniu języka naturalnego

Zasoby

- [Tu and Roth(2011), Tu and Roth(2012), Cook *et al.*(2008), Hashimoto and Kawahara(2008), Savary and Cordeiro(2018), Ehren *et al.*(2020)]: wystąpienia dosłowne vs. idiomatyczne (skala **binarna**)
- [Bott *et al.*(2016), Ramisch *et al.*(2016), Cordeiro *et al.*(2019)] stopień kompozycyjności (skala **kilkustopniowa**)

Ujednoznacznianie

- [Hashimoto and Kawahara(2008), Fazly *et al.*(2009), Peng *et al.*(2014), Peng and Feldman(2016), Köper and Schulte im Walde(2016), Constant *et al.*(2017b), Ehren *et al.*(2020)]: Automatyczne odróżnianie znaczeń idiomatycznych od dosłownych jako jedno **najistotniejszych wyzwań** w PJN
- [Waszczuk *et al.*(2016), Savary and Cordeiro(2018)]: **Analiza ilościowa** wystąpień znaczeń dosłownych (**znikomość** dla polskiego)

Problematyka

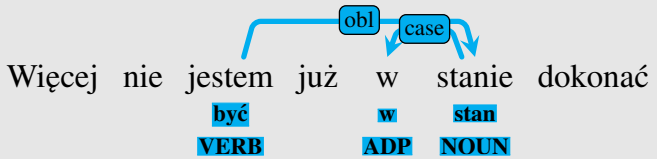
- Jak **zdefinować** wystąpienie dosłownego znaczenia JW?
- Jak **częste** są wystąpienia dosłownych znaczeń JW?
- Czym **charakteryzują** się wystąpienia dosłownych znaczeń JW?

Kontekst

- Baskijski, grecki, niemiecki, polski, portugalski (5 rodzin)
- Korpus PARSEME, morfoskładnia zgodna z **Universal Dependencies**

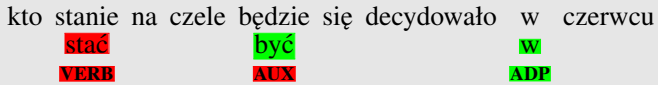
Definicje

Wystąpienie znaczenia dosłownego JW (w skrócie: **wystąpienie dosłowne**)



Dla danej JW $j = j_1, \ldots, j_n$:

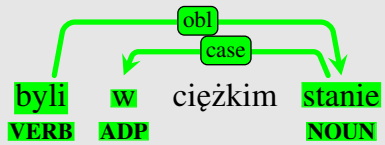
(a) Formy podstawowe **wszystkich składników** $j_1, \ldots, j_n$ występują razem



(b) Mają tę samą **formę podstawową** i **część mowy**

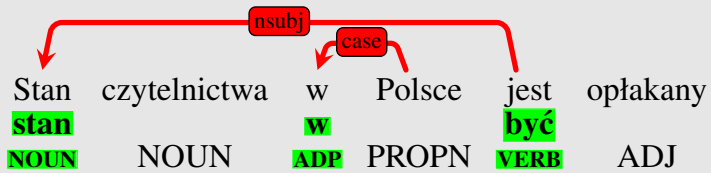
(c) **Relacje zależnościowe** zmiędzy nimi są **takie same lub równoważne** tym w formie kanonicznej

(d) Znaczenie **idiomatyczne nie występuje**



Przypadkowe współwystąpienie składników JW (w skrócie: **wystąpienie przypadkowe**)

Warunek (c) nie jest spełniony:



Wydobywanie wystąpień dosłownych i przypadkowych

- Korpus PARSEME v 1.1 DE, EL, EU, PL, PT - **częściowo automatyczna** anotacja morfologii i składni
- Regułowe **wydobywanie** kandydatów z tym samym zestawem **form podstawowych** lub form **odmiany**, co zaanotowane CzJW
- Regułowe **filtrowanie**:
  - Człony odległe o najwyżej **2 wtrącenia**,
  - Człony tworzące **spójny** graf składniowy.
- Klasyfikacja **ręczna**:
  - Błąd korpusu.
  - Brakująca anotacja CzJW
  - Kontekst niedostateczny do rozstrzygnięcia
  - Wystąpienie dosłowne (WD),
  - Wystąpienie przypadkowe (WP),

Wyniki

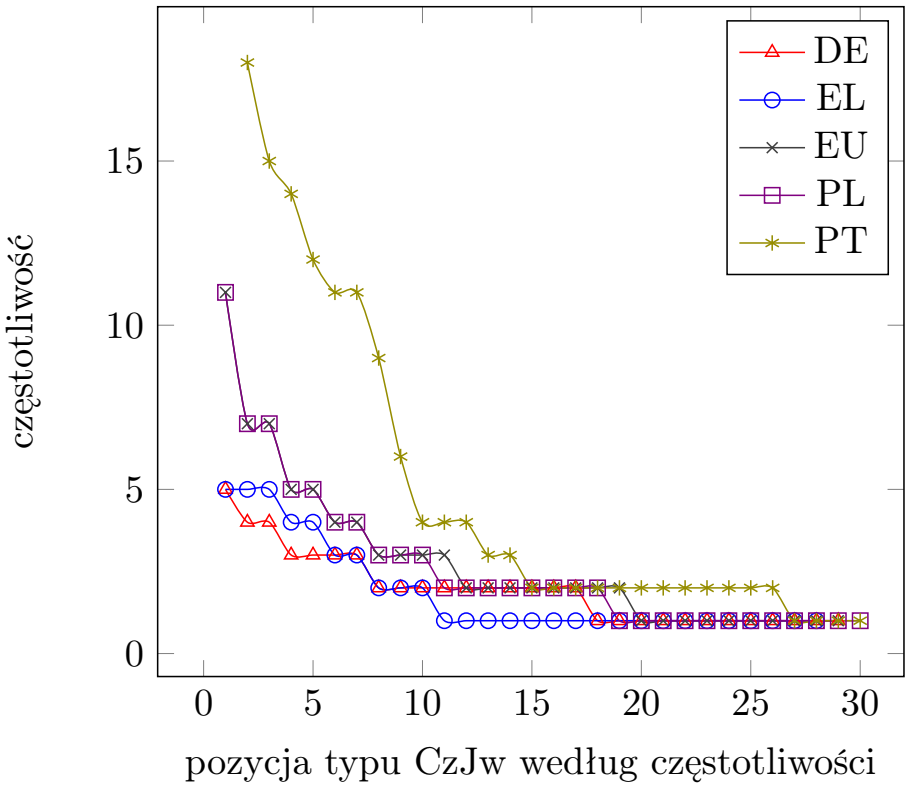
	DE	EL	EU	PL	PT
Zaanotowane CzJW	3,823	2,405	3,823	4,843	5,536
Wydobyte kombinacje	926	451	2,618	332	1,997
Błędy	61.6% (570)	12.9% (58)	36.4% (952)	1.8% (6)	42.3% (845)
Brakujące CzJW	27.0% (250)	47.5% (214)	17.3% (453)	5.4% (18)	10.7% (213)
Brak kontekstu	0.3% (3)	0.2% (1)	0.5% (12)	2.1% (7)	0.7% (13)
Wyst. przypadkowe	2.6% (24)	27.9% (126)	42.4% (1110)	61.1% (203)	33.5% (668)
Wyst. dosłowne	8.5% (79)	11.5% (52)	3.5% (91)	29.5% (98)	12.9% (258)
Stopień idiomatyczności	**98%**	**98%**	**98%**	**98%**	**96%**

Stopień idiomatyczności: $\frac{|CzJW|}{|CzJW|+|WD|}$ (wykryte brakujące anotacje $\in$ CzJW)

Konkluzja

Gdy wymagania morfoskładniowe danej CzJW są spełnione, wystąpienie jest **prawie zawsze idiomatyczne**, a nie dosłowne.

Dystrybucja dosłownych wystąpień



Prawo Zipfa

Większość wystąpień dosłownych dotyczy **niewielkiej** ilości CzJW.

Dosłowne wystąpienia IRV

W polskim, wystąpienia dosłowne najczęściej dotyczą jednostek IRV. Wynika to z częstego użycia **zwrotnego**, **wzajemnego**, **bezosobowego** i **medialnego** (*middle passive*) zaimków zwrotnych.

*Polityk **dopuszczał się** bezprawia.*

Dopuszcza się inną działalność niż gastronomiczna.

Dosłowne wystąpienia LVC

- LVC są w szarej strefie między jednostkami idiomatycznymi a regularnymi.
- Wystąpienia dosłowne = nie spełniające testów z diagramów decyzyjnych, np. **rzeczownik niepredykatywny**.

  *Nie **mają** wymaganego **zezwolenia** na pracę.*

  Kierowcy mieli sfałszowane zezwolenia.

Systematyczna niejednoznaczność między LVC a negacją **egzystencjalnego** użycia *być*:

*(Klient) nie **ma powodów** do satysfakcji.*

Nie ma powodów do satysfakcji.

Dosłowne wystąpienia VID

Obrazowe VID mają silny potencjał interpretacji dosłownych, ale są rzadkie.

*Służenie nam **mają we krwi**.*

Miał we krwi ponad 1,5 promila alkoholu.

Podsumowanie

Korpus PARSEME

- Reguły anotacji PARSEME dla **czasownikowych** jednostek wielowyrazowych są **ujednolicone** dla 26 języków (w tym dla **polskiego**), mają stosunkowo niewiele fragmentów **specyficznych** dla pojedynczych języków.
- Anotacja opiera się na **diagramach decyzyjnych** co przyczynia się do **powtarzalności** decyzji anotatorek
- Testy oparte są na **strukturze składniowej**
- Niekompozycyjność jest kwestią **skali**, ale decyzje są **binarne**
- Głównym celem jest oddanie **niekompozycyjności**, ale jest testowanie **wprost** jest **trudne**.
- Dlatego jest ona **przybliżana** poprzez **wariantywność** leksykalną i morfoskładniową
- Korpus PARSEME pozwala na studia nad CzJW

Podsumowanie

Wystąpienia dosłowne

- Wystąpienia dosłowne należy definiować na podstawie ich własności **semantycznych** i **składniowych** (zapotrzebowanie na korpusy drzewiaste)
- **Wystąpienia dosłowne** są **rzadkie** (2-4%)
- Wystąpienia **przypadkowe** są **częstsze**, ale mogą być wyeliminowane dzięki dobrym **parserom**
- Proste **heurystyki** mają duże szanse powodzenia w **identyfikacji kontrolowanych CzJW** [Pasquer *et al.*(2020)]
- Procedury dostosowane do języka mogą pomóc w eliminacji **kilku częstych** przypadków

Bibliography I

Abeillé, A. and Schabes, Y. (1989).
Parsing idioms in lexicalized tags.
In H. L. Somers and M. M. Wood, eds., *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester*, pp. 1–9. The Association for Computer Linguistics.

Baldwin, T. and Kim, S. N. (2010).
Multiword expressions.
In N. Indurkhya and F. J. Damerau, eds., *Handbook of Natural Language Processing*, pp. 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.

Bott, S., Khvtisavrishvili, N., Kisselew, M., and Schulte im Walde, S. (2016).
G$_h$ost-PV: A Representative Gold Standard of German Particle Verbs.
In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon*, Osaka, Japan.

Cacciari, C. and Corradini, P. (2015).
Literal analysis and idiom retrieval in ambiguous idioms processing: A reading-time study.
Journal of Cognitive Psychology, **27**(7), 797–811.

Chomsky, N. (1975).
Reflections on Language. Temple Smith, London.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017a).
Multiword Expression Processing: A Survey.
Computational Linguistics, **43**(4), 837–892.

Bibliography II


Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017b).
Multiword expression processing: A survey.
Computational Linguistics, to appear.


Cook, P., Fazly, A., and Stevenson, S. (2008).
The vnc-tokens dataset.
In *Proceedings of the Workshop on Multiword Expressions*.


Cordeiro, S., Villavicencio, A., Idiart, M., and Ramisch, C. (2019).
Unsupervised compositionality prediction of nominal compounds.
Computational Linguistics.
(to appear).


de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021).
Universal Dependencies.
Computational Linguistics, 47(2), 255–308.


Ehren, R., Lichte, T., Kallmeyer, L., and Waszczuk, J. (2020).
Supervised disambiguation of German verbal idioms with a BiLSTM architecture.
In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 211–220, Online. Association for Computational Linguistics.


Evans, N. and Levinson, S. C. (2009).
The myth of language universals: Language diversity and its importance for cognitive science.
Behavioral and Brain Sciences, 32(5), 429–448.

Bibliography III


Fazly, A., Cook, P., and Stevenson, S. (2009).
Unsupervised type and token identification of idiomatic expressions.
Computational Linguistics, 35(1), 61–103.


Geeraert, K., Baayen, R. H., and Newman, J. (2018).
"Spilling the bag" on idiomatic variation.
In S. Markantonatou, C. Ramisch, A. Savary, and V. Vincze, eds., *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pp. 1–33. Language Science Press., Berlin.


Greenberg, J. H., ed. (1996).
Universals of language. MIT Press.


Grice, H. P. (1989).
Studies in the Way of Words. Harvard University Press, Cambridge, Mass.


Gross, G. (1988).
Degré de figement des noms composés.
Langages, 90, 57–72.


Hashimoto, C. and Kawahara, D. (2008).
Construction of an idiom corpus and its application to idiom identification based on wsd incorporating idiom-specific features.
In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 992–1001. Association for Computational Linguistics.

Bibliography IV


Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2018).
UniMorph 2.0: Universal Morphology.
In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).


Köper, M. and Schulte im Walde, S. (2016).
Distinguishing literal and non-literal usage of german particle verbs.
In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 353–362, San Diego, California.


Kupść, A. (1999).
Haplology of the Polish reflexive marker.
In R. D. Borsley and A. Przepiórkowski, eds., *Slavic in Head-Driven Phrase Structure Grammar*, pp. 91–124. CSLI Publications, Stanford, CA.


Lichte, T. and Kallmeyer, L. (2016).
Same syntax, different semantics: A compositional approach to idiomaticity in multi-word expressions.
In C. Piñón, ed., *Empirical Issues in Syntax and Semantics 11*, pp. 111–140.


Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., Zeldes, A., and Zeman, D. (2022).
CorefUD 1.0: Coreference meets Universal Dependencies.
In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4859–4872, Marseille, France. European Language Resources Association.

Bibliography V

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020).
Verbal multiword expression identification: Do we need a sledgehammer to crack a nut?
In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pausé, M.-S. (2017).
Structure lexico-sentaxique des locutions du français et incidence sur leur combinatoire.
Ph.D. thesis, Université de Lorraine, Nancy, France.

Peng, J. and Feldman, A. (2016).
Automatic idiom recognition with word embeddings.
In *SIMBig (Revised Selected Papers)*, pp. 17–29. Springer.

Peng, J., Feldman, A., and Vylomova, E. (2014).
Classifying idiomatic and literal expressions using topic models and intensity of emotions.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Poesio, M., Zeldes, A., Nedoluzhko, A., Khosla, S., Manuvinakurike, R., Moosavi, N., Ng, V., Ogrodniczuk, M., Pradhan, S., Rose, C., Strube, M., Yu, J., Grishina, Y., Hou, Y., and Landragin, F. (2023).
Universal Anaphora 1.0 – Proposal for Discussion.
work in progress.

Popiel, S. J. and McRae, K. (1988).
The figurative and literal senses of idioms, or all idioms are not used equally.
Journal of Psycholinguistic Research, 17(6), 475–487.

Bibliography VI

Ramisch, C., Cordeiro, S., Zilio, L., Idiart, M., Villavicencio, A., and Wilkens, R. (2016).
How naked is the naked truth? A multilingual lexicon of nominal compound compositionality.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 156–161, Berlin, Germany. ACL.
CORE2018 rank: A*. https://aclweb.org/anthology/P16-2026.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018).
Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.
In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pp. 222–240. Association for Computational Linguistics.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., and Xu, H. (2020).
Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions.
In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pp. 107–118, online. Association for Computational Linguistics.

Recanati, F. (1995).
The alleged priority of literal interpretation.
Cognitive Science, 19, 207–232.

Bibliography VII

Savary, A. and Cordeiro, S. (2018).
Literal readings of multiword expressions: as scarce as hen's teeth.
In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT 16), Jan 2018, Prague, Czech Republic*, pp. 64 – 72, Prague, Czech Republic.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017).
The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.
In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pp. 31–47, Valencia, Spain. Association for Computational Linguistics.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., and Vincze, V. (2018).
PARSEME multilingual corpus of verbal multiword expressions.
In S. Markantonatou, C. Ramisch, A. Savary, and V. Vincze, eds., *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pp. 87–147. Language Science Press., Berlin.

Savary, A., Cordeiro, S. R., Lichte, T., Ramisch, C., nurrieta, U. I., and Giouli, V. (2019).
Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir.
The Prague Bulletin of Mathematical Linguistics, 112, 5–54.

Bibliography VIII

Savary, A., Ben Khelil, C., Ramisch, C., Giouli, V., Barbu Mititelu, V., Hadj Mohamed, N., Krstev, C., Liebeskind, C., Xu, H., Stymne, S., Güngör, T., Pickard, T., Guillaume, B., Bejček, E., Bhatia, A., Candito, M., Gantar, P., Iñurrieta, U., Gatt, A., Kovalevskaite, J., Lichte, T., Ljubešić, N., Monti, J., Parra Escartín, C., Shamsfard, M., Stoyanova, I., Vincze, V., and Walsh, A. (2023).
PARSEME corpus release 1.3.
In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pp. 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Sheinfux, L. H., Greshler, T. A., Melnik, N., and Wintner, S. (2019).
Verbal MWEs: Idiomaticity and flexibility.
In Y. Parmentier and J. Waszczuk, eds., *Representation and Parsing of Multiword Expressions*, pp. 5–38. Language Science Press, Berlin.

Tu, Y. and Roth, D. (2011).
Learning English light verb constructions: Contextual or statistical.
In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, pp. 31–39. Association for Computational Linguistics.

Tu, Y. and Roth, D. (2012).
Sorting out the most confusing English phrasal verbs.
In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*, pp. 65–69. Association for Computational Linguistics.

Bibliography IX

Waszczuk, J., Savary, A., and Parmentier, Y. (2016).
Promoting multiword expressions in A* TAG parsing.
In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pp. 429–439.

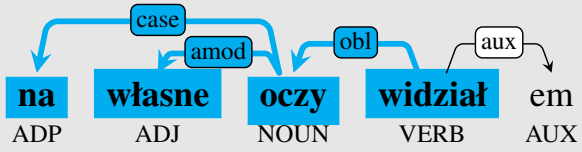
Oprogramowanie korpusu PARSEME

- PARSEME ▸ [wiki] - obszerna dokumentacja korpusu i narzędzi
- Podręcznik anotatora
- Podręcznik użytkownika platformy anotacyjnej
- Repozytoria Gitlab dla ▸ [26 języków]
- Narzędzia do automatycznej walidacji, konwersji, filtrowania i publikacji korpusu
- Wspieranie jakości
  - "Pionowe" testy spójności anotacji ▸ [zob. dla polskiego]
- Przeglądarka korpusowa ▸ [Grew-match]

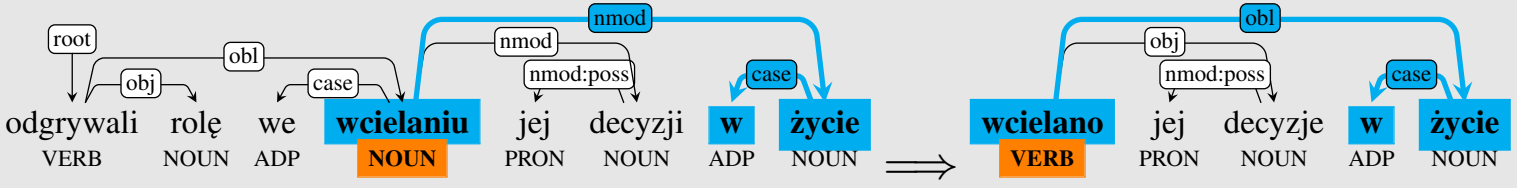
Język polski

Specyficzne zjawiska

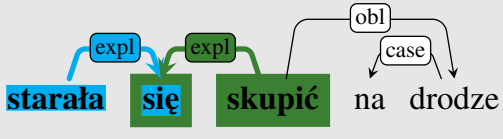
- Częstość wielosłowowych form czasownikowych:



- Gerundia jako warianty CzJW :



- Haplologia zaimka dzierżawczego [Kupść(1999)]



- Czasowniki quasi-zwrotne (cf. test IRV.1)

  delektować się (piwkiem) (563 wystąpień w NKJP)
  delektuje nas (znakomitymi zdjeciami) (3 wystąpienia w NKJP)

- Produktywna idiomatyczność

  na+VERB + *się* ⇒ IRV (cf. test IRV.3)
  nacierpieć się, ***naczytał się***, ***nasiedziała się***, ***nazamiataliśmy się***