

# Wytyczne KPWr

## *Koreferencja*

Osoba odpowiedzialna	Marek Maziarz
Udział	Marek Maziarz, Michał Marcińczuk, Marcin Oleksy, Maciej Piasecki, Adam Radziszewski, Joanna Nowak, Adam Wardyński, Jan Wiczorek

[Skróty](#)

[Wstępne założenia](#)

[Frazy łączone koreferencją](#)

[typ PN-PN](#)

[AgP - chunki poziomu podstawowego](#)

[NP - \(maksymalne\) frazy nominalne chunkera](#)

[Frazy AgP czy głowy fraz AgP?](#)

[Przypadki graniczne - zasady szczegółowe](#)

[Fraza nominalna nadrzędna a frazy nominalne podrzędne \(zagnieżdżone\)](#)

[Apozycje nieuzgodnione - casus "rzeki Bystrzyca"](#)

[Fraza nominalna a zdanie względne](#)

[Koreferencja a metonimia](#)

[Anafora typu konotacyjnego \(sense\)](#)

[Koreferencja a sytuacji \(zdarzenia\)](#)

[Użycia predykatywne](#)

[Lista predykatów, które uznajemy za ośrodki predykcji](#)

[Użycia dzierżawcze](#)

[Znakowanie podmiotów domyślnych](#)

[Założenia](#)

[Uwagi techniczne](#)

[Analiza przypadków szczegółowych](#)

## Skróty

ref – koreferencja

ident – koreferencja identycznościowa

bridg – bridging, relacja semantyczna oparta na metonimii, bliska mero-/holonomii<sup>1</sup>

PN – nazwa własna (*proper name*)<sup>2</sup>

NP – fraza rzeczownikowa definiowana zgodnie z wytycznymi do anotacji na potrzeby chunkera Adama Radziszewskiego

AgP – fraza uzgodniona (*agreement phrase*), fraza uzgodniona pod względem przypadku, rodzaju i liczby z ew. dodatkiem wyrazów nieodmiennych ze wzgl. na przypadek, rodzaj i liczbę (adv lub qub),

definiujemy ją zgodnie z wytycznymi do anotacji na potrzeby chunkera Adama Radziszewskiego

Pron – zaimek osobowy (*ja, ty, my, wy, on, ona, ono*) lub wskazujący (*ten, ta, to; ów, owa, owo*), także zaimki przysłowne określone oznaczające miejsca (*tu, tam, stamtąd, stąd, tamtędy, tędy*)

Ppron – zaimek osobowy (*ja, ty, my, wy, on, ona, ono*)

∅ – podmiot domyślny

DE – poziom znakowania koreferencyjnego (od *discourse entity*).

podkreślenie - podkreślenie oznacza anaformik

[ ] - nawiasy określają granice fraz

\* - gwiazdką oznaczamy niewłaściwe anotacje (przykłady błędne)

CN - rzeczownik pospolity (od *common noun*)

APP - apozycja, definicja w zgodzie z wytycznymi do anotacji na potrzeby chunkera Adama Radziszewskiego

DD - deskrypcja określona

ID - deskrypcja nieokreślona

## Wstępne założenia

Przyjmujemy następujące kryteria wyznaczania relacji anaforycznych:

- Znakować będziemy korpus powstający w ramach projektu SYNAT, tj. InfiKorp. Próbkki mają długość ok. 300 wyrazów tekstowych.
- Znakujemy wyłącznie podtyp identycznościowy typu *ref* (identyczność referenta dwóch wyrażen języka, = podtyp *ident*).

---

<sup>1</sup> *bridging* to relacja pomiędzy zbiorem, a elementem tego zbioru, np. w konstrukcji:

wielcy polscy poeci:

- Juliusz Słowacki
- Adam Mickiewicz
- Wincenty Pol

to np. relacja pomiędzy [Juliusz Słowacki] a [poeci]. Por. [Koreferencja a metonimia](#).

<sup>2</sup> PN to w zasadzie nazwa własna, ale zdarzają się np. nazwy narodowości, typu *Francuz*. Szerzej o tym w podrozdziale poświęconym frazom NER-a (poniżej).

- Ponieważ nie dysponujemy użytecznym parserem głębokim dla języka polskiego, nie będzie możliwe znakowanie relacji pomiędzy równoznacznymi (chodzi tu o znaczenie referencyjne, tj. mającymi ten sam referent) frazami nominalnymi (nazwa własna - fraza rzeczownikowa). Z całego drzewa relacji nie będzie bowiem możliwości wyciągnięcia dowolnego poddrzewa składniowego. Skorzystając z dostępnych parserów płytkich, tj. chunkera Adama Radziszewskiego i Spejda. Zdecydowaliśmy się na użycie politechnicznego chunkera, Spejd będzie wykorzystywany na etapie uczenia mechanizmów sztucznej inteligencji. Znakować będziemy zatem pomiędzy możliwymi chunkami (poziom parsera płytkiego) a nazwami własnymi (poziom NER).
- Zakładamy, że jednostkami podstawowymi na poziomie chunkingu będą najmniejsze uzgodnione frazy rzeczownikowe lub przyimkowo-rzeczownikowe nazwane przez nas AgP. Głowa AgP musi być tożsama z głową frazy, którą by zaznaczył parser głęboki. Na przykład we fragmencie

*Piotr Wielki był carem Rosji. Władca największego państwa w Europie lubił kąpać się w Morzu Czarnym*

nie możemy oznaczyć jako jednej frazy *Władca największego państwa w Europie* ani *największego państwa w Europie*. Nie przewidują tego nasze wytyczne do znakowania na potrzeby płytkiego parsingu. Poniżej to “idealne” parsowanie w przypadku koreferencji oznaczamy przez gwiazdkę (na poziomie DE\*):

PN: *[Piotr Wielki] był carem [Rosji]. Władca największego państwa w [Europie] lubił kąpać się w [Morzu Czarnym].*

AgP: *[Piotr Wielki] był [carem] [Rosji]. [Władca] [największego państwa] [w Europie] lubił kąpać się [w Morzu Czarnym].*

DE\*: *[Piotr Wielki - 2] był carem [Rosji - 1]. \*{Władca {największego państwa w Europie - 1} - 2} lubił kąpać się w Morzu Czarnym.*

DE: *[Piotr Wielki - 2] był carem [Rosji - 1]. Władca-2 największego państwa-1 w Europie lubił kąpać się w Morzu Czarnym*

Wyjaśnienia skrótów: PN - nazwa własna (poziom NER), AgP - rzeczownikowa bądź przyimkowo-rzeczownikowa fraza nominalna (poziom płytkiego parsera), DE - poziom znakowania anafory, na tym poziomie nawiasem wydzielamy poprzednik, a podkreśleniem - anafornik, numerki oznaczają, który anafornik odpowiada któremu poprzednikowi.

Nawiasem wąsiastym { } oznaczamy granice nakładających się “idealnych” anaforników. Niestety, nie dysponujemy idealnym parserem, który doprowadziłby nas do fraz zagnieżdżonych (do nawiasów wąsiastych). Na poziomie tego nieistniejącego parsera głębokiego frazy *Władca największego państwa w Europie* i podrzędna fraza *największego państwa w Europie* (tłustym drukiem zaznaczamy głowy) byłyby dostępne (warstwa DE\*), my - niestety - zatrzymujemy się z przyczyn obiektywnych na poziomie fraz chunkerowych, które zawierają obie te głowy (warstwa DE). Podstawą jest tu zatem poziom AgP:

AgP: [*Władca*] [*największego państwa*] [*w Europie*].

Pierwszą frazę łączymy koreferencją z nazwą własną *Piotr Wielki*, drugą frazę łączymy koreferencją z nazwą *Rosja*. Rozbudowana definicja fraz AgP pojawia się w podrozdziale poświęconym podtypowi koreferencji PN-AgP.

- Zajmujemy się wyłącznie relacją pomiędzy nazwą własną (proper name = PN) a frazą AgP (uzgodnioną, definiowaną identycznie jak w płytkim parsingu) oraz podmiotem domyślnym.
- Z odmian referencji identycznościowej (*ident*) bierzemy pod uwagę następujące typy relacji anaforycznych:
  - relację pomiędzy dwiema frazami nominalnymi, z których jedna jest nazwą własną, a druga frazą uzgodnioną chunkera politechnicznego (typ. PN-AgP);
  - relację pomiędzy nazwą własną a podmiotem domyślnym (typ PN-Ø);
  - relację pomiędzy nazwą własną a frazą uzgodnioną zawierającą zaimek osobowy (Ppron12 oraz Ppron3), wskazujący (*ten, ta, to, ów, owa, owo, tamten, tamta, tamto*) oraz określone zaimki przysłowne oznaczające miejsce (*tu, tam, stąd i stamtąd, tamtędy, tędy*) (typ PN-Pron);
  - relację pomiędzy dwiema nazwami własnymi (typ PN-PN).
- Jeżeli żadna z fraz nominalnych nie jest nazwą własną, koreferencja pomiędzy frazami nie będzie znakowana.
- Powyższy warunek w szczególności oznacza, że:
  - nie łączymy frazy nominalnej niebędącej nazwą własną z deskrypcją określoną;
  - nie łączymy frazy nominalnej niebędącej nazwą własną z deskrypcją nieokreśloną;
  - nie łączymy frazy nominalnej niebędącej nazwą własną z rzeczownikiem pospolitym;
  - podobne ograniczenia nakładamy na odmianę koreferencji typu PN-Pron.
- Łączymy za to:
  - w podtypie PN-PN:
    - nazwę własną z inną nazwą własną;
  - w podtypie PN-AgP:
    - nazwę własną z deskrypcją określoną (PN-AgP:DD),
    - nazwę własną z deskrypcją nieokreśloną (PN-AgP:ID),
    - nazwę własną z rzeczownikiem pospolitym (PN-AgP:CN);
  - w podtypie PN-Pron:
    - nazwę własną z zaimkiem;
  - w podtypie PN-Ø:
    - nazwę własną z podmiotem domyślnym (ten podtyp omawiamy bardziej szczegółowo poniżej).
- Warunkiem koniecznym jest wystąpienie powyższych fraz i wyznaczników w identycznym znaczeniu referencyjnym, tj. zgodność referentów wyrażen języka.

Statystyki podtypów przedstawia poniższa tabela (sporządzone przez Adama Wardyńskiego na podstawie 10 próbnie zaanotowanych próbek z korpusu InfiKorp, podkorpus Wikipedii, URL: <http://nlp.pwr.wroc.pl/inforex/index.php?page=report&corpus=7&id=99883>).

Podtyp anafory	liczba instancji w dziesięciu próbkach InfiKorp	odsetek instancji ref ident
ref ident PN-PN	36	20,2%
ref ident PN-AgP	46	25,8%
ref ident PN-Pron	7	3,9%
ref ident PN-Ø	89	50,0%

### Znakowanie odbywa się zawsze do pierwszej jednostki w tekście!!!

W przypadku, gdy pojawia się jednostka wcześniej, np. w nagłówku artykułu z Wikipedii, znakujemy i tak do pierwszej jednostki z głównego tekstu (a więc drugiej w próbce), a dopiero potem od nagłówka do tej wyróżnionej jednostki z tekstu właściwego.

## Frazy łączone koreferencją

W koreferencji łączyć będziemy warstwę bytów nazwanych z chunkami. Sam proces znakowania poprzedzony będzie fazą bootstrappingu, na tym etapie zaznaczymy nazwy własne (mechanizm NER-a) oraz wybrane chunki (płytki parser Adama Radziszewskiego, obecnie ma 85% skuteczności w wykrywaniu fraz uzgodnionych, trwają prace nad bardziej skutecznymi algorytmami). Z chunków liczyć się będą przede wszystkim frazy AgP (w przypadku trzech podtypów: PN-AgP oraz PN-Pron), a także VP (w przypadku znakowania podtypu PN-Ø).

## typ PN-PN

Ten podtyp ogranicza się do połączeń pomiędzy nazwami własnymi (warstwa NER):

PN: *[Bronisław Komorowski] przeciął wstęgę. [Komorowski] zapamiętał tę uroczystość.*

DE: *[Bronisław Komorowski] przeciął wstęgę. Komorowski zapamiętał tę uroczystość.*

PN: *Ukończył w 1991 studia na Wydziale Prawniczym [Katolickiego Uniwersytetu Lubelskiego].  
(...) Pracuje jako asystent w Katedrze Prawa Cywilnego na [KUL].*

DE: *Ukończył w 1991 studia na Wydziale Prawniczym [Katolickiego Uniwersytetu Lubelskiego].  
 (...) Pracuje jako asystent w Katedrze Prawa Cywilnego na KUL*

Często te połączenia będą miały charakter tożsamościowy:

PN: *W swoim pierwszym sezonie w NBA w barwach [Suns] wystąpił w 49 meczach. (...) Rozegrał w barwach [Suns] zaledwie 3 mecze.*

DE: *W swoim pierwszym sezonie w NBA w barwach [Suns] wystąpił w 49 meczach. (...) Rozegrał w barwach Suns zaledwie 3 mecze.*

Przed przystąpieniem do znakowania tego typu koreferencji automatycznie połączymy równobrzmiące nazwy własne (bootstrapping). Ten podtyp obejmuje ok. 50% wszystkich instancji koreferencji.

## AgP - chunki poziomu podstawowego

Zanim przejdziemy do opisanego kolejnych dwóch podtypów koreferencji, tj. PN-AgP oraz PN-Prep, musimy w pierwszej kolejności zapoznać się z podstawowymi typami fraz, które mają być rozpoznawane przez chunker. Fraza AgP jest przez nas definiowana jako fraza przyimkowo-rzeczownikowa uzgodniona na przypadku, liczbie i rodzaju. W szczególności w jednym AgP znajdują się:

- rzeczownik i jego modyfikatory przymiotnikowe:

AgP: *[czerwony zachód] [słońca], [zielone jabłuszko], [kreatywna księgowość];*

- rzeczownik w związku zgody z liczebnikami porządkowymi:

AgP: *[pierwsza fuga] [Beethovena], [setna prośba] [drugiego premiera];*

- rzeczownik z modyfikatorem imiesłowowym:

AgP: *[pogiębiony aktor], [zasłuchana primadonna], [rozochocony satyr];*

- rzeczowniki w apozycji (także z określnikami):

AgP: *[pan doktor konik polny], [rozochocona pani doktor];*

- wymienione powyżej części mowy w połączeniu z nieodmiennymi (przez przypadek, liczbę i rodzaj) częściami mowy (tj. qub lub adv), jeśli tylko są konotowane przez inne składniki frazy AgP:

AgP: *[pogiębiony bardzo doktor], [zasłuchana wciąż primadonna], [nieco kreatywna księgowa];*

- do tych fraz zaliczamy też przyimki rządzące przypadkiem rzeczownika:

AgP: *[doktor pognębiony] [przez lata niewoli];*

- liczebniki główne (klasa num w KIPI) traktujemy jak rzeczowniki, mogą być zatem ośrodkami AgP:

AgP: *[dwoma żołnierzami] interesowały się [dwie panienki] [z okienka],*

NP: *[dwoma żołnierzami] interesowały się [dwie panienki z okienka];*

- podobnie jako rzeczowniki są traktowane zaimki rzeczowne *ktoś, coś* (klasa subst w KIPI):

AgP: *[Ktoś] powiedział: [Coś] tam widać?*

NP: *[Ktoś] powiedział: [Coś] tam widać?<sup>3</sup>*

W rzeczywistości frazy AgP zawierające nadrzędny składniowo przyimek to frazy przyimkowe (PrepP). Frazy tniemy w miejscu, gdzie zmienia się charakter uzgodnienia (np. przypadek, rodzaj). Przyimkowo-rzeczownikowe frazy AgP mogą być nieciągłe, takie frazy zszywamy:

AgP: *[pobity bardzo - 1] [kijem bejsbolowym] [pan doktor - 1].*

## NP - (maksymalne) frazy nominalne chunkera

W anotacji składniowej nadrzędne frazy przyimkowo-rzeczownikowe (lub rzeczownikowe) nazywamy NP (*Nominal Phrase*). Definiujemy je jako największe w hierarchii składniowej poddrzewa mające dystrybucję rzeczownikową (przymiotnikową). Pełnią one w strukturze predykatowo-argumentowej rolę argumentów.

**Nie znakujemy koreferencji w obrębie jednej frazy nominalnej**, np. pomiędzy PN zawierającą głowę nadrzędnej frazy rzeczownikowej a AgP zawierającą głowę podrzędnej frazy nominalnej (gwiazdką w wierszu DE\* oznaczamy niewłaściwe anotacje, właściwe anotacje podajemy w wierszu DE):

DE\*: *\*[Piotr Wielki] zwany \*Carem Uwodzicielem*

DE: *Piotr Wielki zwany Carem Uwodzicielem*

NP: *[Piotr Wielki zwany Carem Uwodzicielem].*

W przykładzie podanym uprzednio

*Piotr Wielki był carem Rosji. Władca największego państwa w Europie lubił kąpać się w Morzu Czarnym*

<sup>3</sup> Tam to zaimek przysłowny (przypomina przysłówki), dlatego nie może być traktowany jak rzeczownik (należy do klasy qub w tagsecie KIPI).



na poziomie NP otrzymamy (dla porównania podajemy warstwę AgP, wytłuszczonym drukiem):

NP: [*Piotr Wielki*] był [*carem Rosji*]. [*Władca największego państwa w Europie*] lubił kąpać się [w *Morzu Czarnym*]

AgP: [*Piotr Wielki*] był [*carem*] [*Rosji*]. [*Władca*] [*największego państwa*] [w *Europie*] lubił kąpać się [w *Morzu Czarnym*].

Teoretycznie więc moglibyśmy znakować koreferencję w ten sposób:

DE\*: [*Piotr Wielki* - 2] był carem [*Rosji* - 1]. \*[Władca [największego państwa AgP-1] w Europie - NP-2] lubił kąpać się w Morzu Czarnym

tj. koreferencję *Władca największego państwa w Europie* <> *Piotr Wielki* na poziomie NP, a koreferencję *największe państwo w Europie* <> *Rosja* w okrojony sposób na poziomie AgP (*największe państwo* <> *Rosja*)<sup>4</sup>. W rzeczywistości jednak do poziomu NP możemy w prosty sposób dojść, korzystając z frazy AgP zawierającej głowę nadrzędnej frazy NP. Jeżeli bowiem jako anafornik do poprzednika *Piotr Wielki* wybierzemy rzeczownik *władca* (AgP), który to wyraz jest nadrzędnikiem całej frazy *władca największego państwa w Europie* (NP), to możemy być pewni, że anafornikiem jest cała ta fraza NP:

AgP: [*władca*] [*największego państwa*] [w *Europie*]

|  
∨

NP: [*władca największego państwa w Europie*]

Z tego powodu skupiamy się wyłącznie na frazach AgP w przypadku podtypów PN-AgP, PN-Pron.

## typ PN-AgP

Ten podtyp koreferencji (PN-AgP) obejmuje ok. ¼ wszystkich instancji koreferencji w tekstach (oszacowanie na podstawie Wikipedii). Oto garść przykładów:

PN: [*Bronisław Komorowski*] był nieuchwytny. *Prezydent cały dzień unikał dziennikarzy.*

AgP: [*Bronisław Komorowski*] był nieuchwytny. [*Prezydent*] [*cały dzień*] unikał [*dziennikarzy*].

VP: *Bronisław Komorowski [był] nieuchwytny. Prezydent cały dzień [unikał] dziennikarzy.*

DE: [*Bronisław Komorowski*] był nieuchwytny. *Prezydent cały dzień unikał dziennikarzy.*

PN: [*Władimir Putin*] na wakacje jeździ nad [*Morze Czarne*]. *Premier [Rosji] lubi pływać.*

AgP: [*Władimir Putin*] [*na wakacje*] jeździ [*nad Morze Czarne*]. [*Premier*] [*Rosji*] lubi pływać.

VP: *Władimir Putin na wakacje [jeździ] nad Morze Czarne. Premier Rosji [lubi pływać]*

<sup>4</sup> Przez *okrojony* mamy tu na myśli to, że fraza AgP *największe państwo* stanowi tylko część frazy “idealnej”, tj. *największe państwo w Europie*.

DE: *[Władimir Putin] na wakacje jeździ nad Morze Czarne. Premier Rosji lubi pływać.*

PN: *[Napoleon Bonaparte] był cesarzem [Francuzów]. Władca nie respektował traktatów.*<sup>5</sup>

AgP: *[Napoleon Bonaparte] był [cesarzem] [Francuzów]. [Władca] nie respektował [traktatów].*

VP: *Napoleon Bonaparte [był] cesarzem Francuzów. Władca [nie respektował] traktatów.*

DE: *[Napoleon Bonaparte] był cesarzem Francuzów. Władca nie respektował traktatów.*

W powyższych przykładach pojawia się fraza czasownikowa VP, szerzej omówimy ją przy okazji analizy podtypu ref ident PN-Ø. Znakujemy na poziomie DE *Premier*, a nie *\*Premier Rosji*, ponieważ odnosimy się wyłącznie do fraz AgP zawierających głowę frazy nominalnej, której referent jest identyczny z referentem nazwy własnej *[Władimir Putin]*.

Na późniejszym etapie uczenia maszynowego na korpusie takie okrojone frazy AgP będzie można rozszerzyć do rozpoznawanych przez chunker fraz NP na zasadzie tożsamości głów obu zagnieżdżonych fraz (wytluszczonym drukiem zaznaczamy głowy fraz<sup>6</sup>):

AgP: *[Władimir Putin] [na wakacje] jeździ [nad **Morze Czarne**]. [**Premier**] [**Rosji**] lubi pływać.*

|

∨

NP: *[**Władimir Putin**] [na wakacje] jeździ [nad **Morze Czarne**]. [**Premier Rosji**] lubi pływać*

DE: *[Władimir Putin] na wakacje jeździ nad Morze Czarne. Premier Rosji lubi pływać.*

Strzałka oznacza możliwe przejście od frazy AgP do frazy NP, o ile mają tę samą głowę. Na poziomie DE' uzyskalibyśmy koreferencję:

DE': *[Władimir Putin] na wakacje jeździ nad Morze Czarne. Premier Rosji lubi pływać*

Prim w DE' oznacza, że taką koreferencję uzyskujemy w wyniku transformacji AgP -> NP (przy tożsamości głów).

## typ PN-Pron

Szczególnym przypadkiem frazy AgP naszego chunkera jest fraza zbudowana na bazie zaimków osobowych i wskazujących. Tego typu frazy AgP określać będziemy na potrzeby anotacji anafory skrótem *Pron*. Ze względów oczywistych zaimki łączyć się będą wyłącznie z nazwami własnymi (znakujemy wyłącznie relacje typu PN-X):

AgP: *[Bronisław Komorowski] szedł [pewnym krokiem]. Tylko [on] znał [drogę] [na skróty].*

NP: *[Bronisław Komorowski] szedł [pewnym krokiem]. Tylko [on] znał [drogę na skróty].*

<sup>5</sup> Co prawda *Francuz* nie jest nazwą własną (może występować w zdaniu *X jest Francuzem* w pozycji predykatu), jednak na poziomie NER-a jest znakowany jako fraza PN.

<sup>6</sup> W przypadku apozycji przyjmujemy arbitralnie, że głową frazy jest pierwszy z lewej rzeczownik.

PN: *[Bronisław Komorowski] szedł pewnym krokiem. Tylko on znalazł drogę na skróty.*

DE: *[Bronisław Komorowski] szedł pewnym krokiem. Tylko on znalazł drogę na skróty.*

AgP = NP: *[Adam] powiedział: „To [ja]!”*

PN: *[Adam] powiedział: „To ja!”*

DE: *[Adam] powiedział: „To ja!”*

Formy *jego*, *jej*, *ich* użyte w funkcji przydawki Laskowski (1998: 339) uznaje za zaimki dzierżawcze nieodmienne. Morfeusz nie czyni rozróżnienia pomiędzy użyciem przydawkowym (*z jego fuzją*, *z jej dwulufką*, *z ich bronią krótką*), a użyciem dopełnieniowym (*nie zastrzelił ani jego*, *ani jej*, *ani tym bardziej ich*), traktując oba te użycia jako formy jednego leksemu (*on*, *ona*, *oni*, ppron3).

My idziemy tropem Morfeusza, uznając użycia przydawkowe za dopełniacze odpowiednich zaimków osobowych. Takie użycia znakujemy:

[Piotr-1] miał fuzję, a [Ania-2] dwulufkę. Jego-1 broń wystrzeliła, a jej-2 - nie.

Por:

[On] miał fuzję, a [ona] dwulufkę. Piotra-1 broń wystrzeliła, a Ani-2 - nie.

Analogicznie znakujemy użycia dzierżawcze (dopełniacz dzierżawczy) od rzeczowników pospolitych i nazw własnych (por. niżej).

Wyjątkowo traktujemy zaimki przysłówne określone *tu* i *tam*, *stamtąd* i *stąd*, *tędy*, *tamtędy*. Jako przysłówki (klasa  $\text{qub}$  tagsetu KIPI) nie może tworzyć fraz AgP, NP czy AdjP albo VP (nie może być ich głową), może jednak - jako określenie czasownika czy imiennej części mowy - wchodzić w ich skład:

VP: *Wyspa Opatowicka [zaludniła się] pijanymi studentami. Spokojni turyści [uciekali stamtąd].*

AgP: *[Wyspa Opatowicka] zaludniła się [pijanymi studentami]. [Spokojni turyści] uciekali stamtąd.*

PN: *[Wyspa Opatowicka] zaludniła się pijanymi studentami. Spokojni turyści uciekali stamtąd.*

DE: *[Wyspa Opatowicka] zaludniła się pijanymi studentami. Spokojni turyści uciekali stamtąd.*

Nazwę *Wyspa Opatowicka* będziemy łączyć z tymi zaimkami przysłownymi na poziomie tokenów, a nie fraz.

## typ PN-Ø

Ostatnim wyróżnionym przez nas typem anafory jest relacja z podmiotem domyślnym. Relację tę na poziomie anotacji znakować będziemy pomiędzy orzeczeniem z podmiotem domyślnym a nazwą własną, której podmiot domyślny odpowiada. Na etapie maszynowego uczenia do orzeczeń dodane zostaną elementy zerowe, a połączenia z nazwą własną zostaną na nie przesunięte z fraz czasownikowych chunkera (VP)

AgP=NP: *[Bronisław Komorowski] zreferował [trzy projekty]. Wyglądał profesjonalnie.*  
 VP: *Bronisław Komorowski [zreferował] trzy projekty. [Wyglądał] profesjonalnie.*  
 DE: *[Bronisław Komorowski] zreferował trzy projekty. Wyglądał profesjonalnie.*

Ponieważ na poziomie znakowania składniowego nie uwzględniamy połączeń podmiotu w zdaniu nadrzędnym z orzeczeniem w wypowiedzeniu podrzędnym, takie połączenia możemy znakować (= znakujemy) jako PN-Ø:

NP: *[Szpital Dzieciątka Jezus], [który] mieścił się [przy ulicy Lindleya], przeniesiono [poza miasto].*  
 DE: *[Szpital Dzieciątka Jezus], który mieścił się przy ulicy Lindleya, przeniesiono poza miasto.*

Tak samo należy postąpić w wypowiedzeniu złożonym współrzędnie:

NP:  
 - *[Witek]! [Mój złoty Witek]!*  
*[Chłopak] się obejrzał, [źrebaka], [na którym] jechał, powstrzymał i mruknął: – Cegój! A gdzie!... To psinorody dopiero, zaro w skodę kiej swynie!*  
*Pchnął [swego konia], obegnał [rozpraszające się] [po drodze i zbożu] [stado źrebiąt], spędził [w kupę] i podjechał znowu [pod parkan okalający] [ogród].*

DE:  
 - *[Witek]! Mój złoty [Witek]!*  
*Chłopak się obejrzał, źrebaka, na którym jechał, powstrzymał i mruknął: – Cegój! A gdzie!... To psinorody dopiero, zaro w skodę kiej swynie!*  
*Pchnął swego konia, obegnał rozpraszające się po drodze i zbożu stado źrebiąt, spędził w kupę i podjechał znowu pod parkan okalający ogród.*

## Frazy AgP czy głowy fraz AgP?

Znakowanie składniowe, na potrzeby NER-a oraz koreferencji przebiega równolegle. Z tego względu nie jesteśmy w stanie zagwarantować, że znakowanie koreferencji będzie poprzedzone znakowaniem składniowym. Chcemy mieć za to teksty korpusowe oznakowane pod kątem bytów nazwanych przed anotowaniem koreferencją.

Rodzi to pewne komplikacje. O ile bowiem jedna część procesu anotowania (byty nazwane) będzie w anotorze (Inforex) dostępna, o tyle drugą część (poziom składniowy chunkera) trzeba będzie sobie odtwarzać. By uniknąć zamieszania, anotor powinien mieć doświadczenie w znakowaniu chunków.

W przypadku gdy tekst danego podkorpusu ma oznakowane frazy AgP czy VP, postępujemy zgodnie z tekstem tej instrukcji. Natomiast gdy znakowania składniowego jeszcze nie przeprowadzono, znakujemy wyłącznie głowy (za głowę uznajemy pojedynczy token), mając w pamięci kształt właściwej frazy AgP

czy VP, tj. tej, której głową jest znakowany token. Po “wyobrażeniu” sobie docelowego kształtu AgP czy VP możemy z powodzeniem stosować wszystkie reguły tej instrukcji.

## Przypadki graniczne - zasady szczegółowe

### zazębianie się typów PN-AgP oraz PN-PN

1. Zdarzyć się może, że w anotowanym tekście pojawi się możliwość znakowania na kilka możliwych sposobów, chociażby w takim przykładzie:

*Achilles szalejący zabił Hektora. Wojownik Achilles pastwił się nad ciałem.*

Mamy tu dwie frazy nominalne powiązane relacją identycznościową:

*Achilles szalejący - wojownik Achilles*

oraz

*Hektor - ciało*

O ile w drugim przypadku sprawa jest jasna - mamy do czynienia z typem PN-AgP, o tyle w pierwszym przypadku nie do końca wiadomo, czy znakować typ PN-PN, czy może PN-AgP. Możliwości przedstawiamy poniżej:

AgP: [*Achilles szalejący*] ... [*wojownik Achilles*]

PN: [*Achilles*] szalejący ... wojownik [*Achilles*]

DE: [*Achilles*] szalejący ... wojownik Achilles - typ PN-AgP

[*Achilles*] szalejący ... wojownik Achilles - typ PN-PN

We wszystkich tych przypadkach będziemy wybierać znakowanie PN-PN:

#### **PIERWSZEŃSTWO MA ZNAKOWANIE POMIĘDZY PN (PPN)**

▣ W związku z ograniczeniami nakładanymi przez NER i chunking na kształt fraz (dysponujemy frazami PN, AgP, niekiedy również NP) pojawia się możliwość znakowania koreferencji na kilka różnych sposobów. Jeżeli jest to możliwe, znakujemy najpierw pomiędzy frazami NER. ▣

Zgodnie z tą zasadą wybieramy podtyp PN-PN.

Inny przykład:

AgP: [wódz Achilles] ... [wojownik Achilles]  
 PN: wódz [Achilles] ... wojownik [Achilles]  
 DE: [wódz Achilles] ... wojownik Achilles - typ AgP-PN  
       wódz [Achilles] ... wojownik Achilles - typ PN-AgP  
       wódz [Achilles] ... wojownik Achilles - typ PN-PN

Z trzech możliwych sposobów znakowania koreferencji, tj. AgP-PN, PN-AgP i PN-PN, wybieramy podtyp PN-PN.

## 2. Rozbudujemy przykład "Homerycki":

*Achilles szalejący zabił Hektora. Wojownik Achilles pastwił się nad ciałem. Wódz Myrmidonów wzruszył się potem nad niedolą Priamidy, syna Priama.*

Teraz nasze koreferencje rozbudowały się do dwóch przeplatających się łańcuchów:

(1) *Achilles szalejący* - (2) *wojownik Achilles* - (3) *wódz Myrmidonów*

oraz

(1) *Hektor* - (2) *ciało* - (3) *Priamida, syn Priama*.

**2.1.** Łańcuch Hektorowy jest niejednoznaczny. Co prawda, w pierwszej parze (1←2) koreferencja przyjmuje jednoznaczną postać:

AgP: [Hektor] - [ciało]  
 NP: [Hektor] - [ciało]  
 PN: [Hektor] - ciało  
 DE: [Hektor] - ciało - typ PN-AgP.

Ale w przypadku trzeciej frazy rzecz się komplikuje. Mamy możliwość podpięcia się pod *Hektora* (1←3) bądź pod *ciało* (2←3):

AgP: [Hektor] - [Priamida], [syn] [Priama]  
 NP: [Hektor] - [Priamida, syn Priama]  
 PN: [Hektor] - [Priamida], syn [Priama]  
 DE: [Hektor] - Priamida, syn Priama - typ PN-PN  
       [Hektor] - Priamida, syn Priama - typ PN-AgP

Zwróćmy uwagę, że znakowanie *Priamida - syn* nie jest poprawne, ponieważ zmuszałoby nas do znakowania relacji WEWNĄTRZ tej samej frazy przyimkowo-rzeczownikowej NP, co z zasady wykluczamy:

NP: [Priamida, syn Priama]  
 AgP: [Priamida], [syn] [Priama]  
 DE\*: [ ] \*PN-AgP

Precyzuje to reguła:

### **NIE ZNAKUJEMY KOREFERENCJI WEWNĄTRZ TEJ SAMEJ NP (WEW-NP)**

▣ Nawet jeśli możliwe jest powiązanie AgP z PN wewnątrz tej samej NP (nadrzędnej frazy przyimkowo-rzeczownikowej chunkera), zabraniamy tego. W szczególnych przypadkach unikamy łączenia głowy frazy NP z frazami związanymi z dopowiedzeniami, wtrąceniami, wyznaczanymi przez nawiasy, przecinki i myślniki. Szerzej o włączaniu dopowiedzeń i wtrąceń do nadrzędnej NP piszemy w opisie anotacji na potrzeby chunkera. Zasada nie zabrania jednak łączenia dowolnej frazy podrzędnej względem NP (typu AgP, NER) z dowolną frazą s p o z a NP. ▣

Na podstawie reguły WEW-NP wykluczamy znakowanie *Priamida* ← *syn*. Możliwe są jednak jeszcze połączenia *syn* → *Hektor* (3b→1) oraz *Priamida* → *Hektor* (3a→1). Wybieramy znakowanie typu PN-PN.

Pozostaje jeszcze problem możliwości połączenia rzeczownika *ciało* z jedną z fraz PN. Mamy dwie możliwości podłączenia koreferencją ostatniej frazy *Priamida, syn Priama* albo z *Hektorem* (3a→1), albo z *ciałem* (3a→2):

DE: [*Hektor*] - [*ciało*] - [*Priamida, syn Priama*]

| [ ] [ ] AgP-PN  
 | [ ] PN-PN

Obie możliwości są równoważne. Którą wybrać? Wybieramy połączenie z pierwszym użyciem PN o danej referencji w tekście. W tym przypadku z *Hektorem*.

### **ŁĄCZYMY ZAWSZE Z PIERWSZYM WYSTĄPIENIEM PN W DANYM ZNACZENIU W TEKŚCIE.**

Uwaga: pomijamy jednak PN występujące w nagłówku tekstu!

## **2.2. Łańcuch Achillesowy (1) *Achilles szalejący* - (2) *wojownik Achilles* - (3) *wódz Myrmidonów***

AgP: [*Achilles szalejący*] ... [*wojownik Achilles*] ... [*wódz*] [*Myrmidonów*]  
 PN: [*Achilles*] *szalejący* ... *wojownik* [*Achilles*] ... *wódz* [*Myrmidonów*]  
 NP: [*Achilles szalejący*] ... [*wojownik Achilles*] ... [*wódz Myrmidonów*]

również będziemy analizować od lewej do prawej (w porządku czytania), najpierw koreferencję 1←2:

DE: *[Achilles] szalejący ... wojownik Achilles* - typ PN-AgP  
*[Achilles szalejący] ... wojownik Achilles* - typ AgP-PN  
*[Achilles] szalejący ... wojownik Achilles* - typ PN-PN

Wybieramy typ PN-PN.

Znakować będziemy tak (po uwzględnieniu całego ciągu koreferencyjnego):

DE: *[Achilles] szalejący - wojownik Achilles - wódz Myrmidonów*  
 | \_\_\_\_\_ | PN-PN |  
 | \_\_\_\_\_ | PN-AgP

Czyli wybieramy znakowanie PN-PN oraz PN-AgP do pierwszej frazy NER-a.

### 3. Gdyby w tekście pojawiły się frazy

(1) *Achilles szalejący* - (2) *boski Achilles* - (3) *wódz Myrmidonów Achilles*,

mające w NER i w anotacji “chunkerowej” następujące opisy

PN: *[Achilles] szalejący - boski [Achilles] - wódz [Myrmidonów] [Achilles]*  
 AgP: *[Achilles szalejący] - [boski Achilles] - [wódz - 1] [Myrmidonów] [Achilles - 1]*  
 NP: *[Achilles szalejący] - [boski Achilles] - [wódz Myrmidonów Achilles],*

zanalizować je moglibyśmy następująco (idziemy od lewej, tj. 1←2):

DE: *[Achilles szalejący] - boski Achilles* - typ AgP-PN  
*[Achilles] szalejący - boski Achilles* - typ PN-PN  
*[Achilles] szalejący - boski Achilles* - typ PN-AgP.

Wybieramy podtyp PN-PN.

Dołączmy teraz do łańcucha koreferencyjnego ogniwo kolejne, tj. frazę (3) *wódz Myrmidonów Achilles*. Z pierwszą frazą moglibyśmy połączyć tę frazę na trzy sposoby:

DE: *[Achilles] szalejący - wódz Myrmidonów Achilles* - typ PN-PN  
*[Achilles] szalejący - wódz-1 Myrmidonów Achilles-1* - typ PN-AgP  
*[Achilles szalejący] - wódz Myrmidonów Achilles* - typ AgP-PN

Również wybieramy możliwość PN-PN. Pamiętajmy, że nie będziemy mogli już znakować koreferencji pomiędzy wyrazem *wódz* a innymi frazami. Ponieważ referent został uwzględniony w nazwie własnej (wybraliśmy z frazy *wódz Myrmidonów Achilles - Achillesa*).



Frazę *wódz Myrmidonów Achilles* z frazą *boski Achilles* (2←3) również możemy złączyć na trzy sposoby:

DE: [*boski Achilles*] - *wódz Myrmidonów Achilles* - typ AgP-PN  
*boski [Achilles]* - *wódz-1 Myrmidonów Achilles-1* - typ PN-AgP  
*boski [Achilles]* - *wódz Myrmidonów Achilles* - typ PN-PN

I znowu z tych trzech możliwości wybieramy trzecią (PN-PN).

Ostatecznie zatem nasz ciąg koreferencyjny będzie wyglądać następująco:

DE: [*Achilles*] *szalejący* - *boski Achilles* - *wódz Myrmidonów Achilles*  
 \_\_\_\_\_| PN-PN |  
 \_\_\_\_\_| PN-PN

### Fraza nominalna nadrzędna a frazy nominalne podrzędne (zagnieżdżone)

Jak już wspomnieliśmy (zasada WEW-NP) nie znakujemy relacji w obrębie jednej frazy nominalnej, np. pomiędzy PN zawierającą głowę nadrzędnej frazy rzeczownikowej a AgP zawierającą głowę podrzędnej frazy nominalnej (gwiazdką w wierszu DE\* oznaczamy niewłaściwe anotacje, właściwe anotacje podajemy w wierszu DE):

DE\*: \*[Piotr Wielki] zwany \*Carem Uwodzicielem  
 DE: Piotr Wielki zwany Carem Uwodzicielem  
 NP: [Piotr Wielki zwany Carem Uwodzicielem]

Ponieważ **fraza szeregową** wchodzi w skład NP tylko wtedy, gdy szereg ma nadrzędnik nominalny, nie możemy łączyć koreferencją frazy nominalnej z szeregiem NP. Tworzymy wtedy dwie relacje ref:ident: [Albert Einstein-1,2] był wielkim fizykiem. Ten wspaniały muzyk-1 i naukowiec-2 miał dwóch synów. Jeżeli fraza szeregową składa się z dwóch nazw własnych połączonych spójnikiem równorzędnym, a nadrzędnik frazy nominalnej jest podany w liczbie mnogiej albo jest rzeczownikiem zbiorowym, to takiej sytuacji nie możemy uwzględnić na poziomie ref:

NP: [Eistein] i [Bohr] pili razem [piwo]. [Ci najslawniejsi fizycy swoich czasów] uwielbiali [dobrą muzykę].  
 CNP: Eistein i Bohr pili razem piwo. Ci najslawniejsi fizycy swoich czasów uwielbiali dobrą muzykę.  
 DE: Eistein i Bohr pili razem piwo. Ci najslawniejsi fizycy swoich czasów uwielbiali dobrą muzykę.

Jeżeli fraza szeregową składa się z dwóch nazw własnych połączonych spójnikiem równorzędnym, a nadrzędnik frazy nominalnej jest podany w liczbie mnogiej albo jest rzeczownikiem zbiorowym, to takiej sytuacji nie możemy uwzględnić na poziomie ref:

NP: [Eistein] i [Bohr] pili razem [piwo]. [Ci najsłynniejsi fizycy swoich czasów] uwielbiali [dobrą muzykę].

CNP: Eistein i Bohr pili razem piwo. Ci najsłynniejsi fizycy swoich czasów uwielbiali dobrą muzykę.

DE: Eistein i Bohr pili razem piwo. Ci najsłynniejsi fizycy swoich czasów uwielbiali dobrą muzykę.

Jeżeli w NP **zagnieżdżona fraza szeregową** oznacza dwa różne byty, wtedy łączymy koreferencją tylko ten składnik nazwy, który jest w stosunku identycznościowym do drugiej frazy nominalnej:

NP.: [Grupa złożona z fizyka i trzech matematyków] poszła na lunch. Albert Einstein był głodny.

CNP: Grupa złożona z [fizyka i trzech] matematyków poszła na lunch. Albert Einstein był głodny.

DE: Grupa złożona z [fizyka] i trzech matematyków poszła na lunch. Albert Einstein był głodny.

**Nie łączymy** głowy danej frazy nominalnej z inną frazą nominalną, jeśli nie jest powiązana z jej (głowy) określeniami (podrzędnikami):

DE\*: [A. E.] był wielkim naukowcem. Fizyk powierzchni to nie to samo, co fizyk teoretyk.

DE: [A. E.] był wielkim naukowcem. Fizyk powierzchni to nie to samo, co fizyk teoretyk

Einstein był co prawda fizykiem, ale nie był fizykiem powierzchni. Nazwa własna *Albert Einstein* łączy się semantycznie z frazą *fizyk*, ale już nie ze złożoną NP [fizyk [powierzchni]]. Zabronione jest zatem wycinanie z fraz nominalnych ich integralnych części (np. podrzędników głowy). Można sobie wyobrazić taką sytuację, że ktoś stara się zmienić referencję frazy rzeczownikowej przez wycięcie jej fragmentu. Taka sytuacja jest niedopuszczalna. Przez integralną część frazy nominalnej rozumiemy to wszystko, co znajduje się w hierarchii składniowej poniżej głowy.

## Apozycje niezgodnione - casus "rzeki Bystrzyca"

Mamy casus, w którym występuje rzeczownik pospolity 'rzeka' oraz nazwa własna 'Bystrzyca' tworzące niezgodnioną apozycję (apozycję, której jeden składnik jest nieodmienny bądź błędnie uznany za nieodmienny). W tekście użycie tej formy jest zdublowane:

[[Zespół]<sub>AgP</sub> [czterech mostów stanowiących]<sub>AgP</sub> [przeprawę]<sub>AgP</sub> [nad rzeką]<sub>AgP</sub> [Bystrzyca]<sub>AgP</sub>]<sub>NP</sub>.  
 (...) [[Mosty]<sub>AgP</sub>]<sub>NP</sub> [[przerzucone]<sub>AgP</sub> -1]<sub>AdjP</sub> [są]<sub>VP</sub> [[nad głównym korytem]<sub>AgP</sub> [rzeki]<sub>AgP</sub> [Bystrzyca]<sub>AgP</sub> -1]<sub>AdjP</sub>.

PN: Zespół czterech mostów stanowiących przeprawę nad rzeką [Bystrzyca]. (...) Mosty przerzucone są nad głównym korytem rzeki [Bystrzyca].

Koreferencja może być w takim przypadku znakowana trojako<sup>7</sup>:

DE?: *nad [rzeką] Bystrzyca - nad głównym korytem rzeki Bystrzyca*

DE?: *nad rzeką [Bystrzyca] - nad głównym korytem rzeki Bystrzyca*

DE?: *nad rzeką [Bystrzyca] - nad głównym korytem rzeki Bystrzyca*

Ponieważ każda z tych relacji odnosi się do jednego referenta (czyli rzeki Bystrzycy) nie możemy oznaczać wszystkich tych połączeń jednocześnie. Wybieramy tylko jedną możliwość (będzie to PN-PN). Precyzuje to poniższa zasada:

### W PRZYPADKU APOZYCJI - JEDNA RELACJA (APP-1)

▣ Ponieważ apozycja nieuzgodniona rozpada się na niezależne AgP, pojawia się pokusa znakowania relacji pomiędzy częściami apozycji a innymi frazami. Ponieważ jednak referent apozycji jest jeden i ten sam, nie ma potrzeby dublowania relacji. ▣

Apozycje stanowiące połączenie rzeczowników:

- o różnych rodzajach, np: *książka Potop; film Piękna i Bestia;*
- o różnych liczbach: *książka Krzyżacy, samochody Nissan,*
- z wyrażeniem przyimkowym: *ulica Na Ostatnim Groszu; lektura Nad Niemnem;*

- mogą nastęrczać podobnych trudności.

## Fraza nominalna a zdanie względne

Zdania względne nie wchodzą w skład fraz nominalnych, nie są też przez nas dołączane do nadrzędnych względem nich fraz NP:

AgP: [Albert Einstein], [który] zawsze grał [na skrzypcach], umarł. [Ten fizyk] był wspaniały

DE\*: [Albert Einstein, który zawsze grał na skrzypcach], umarł. [Ten fizyk] był wspaniały

DE: [Albert Einstein], który zawsze grał na skrzypcach, umarł. Ten fizyk był wspaniały

Ponieważ łączymy ze sobą wyłącznie frazy nominalne bez ich zdań podrzędnych, nie możemy patrzeć na referencję połączenia frazy rzeczownikowej i jej zdania podrzędnego, lecz wyłącznie na „czyste” frazy nominalne.

- nie znakujemy relacji pomiędzy zaimkami względnymi a ich odpowiednikami
- możemy znakować relacje koreferencji pomiędzy częściami składniowymi zdań składowych.

Rodion zabił [Lizawietę] siekierą, na którą spływała później jej krew.

<sup>7</sup> Relacja *Bystrzyca - rzeka* w obrębie fraz NP:

DE\*: *Zespół czterech mostów stanowiących przeprawę nad [rzeką] Bystrzyca*

DE\*: *Mosty przerzucone są nad głównym korytem [rzeki] Bystrzyca*

- nie może być znakowana ze względu na zasadę WEW-NP.

## Koreferencja a metonimia

Nie łączymy ze sobą elementów zbiorów i samych zbiorów. Szczególnym przypadkiem takiej reguły jest zakaz łączenia ze sobą nazw elementów z nazwą całości

DE\*: Grupa złożona z [Einsteina] i kolegów poszła na spacer. Fizycy byli szczęśliwi.

DE: Grupa złożona z Einsteina i kolegów poszła na spacer. Fizycy byli szczęśliwi.

W tym przypadku nie mamy możliwości połączenia nazwy własnej z frazą nominalną (relacja część – całość, której nie znakujemy).

## Anafora typu konotacyjnego (sense)

Nie znakujemy anafory typu konotacyjnego (identyczność znaczenia konotacyjnego przy nietożsamości referentów):

DE\*: Piotr ma fajną dziewiętnastoletnią [Olę], ale \*dziewiętnastka Pawła też jest niezła.

DE: Piotr ma fajną dziewiętnastoletnią Olę, ale dziewiętnastka Pawła też jest niezła.

NP: [Piotr] ma [fajną dziewiętnastoletnią Olę], ale [dziewiętnastka Pawła] też jest niezła.

## Koreferencja a sytuacje (zdarzenia)

Nie łączymy rzeczownikowych określeń sytuacji (zdarzeń) z ich odpowiednikami wyrażonymi frazami werbalnymi, ponieważ nie znakujemy anafory typu *deixis*. Np. nie połączymy gerundium z jego odpowiednikiem tekstowym wyrażonym frazą czasownikową:

DE\*: Prezydent \*[przemawiał długo i zawile wykladał] przyczyny zaproszenia Wojciecha Jaruzelskiego na spotkanie prezydentów. \*Przemówienie Komorowskiego nie znalazło uznania w społeczeństwie.

Oznaczmy za to nazwy sytuacji (zdarzeń), jeżeli będą wyrażone rzeczownikami i jeśli jedna z fraz będzie wyrażona nazwą własną:

DE: [3 Maja] to wspaniałe wydarzenie. Najlepszy moment w historii Polski trwał chwilę.

## Użycia predykatywne

Nie znakujemy koreferencji typu AgP-PN wtedy, gdy mamy do czynienia z użyciem frazy nominalnej w funkcji predykatywnej, której najbardziej typowym przykładem jest funkcja orzecznika w orzeczeniu

złożonym.

Orzeczenie złożone wyrażane jest trzema schematami składniowymi:

1. X (nom.) jest Y-em (inst.)
2. X (nom.) to Y (nom.)
3. X (nom.) to jest Y (nom.)

Jeżeli Y jest nazwą własną to nie może pojawić się w schemacie 1. Pojawienie się nazwy własnej w schemacie 1. wskazuje na jej użycie predykatowe, w związku z czym dochodzi do jej deprioprializacji. często genetycznie nazwa własna (a w tekście już użyta predykatywnie) pisana jest małymi literami.

Np. jednostka leksykalna *Einstein* będącą nazwą własną nie może być użyta w funkcji orzecznika, chyba że w szczególnym przypadku, gdy zaczyna pełnić funkcję predykatywną:

(&) *Piotr (nie) jest Einsteinem.*

Użycia fraz nominalnych pełniących zazwyczaj funkcję nazw własnych w roli orzecznika wykracza jednak poza funkcję nazwy własnej. Użyty w taki sposób wyraz przestaje być nazwą własną i staje się wyrazem pospolitym:

„Zgodnie z tym kryterium [Russella], NW [nazwy własne] Nixon, Tomasz z Akwinu, Sokrates, użyte w funkcji orzecznika pełnionej w języku polskim przez nazwy w narzędniku w zdaniach takich jak

*Nie jest pan Nixonem*

*Nie jestem Tomaszem z Akwinu* (parafraza słów Williama z Baskerville, bohatera powieści „Imię Róży” U. Eco)

*Aby być Sokratesem, musiałby pan wypić cykute*

nie są logicznymi NW” (Karolak 1995/2001, 298-9).

To zjawisko w językoznawstwie polskim znane jest bardzo dobrze. Doczekało się nawet dwóch reguł ortograficznych (www.pwn.pl, Słownik ortograficzny języka polskiego: Zasady pisowni i interpunkcji:

„20.22. Rzeczowniki utworzone od imion własnych, używane jako nazwy pospolite:

bajronista, garybaldczyk, heglista, kościuszkowiec, marksista, piłsudczyk, stachanowiec.

UWAGA: Zaliczamy tu także nazwiska osób znanych z życia publicznego, używane w liczbie mnogiej (np. Nieporadne rządy gomułków i bierutów) w celu poniżenia, okazania lekceważenia bądź pogardy wobec postaw, idei czy też osób związanych z nosicielem danego nazwiska.”

20.23. Rzeczowniki utworzone od nazw własnych ludzi oraz istot mitologicznych, używane w znaczeniu pospolitym:

hamlet (= człowiek niezdecydowany), kozak (= taniec), krezus (= bogacz), ksantypa (= kobieta kłótniwa, zrzędna), łazarz (= człowiek chory, opuszczony), szwajcar (= odźwierny).”

W świetle tych reguł zdanie (&) moglibyśmy zapisać tak:

(&') *Piotr (nie) jest einsteinem.*

Przyjmujemy w niniejszym opracowaniu zasad anotacji korpusu odniesieniami anaforycznymi, że frazy

nominalne, których użyć można w funkcji orzecznika w zdaniach (atomowych), nie są nazwami własnymi.

Nazwa własna może za to pojawić się w schemacie 2. i 3. Funkcję tych schematów składniowych moglibyśmy określić mianem *identyfikującej*. Nazwa własna na pozycji Y w przeciwieństwie do schematu 1. nie traci swojego statusu.

Użycie w schemacie 1. w funkcji orzecznika rzeczownika pospolitego lub deskrypcji określonej ma charakter predykatywny. Tego typu użyć NIE ZNAKUJEMY.

Przykłady:

**Schemat: 1. X (nom.) jest Y-em (inst.)**

AgP: [Piotr Wielki] był [carem] [Rosji].

NP: [Piotr Wielki] był [carem Rosji].

PN: [Piotr Wielki] był carem [Rosji].

DE: Piotr Wielki był carem Rosji.

AgP: [Piotr Wielki] był [carem].

NP: [Piotr Wielki] był [carem].

PN: [Piotr Wielki] był carem.

DE: Piotr Wielki był carem.

Użycie w funkcji orzecznika w obrębie schematu 2. i 3. rzeczownika pospolitego lub deskrypcji określonej ma również charakter predykatywny.

AgP: [Piotr Wielki] to (jest) [car].

NP: [Piotr Wielki] to (jest) [car].

PN: [Piotr Wielki] to (jest) car.

DE: Piotr Wielki to (jest) car.

AgP: [Piotr Wielki] to (jest) [car] [Rosji].

NP: [Piotr Wielki] to (jest) [car Rosji].

PN: [Piotr Wielki] to (jest) car [Rosji].

DE: Piotr Wielki to (jest) car Rosji.

Podsumujmy to w tabelce.

schemat \ Y =	nazwa własna	rzeczownik pospolity	deskrypcja określona
---------------	--------------	----------------------	----------------------

X jest Y-em	P	P	P
X to Y	I	P	P
X to jest Y	I	P	P

Legenda: P - użycie predykatywne; I - użycie identyfikujące

A zatem spośród tych trzech schematów jedynie dwa ostatnie mogą być znakowane na poziomie koreferencji (relacja łącząca X i Y) i to tylko w przypadku gdy orzecznik jest nazwą własną.

Wśród innych schematów predykatywnych wymienić można:

**X zwany Y, X zwie się Y, X nazywa się Y, X nazywany Y, X mieni się Y, X o nazwie Y, X o mianie Y, X noszący miano Y;**

TAKICH UŻYĆ RÓWNIEŻ NIE ZNAKUJEMY JAKO KOREFERENCJI MIĘDZY X a Y.

Nie łączymy użytych w funkcji predykatywnej głów fraz nominalnych z innymi frazami. Np.

DE\*: *[Projekt ten] spalił na panewce. PZL-13 ochrzczono mianem projektu nieudanego.*

DE: *[Projekt ten] spalił na panewce. PZL-13 ochrzczono mianem projektu nieudanego.*

Elementy frazy NP czy AdjP niebędące głowami mogą za to wchodzić w koreferencję z innymi frazami:

NP: *[Mariusz Kamiński - polarnik i zdobywca Bieguna Północnego]. Jak wspomina: [Zdobycie pierwszego bieguna] było najłatwiejsze. Potem poszło już [z **górk**].*

PN: *[Mariusz Kamiński] - polarnik i zdobywca [Bieguna Północnego]. Jak wspomina: Zdobycie pierwszego bieguna było najłatwiejsze. Potem poszło już z **górk**.*

DE: *Mariusz Kamiński - polarnik i zdobywca [Bieguna Północnego]. Jak wspomina: Zdobycie pierwszego bieguna było najłatwiejsze. Potem poszło już z **górk**.*

Ukryta predykacja

Zdarzają się sytuacje, gdy użycie przedykatowe nie przyjmuje znanych nam ze schematów form. Przykładem takiej sytuacji niech będzie:

DE\*: [Szczepionka pre-pandemiczna] jest niebezpieczna dla zdrowia. Na rynku farmaceutycznym występuje pod nazwą Killhimix i Killherix.

DE: [Szczepionka pre-pandemiczna] jest niebezpieczna dla zdrowia. Na rynku farmaceutycznym występuje pod nazwą Killhimix i Killherix.

DE: [Lidia], którą niektórzy nazywali Żelazną Damą, jeździła kabrioletem.  
W okresie emigracyjnym pełniła funkcję członka Centralnej Rady PPS.

DE\*: [Lidia] była autorką programu marketingu szeptanego w znanej firmie z branży makulaturowej. Swego czasu lider sprzedaży tekturowych krasnali ogrodowych.

DE: Lidia była autorką programu marketingu szeptanego w znanej firmie z branży makulaturowej. Swego czasu lider sprzedaży tekturowych krasnali ogrodowych.

W przykładzie tym nie zaznaczamy żadnych anafor, ponieważ druga fraza jest równoważnikiem zdania, ma formę eliptyczną.

Myślnik jako element predykatopodobny:

W kompleksie znajduje się jedna z najważniejszych galerii sztuki Inuitów - Toronto Art Gallery.

Można by ten myślnik zinterpretować jako “czyli”. Interpretujemy taki przypadek jako podtyp związany z identyfikacją referentów:

DE: W kompleksie znajduje się [jedna] z najważniejszych galerii sztuki Inuitów - Toronto Art Gallery.

## Referencja jako podstawa znakowania

Zdarzają się sytuacje, w których pewna nazwa własna odnosząca się do rzeczywistości pozajęzykowej jest zamieniana w tekście inną nazwą własną. Czasami różnica może być na tyle duża, że możemy mieć wątpliwości w kwestii tożsamości referenta - czy faktycznie w obu przypadkach jest ten sam? Wątpliwości takie mogą pojawiać się szczególnie w kontekście bytów istniejących długi czas, które kilkakrotnie zmieniały nazwy (np. nazwy instytucji, oddziałów wojskowych itp.). Nazwy takie łączymy ze sobą w przypadkach pewności, co do tożsamości referenta, w wypadkach wątpliwych - nie znakujemy.

DE: [11 Pułk Piechoty Liniowej Królestwa Polskiego] został sformowany rozkazem dyktatora gen. Józefa Chłopickiego z dnia 10 stycznia 1831. 1 Pułk Województwa Sandomierskiego otrzymał w ciągu wojny 7 krzyży złotych i 3 srebrne.

W tym przypadku z dalszego brzmienia artykułu można było wywnioskować, że 11 Pułk Piechoty Liniowej Królestwa Polskiego oraz 1 Pułk Województwa Sandomierskiego to na pewno nazwy tego samego bytu.



## Lista predykatów, które uznajemy za ośrodki predykcji

pełnić funkcję + gen.  
 pracować jako + nom.  
 uważać za + Acc.

## Użycia dzierżawcze

Formy *jego*, *jej*, *ich* użyte w funkcji przydawki (z *jego fuzją*, z *jej dwulufką*, z *ich bronią krótką*), znakujemy (typ NP-Pron):

DE: [Piotr-1] miał fuzję, a [Ania-2] dwulufkę. Jego-1 broń wystrzeliła, a jej-2 - nie.

Por:

[On] miał fuzję, a [ona] dwulufkę. Piotra-1 broń wystrzeliła, a Ani-2 - nie.

Analogicznie znakujemy użycia dzierżawcze (dopełniacz dzierżawczy) od rzeczowników pospolitych i nazw własnych:

DE: Płk [Piotr-1] kupił kapelusz. Kapelusz pułkownika-2 był drogi.

Nie znakujemy relacji pomiędzy przymiotnikiem dzierżawczym a nazwą własną:

DE\*: Płk [Piotr] kupił kapelusz. Kapelusz pułkownikowy był drogi.

DE: Płk Piotr kupił kapelusz. Kapelusz pułkownikowy był drogi.

## Znakowanie podmiotów domyślnych

Na potrzeby uczenia maszynowego konieczne jest oznaczenie wszystkich czasowników z podmiotem domyślnym, a więc nie tylko tych, których subiekt przybiera formę nazwy własnej (jednostki identyfikacyjnej) użytej w innym miejscu tekstu. Te bowiem powinny zostać dodatkowo oznaczone jako “wyznacznik”. Wszystkie czasowniki z podmiotem domyślnym (także te będące wyznacznikami) oznaczamy według następujących zasad.

### Założenia

- Znakujemy czasowniki zarówno z subiektem konotowanym końcówką czasownika (a), jak i z subiektem domyślnym *sensu stricto* (b), a więc takim, który został w pełni oznaczony w innym miejscu tekstu, np.
  - subiekt konotowany:
    - *Idę do domu,*
    - *Skąd to **wzięła**|ś?*
  - subiekt domyślny:
    - *<Zmiana jest spowodowana wystąpieniem senator Hillary Clinton, która postanowiła walczyć z producentem gry, Rockstar Games.> **Oskarżyła** ich o celowe pozostawienie w kodzie gry scen erotycznych, które mogą być odblokowane za pomocą specjalnej modyfikacji o nazwie Hot Coffee Mod. (oskarżyła odnosi się do wspomnianej wcześniej Hillary Clinton)*
- Nie znakujemy form blokujących podmiot zgodnie z poniższą klasyfikacją:

## funkcjonalna klasyfikacja orzeczeń

orzeczenia				
nieblokujące podmiotu	blokujące podmiot			
osobowe formy czasownika	leksykalnie		fleksyjnie	poprzez nieprototypowe użycie kategorii osoby
	czasowniki niewłaściwe	predykatywy	bezosobniki	formy osobowe użyte niesobowo
<i>Jem loda. Ty jesz loda. Wiesław je loda.</i>	<i>Mdliło mnie mocno. Już świta.</i>	<i>Nie sposób go opanować. Pora spać. Nie trzeba było go budzić.</i>	<i>Wybito szybę. Wyłano mleko.</i>	<i>Należałoby teraz wyjść. Wczoraj było tak cicho. Zrobiło mi się niedobrze. Ciagle się o tym mówi. Dobrze mi się jeździ</i>

Poprzez nieprototypowe użycie kategorii osoby rozumiemy takie użycia form osobowych, które konwencjonalnie sygnalizują odbiorcy, że w rzeczywistości nie ma subiektu. W języku polskim

są to najczęściej formy 3 osoby (rodzaju nijakiego), np. *pachniało miodem i żywicą, wczoraj było tak cicho* oraz formy ze słowem *się* w funkcji wykładnika bezosobowości, np. *dobrze mi się jeździ, tak się nie robi*.

3. Jeśli podmiot został wyrażony zaimkiem względnym (który, co, jaki...), to wprowadzamy oznaczenie "verb\_null\_in".

## Uwagi techniczne

1. Znakujemy tylko głowy czasowników. Głowę stanowi:
  - dla orzeczenia prostego: token zawierający temat czasownika;
    - > w przypadku form aglutynacyjnych w zakres anotacji nie wchodzi ruchome końcówki czasownika i flektywy trybu przypuszczającego;
    - > w przypadku form z nieciągłym morfemem leksykalnym (np. *bać się*) w zakres anotacji wchodzi tylko segment podstawowy;
    - > w przypadku czasu przyszłego złożonego w zakres anotacji wchodzi tylko słowo posiłkowe "być"

Przykłady:

- *śpiewała*
- *śpiewała|ś*
- *śpiewała|byś*
- *bała|ś się*

- dla orzeczenia złożonego:
  - (słowno-)jimiennego oraz innych orzeczeń z elementem niewerbalnym (frazologizmów, orzeczeń zaprzeczonych) → tylko część werbalną znakowana zgodnie z zasadami opisanymi w podpunkcie a)

Przykłady:

- *został strażakiem*
- *został|by postrzelony*
- *dała ognia*
- *darli|ście koty*

- składającego się z czasownika modalnego lub fazowego + bezokolicznik → tylko czasownik w formie osobowej, a więc modalny bądź fazowy znakowany zgodnie z zasadami opisanymi w podpunkcie a)

Przykłady:

- *może pomieścić*
- *mógł|by pomieścić*
- *zaczęła uciekać*
- *zaczęli|ście uciekać*

2. W pierwszej iteracji interesuje nas automatyczny podział na sentencje i oznaczamy czasowniki, które nie mają podmiotu oznaczonego (wyrażonego) w danej sentencji (w przypadku wypowiedzeń złożonych, ten podział nie pokrywa się z podziałem na zdania składowe).

Kategoria anotacji: `wyznacznik_verb_null`

Przykładowo we fragmencie:

- *Janek wrócił ze sklepu. **Zapomniał** jednak zakupów.*

oznaczylibyśmy czasownik "zapomniał" jako "verb\_null", ale gdyby ten fragment wyglądał tak:

- *Janek wrócił ze sklepu, ale zapomniał zakupów,*

to czasownika "zapomniał" już nie oznaczalibyśmy (mimo że *de facto* jego podmiot nie został wyrażony w tym samym zdaniu), dlatego że ma podmiot oznaczony w zdaniu nadrzędnym (a więc w tej samej sentencji).

3. W drugiej iteracji oznaczamy czasowniki, które mają podmiot oznaczony w danej sentencji, ale nie w danym zdaniu składowym (przykład b).

Kategoria anotacji: `wyznacznik_verb_null_in`

## Analiza przypadków szczególnych

### 1. *Prosić*

Czasownik *prosić* może być znakowany. Istnieje jednak jedno ze znaczeń tego czasownika trudne do sprecyzowania, używane w zwrotach grzecznościowych typu: *proszę państwa, proszę cioci*. W tym znaczeniu czasownik *prosić* nie będzie znakowany. W pozostałych jednak przypadkach, np. w zdaniach: *proszę, zrób mi herbaty; proszę zachować bezpieczną odległość, proś gości do salonu, czy mogę prosić dziekana do telefonu, czy mogę prosić Panią do tańca...* czasownik *prosić* (jeśli ma subiekt konotowany lub domyślny *sensu stricto*) powinniśmy znakować.