Marek Łaziński
m.lazinski@uw.edu.pl

Aspekt czasownika w słownikach i korpusach. Jak i po co tagować pary aspektowe?

(Projekt z grantu Beethoven 2016/23/G/HS2/00922)

Seminarium Przetwarzanie języka naturalnego
IPI PAN, 21.1.2019

Plan wykładu

- Aspekt i para aspektowa. Definicja pary
- Pary aspektowe w słownikach
- Aspekt w korpusach słowiańskich i w NKJP
  - Korpus polsko-niemiecki UW
- Tagowanie par. Zasady ogólne i problemy
  - System trójdzielny: aspekt, wyznacznik, para
  - Homonimia leksykalna i aspektowa - niejednoznaczność przypisania leksemu do pary
  - Trójki aspektowe
- Przykład badania
  - Profile gramatyczne par prefiksalnych i sufiksalnych

Aspekt

- Przypisany wszystkim czasownikom w każdej formie
  - Inwariant dk: zdarzenie, determinacja czasowa
  - Inwariant ndk: trwanie (negacja zdarzenia) lub wielokrotność
  - Wartości: dk, ndk, dwuaspektowe, np. *aresztować*
  - Test: tylko ndk z czas. fazowymi, np. *zacząć jeść *zjeść*
- Para aspektowa
  - Ndk i dk tożsame leksykalnie (cokolwiek to znaczy)
  - Wyznaczniki: sufiks ndk lub dk (ną), prefiks dk
  - Kryterium Masłowa: czas. ndk może oznaczać wiele zdarzeń (jak dk) lub jedno zdarzenie w praesens historicum.
  - Inne kryteria: jeden przekład, proces i efekt, przeczenie w imperat., brak wtórnego ndk od prefiksalnego dk.

Semantyczne typy par

- Pary teliczne: *czytać : przeczytać*
  - Janek czyta X może znaczyć: ‚przeczytał X1 i przeczyta X2'
  - może też znaczyć: czytał, czytał, ale nie przeczytał
- Pary wielokrotne i ingresywne
- gubić : zgubić, widzieć : zobaczyć, mówić : powiedzieć
  - *Janek gubi X* może znaczyć: ‚zgubił i zgubi'
  - nie: **gubił X, ale nie zgubił*
- Trójki aspektowe: *tworzyć : stworzyć : stwarzać*

(naturalny etap rozwoju)

  - *Tworzy* X i *stwarza* X może znaczyć ‚stworzył i stworzy'
  - W trójkach czasowników ruchu różne odpowiedniki PH oraz iteratywne - pójść : iść/chodzić

Czego nie uwzględniamy?

- Klasy akcjonalne (Vendler) warunkujące typ opozycji:
  - State (*stać*), activity (*tańczyć*), accomplishment (wy/schnąć, prze/czytać), achievement (z/gubić, zwyciężyć, -ać), semelfaktywa (*kopnąć*, -ać), por. klasyfikacja GWJP
  - Czasowniki teliczne i nieteliczne
  - Rekategoryzacja, np. *umierać*: achiev. > accompl.
- Znaczenia szczegółowe aspektu
  - Ndk: trwałe, iteratywne, potencjalne, ogólnofakt.
  - Dk: konkretnofaktyczne, przykładowe
- Odpowiedniki delimitatywne na *po*- i inne rodzaje akcji (Aktionsarty)

Pary aspektowe w słownikach polskich (tabela 1)

- Opozycje prefiksalne
  - W dwóch hasłach bez odnośników (słowniki XIX w., SJPDor, SJPPWN, PSPP)
  - Z odnośnikiem dk > ndk (SWJP, WSJP)
  - Z odnośnikiem ndk > dk (ISJP, WSJP)
  - W jednym haśle (SWJP)
- Opozycje sufiksalne
  - Definicja ndk (WSJP)
  - Definicja dk (ISJP)
  - Odrębne hasła (WSJP)
- Nie ma znaczenia typ opozycji telicznej lub nietelicznej

Pary aspektowe w słownikach rosyjskich

- Dal (1861-66)
  - Pary prefiksalne w oddzielnych hasłach bez odnośników (hasła w gniazdach jak u Lindego).
  - Pary sufiksalne w jednym haśle.
- Ušakov (1935>), Ožegov i Švedova (1949>)
  - Pary prefiksalne w oddzielnych hasłach, odośnik od dk do ndk np. (napisat’ > pisat’).
  - Pary sufiksalne: od ndk tylko odośnik do dk.
  - Niekonsekwentne, np. obrisovyvat’ i obrisovat’ tylko odnośniki bez opisu (Uš.).
- Efremova (2000)
  - Pary prefiksalne w odrębnych hasłach bez odnośników, sufiksalne w jednym haśle.

Pary w słownikach i korpusach

- Niekonsekwentny układ par aspektowych w różnych słownikach nie może być wzorcem dla spójnego przedstawienia par w korpusach.

- W korpusach języków słowiańskich aspekt taguje się jako kategorię niezależną od par.

Tagi aspektu w korpusach słowiańskich

	Wartości dk i ndk	Dwuaspektowe	Aspekt gerundium
Polski	+	-	+
Czeski	+	+	-
Rosyjski	+	+	-
Słowacki	+	+	-
Słoweński	+	(+)	-
Bułgarski	+	-	-

Tagowanie par

- Przypisanie każdej formy czasownikowej (reprezentacji leksemu) do pary aspektowej oraz określnie formalnych wyznaczników opozycji pozwala.
  - Lepiej zbadać statystykę użycia aspektu przy tożsamym znaczeniu leksykalnym, np. w przekładach,
  - Porównać zachowanie par prefiksalnych i sufiksalnych
  - Zbadać semantykę poszczególnych przedrostków i sufiksów jako wyznaczników aspektu.
- Automatyczne tagowanie par aspektowych jest możliwe na poziomie słownika.
  - Tagowanie znaczeń szczegółowych wymaga pracy ręcznej (w wybranym podkorpusie.

Korpus polsko-niemiecki UW/JGU

- Korpus równoległy zrównoważony chronologicznie i gatunkowo z podkorpusem tekstów prawnych.
- UW i Uniw. Gutenberga w Moguncji/Germersheim
- Na razie zindeksowane 1,5 mln słów
  - w tym miesięcznik Dialog, starsza klasyka, kodeksy
- Tagowany według standardów NKJP (Poliqarp) i STTS.
- Wyszukuje słowa, leksemy, kategorie gramatyczne.
- Alignment word to word.
- Strony: parasolcorpus.org/KrakowMW; parasolcorpus.org/Beethoven/#!/
- Pary tagowane w korpusie próbnym 500 tys.

Charakterystyka aspektowa

- Każdej formie czasownika tager Poliqarp przypisuje charakterystykę gram., m.in.: perf/imperf
  - (nie ma wartości dwuaspekt., aresztować, amputować, są dk)
- Nowa charakterystyka aspektowa dotyczy leksemu w bazie i jego form w tekście
- [atag] to powtórzona wartość aspektu oraz wyznacznik formalny
  - Np. [atag=i:si] (ndk, simplex), p:pr (dk prefiks)
- [superlemma] odnosi czsownik do pary (lub par)
  - [charakteryzować_scharakteryzować|ucharakteryzować]
  - [stworzyć_tworzyć|stwarzać]

Wartości [atag]

- Część 1: powtarza, uszczegóławia lub koryguje [tag]
  - [i – ndk parzysty, p – dk parzysty, itn– imp. tantum, ptn – perf. tantum, bi – dwuaspekt.]
- Część 2: określa wyznacznik opozycji w parze
  - [si – simplex: czytać, odczytać, pr – przeczytać, su – morfem wyraźnie wzbogacający tylko jeden odpowiednik w parze: odczytywać, mrugnąć, gr – morfem wymienny między odpowiednikami (formalnie sufiks): rzucać, rzucić, mrugać, sp – supletywne: brać, wziąć
- Czyta [atag=i:si], odda [atag=p:si], przeczyta [atag=p:pr], mrugnie [atag=p:su], oddaję [atag=i:su], stoi [atag=itn], każe [atag=bi] (zob. tabela)

Niejednoznaczne przypisanie do pary (tabela 2)

- Różne pary prefikasalne dla różnych znaczeń ndk
  - [charakteryzować_scharakteryzować|ucharakteryzować]
  - Charakteryzuje/scharakteryzował towarzyszkę podróży krótkimi słowami.
  - Aktor charakteryzuje/ucharakteryzował się przed występem.
- Homonimy czasownikowe
  - Pełne: pochodzić [i_tn], pochodzić [p_tn]
  - Pochodzi z Krakowa. Pochodził trochę po pokoju.
  - Supletywizm form: stanie [atag=p:su][stawać_stanąć], stanie (się) [atag=p:su][stawać_stać](się)
  - Pociąg stanie za chwilę. Stanie się to jutro.

Trójki i pary

- *Stworzyć* w parze z *tworzyć* ma charakterystykę [p:pr], w parze ze stwarzać - [p:si]
  - Stworzy [atag=p:si|su][pair=tworzyć|stwarzać_stworzyć]
  - Stwarza [atag=i:su][pair=stwarzać_stworzyć]
  - Tworzy [atag=i_si][pair=tworzyć_stworzyć|utworzyć]
  - Według kryterium formalnego parą prefiks. dla *tworzyć* jest tylko utworzyć (tak w SWJP)

Co można sprawdzić?

- Jakie są proporcje aspektu w parach?
  - W NKJP ndk (w tym itn) : dk = 110025 : 59262 = 1,9
  - Niem. *schreiben* w polskiej części korpusu Parasol odpowiada 349 ndk pisać i 358 dk napisać (ca. 1 : 1).
  - Niem. *kaufen* odpowiada 193 dk *kupić* (75%) i 65 ndk *kupować*.
  - Stosunek ndk : dk w parach aspekt. = 24885 : 25145
    - [atag contains "i(:.*)?"] imp. si/su/gr/sp 24885
    - [atag contains "p(:.*)?"] pef. pr/su/gr/sp 25145
    - [atag contains "i(_.*)?"] imperf. tantum 40693
    - [atag contains "p(_.*)?"] perf. tantum 75
  - W parach aspektowych proporcję frekwencji 1 : 1
  - Czas. achievement, np. kupić/kupować, mają wyższy udział dk.

Pary sufiksalne i prefiksalne

- Pary prefiksalne są uznane za mniej gramatyczne niż sufiksalne (Maslov, Isačenko, Laskowski)
- Janda i Ljaševskaja (2011) zbadały w NKRJa profile gramatyczne tysięcy par prefiksalnych oraz sufiksalnych i potwierdziły jednorodność wszystkich par
  - Aspectual Pairs in the RNC, Scando-Slavica 57 (2011), 201-215.
  - Profil gramatyczny to układ odpowiednich form, np. przeszłych, nieprzeszłych, różnych trybów, imiesłowów, form nieosob.
- Obliczymy proporcje form przeszłych i nieprzesz. w korpusie polskim i porównamy z danymi rosyjskimi.
  - Niejednoznaczność przypisania formy wyrazowej do opozycji leksemów zaciemnia obraz w obu badaniach.

Zapytania o opozycje prefiks.

- Formy przesz. czas. dk prefiksalnych: 2183

[tag=".*praet.*" & atag contains "p.pr"] *napisał*

- Formy przesz. czas. ndk bezsufiks. (si): 2526:

[tag=".*praet.*" & atag contains "i.si"] *pisał*

- Formy nieprzesz. czas. dk prefiksalnych: 512

[tag=".*fin.*" & atag contains "p.pr"] *napisze*

- Formy nieprzesz. czas. ndk bezsufiks. (si): 3633

[tag=".*fin.*" & atag contains "i.si"] *pisze*

Zapytania o opozycje sufiksalne

- Formy przesz. czas ndk sufik. (su lub gr): 983

[tag=".*praet.*" & atag contains "i.(su|gr)"]

- Formy przesz. czas dk bezprefiks.: 3250

[tag=".*praet.*" & atag contains "p.(si|gr)"]

- Formy nieprzesz. cas ndk sufik. (su lub gr): 1519

[tag=".*fin.*" & atag contains "i.(su|gr)"]

- Formy nieprzesz. czas dk bezprefiks.: 517

[tag=".*fin.*" & atag contains "p.(si|gr)"]

Zapytania o opozycje sufiksalne

- Formy przesz. czas ndk sufik. (su lub gr): 1778

[tag=".*praet.*" & atag contains "i.(su|gr)"] *zostawał*

- Formy przesz. czas dk bezprefiks.: 7745

[tag=".*praet.*" & atag contains "p.(si|gr)"] *został*

- Formy nieprzesz. czas ndk sufik. (su lub gr): 3718

[tag=".*fin.*" & atag contains "i.(su|gr)"] *zostaje*

- Formy nieprzesz. czas dk bezprefiks.: 1746

[tag=".*fin.*" & atag contains "p.(si|gr)"] *zostanie*

Opozycje prefiksalne i sufiksalne polskie i rosyjskie

- Dane procentowe w tabeli polskiej i rosyjskiej są podobne.
- Mimo niereprezentatywności obecnego Korpusu P-N świadczy to o podobnej relacji między opozycjami prefiksalnymi i sufiksalnymi w obu językach.
- Tagowanie par aspektowych sprawia, że takie badania są łatwiejsze niż to opisały Janda i Ljaševskaja.