Samsung R&D Poland
Language Support Part

# PolEval – Task 2 Named Entity Recognition

## *Ustalenia i strategie anotacyjne*

na podstawie:

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski i Barbara Lewandowska-Tomaszczyk (red.) **Narodowy Korpus Języka Polskiego**. Wydawnictwo Naukowe PWN, Warszawa.

**Zespół anotatorów:**
Tunia Błachno
Oliwia Ebebenge
Monika Jędras
Estera Małek
Justyna Polak
Justyna Rafalska
Grzegorz Woźniak

## OGÓLNE

---

**1. NERy w nazwach produktów** anotujemy, o ile nie są tożsame z nazwą produktu
za instrukcją: "Nowa **Grochola** jest bardzo fajna" – jako książka, a nie osoba/person, wiec nie anotujemy

[**Microsoft**] $^{orgname}$ Access to jedna z najpopularniejszych baz danych.
- zagnieżdżamy nazwę firmy w nazwie produktu
Wpisałam dane do **Microsofta**.
– bez anotacji, bo wpisałam dane do bazy danych (firmy) Microsoft

Książka pt. „Życie[**Jana Kowalskiego**]$^{persName}$„
– zagnieżdżamy Jana Kowalskiego w nazwie produktu (książki)
Książka pt. „**Jan Kowalski**”
– bez anotacji, Jan Kowalski nazywa książkę, a nie osobę

Wypiłem herbatę [**Liptona**]$^{org}$.
- wypiłem herbatę firmy Lipton
Wypiłem **Liptona**.
- bez anotacji, ponieważ Lipton nazywa napój, a nie organizację (firmę)
ALE UWAGA! Wypiłem kawę **Nescafe.**
- bez anotacji, ponieważ kawa jest <u>marki</u> Nescafe (nie ma takiej firmy/takiego ORGa jak „Nescafe”)

<u>Inne metonimie:</u>
Przelot **Air India** do Pragi byłby najprostszy.
- bez anotacji ORG, ponieważ nie lecę ORG-iem, ale liniami lotniczymi firmy ORG

Buty **[Gino Rossi]**$^{org}$
- bez anotacji persName

---

**2. EUROPA:**
Za instrukcją: „Jedną z najbardziej skomplikowanych kwestii była anotacja jednostek nazewniczych związanych z Europą i Unią Europejska. Przyczyną tej trudności były użycia metonimiczne leksemu Europa oraz przymiotnika europejski. Jednostki te niezmiernie rzadko oznaczały kontynent, jak w przykładzie (9.39). Najczęściej odnosiły się one do mieszkańców lub obywateli Europy jako kontynentu w sensie geograficznym (9.40) lub jako organizacji. W tym ostatnim przypadku nazwa Europa była często synonimem Unii Europejskiej, interpretowanej jako blok państw lub jako instytucja.”

geog – największe miasto Europy
org - sprawczość - Europa nie przejmuje się tym
bloc - jeśli jasno widać, że chodzi o UE, a równocześnie nie ma sprawczości (jeśli jest sprawczość – ORG) - np. państwo w tym roku dołączyło do Europy

**NATO, UNIA EUROPEJSKA, UE:**
bloc - członkostwo (wejść do UE, być członkiem UE), na terenie UE, w UE
org - sprawczość
**EUROPEJSK.*:**
relAdj_geog - od kontynentu
relAdj_bloc - od Unii- czyli w samej nazwie: [Unia [europejska]$^{bloc}$] $^{org/bloc}$ + przepisy europejskie (nie Europy a Unii)
**UNIJN.*:**
relAdj_bloc - od Unii
**KOMISJA EUROPEJSKA**: [Komisja [Europejska] $^{bloc, relAdj}$] $^{org}$
**PARLAMENT EUROPEJSKI**: [Parlament [Europejski] $^{bloc, relAdj}$]$^{org}$
**WSPÓLNOTA EUROPEJSA:** [Wspólnota [Europejska] $^{bloc, relAdj}$]$^{org/bloc}$

---

**3. Województwa, przymiotniki od nazw województw**
Za instrukcją: „otwarta kwestia pozostaje jednak wówczas, jaka baza derywacyjna jest najbardziej uzasadniona w nazwie samego województwa – zob. (9.78).”

> [województwo [lubelskie]$^{settlement, relAdj}$]$^{region}$
> [województwo [pomorskie]$^{geog, relAdj}$]$^{region}$
> w [lubelskim] - to jest elipsa od 'województwie lubelskim', więc tylko [region], bez [settlement, relAdj]

w instytucjach typu “Lubelski Oddział Krwiodawstwa [Lubelski] region+relAdj albo settlement+relAdj, trzeba sprawdzać czy w instytucji chodzi o miasto czy o całe województwo

> [pomorska] firma - raczej geog + relAdj

---

**4. Anotujemy nazwy w innych językach, dopóki powszechne jest jej zrozumienie**
(za instrukcją, str. 154)

> Reductio ad **Hitlerum**
> festiwal **Wratislavia** Cantans
> Król **Lir** – spolszczone, ale anotujemy (tak jak **Szekspir** oraz **Shakespeare**)

---

**5.** Za instrukcją: **NIE anotujmy wariantów stylistycznych nazw własnych**:

> Zahaczył się na dalsze występy w akademickiej drużynie za **Oceanem** (~ Ameryka)
> Wśród **"Jeziorowców"** wyróżnił się Kobe Brayn (~Los Angeles Lakers)
> **Sympatyczna panna S.** (~ Solidarność)

ALE! Cienka granica między derywatem, a wariantem stylistycznym:
„[**stonsi]**$^{persDeriv\_ORG}$ zażyczyli sobie zapłatę za koncert” – anotujemy, ponieważ nazwa stonsi jest spolszczonym derywatem dla nazwy członka zespołu Rolling Stones; tym samym „**Lakersi”** też byliby anotowani

---

**6. NIE anotujemy anafor**, nawet jeśli odnoszą się jednoznacznie do znanego nam ORGa
[Gmina [Kościelisko]$^{settlement}$]$^{org}$ dostała pismo. Po dwóch dniach Gmina odpowiedziała na nie.

---

**7.** W sytuacjach trudnych do roztrzygnięcia najpierw odwołujemy się do wiedzy (**robimy niezbędny research**), a jeśli to nie daje odpowiedzi, **zdajemy się na kontekst oraz intuicję**.

Np. "pędzi wzdłuż **ulicy głuchej"**
*Research:* znalezienie źródła (książka „Ozimina"); poznanie miejsca akcji książki (Warszawa); potwierdzenie istnienia ulicy o nazwie „Głucha” w Warszawie
*Obserwacje:* dominujący w książce szyk przestawny, mało prawdopodobne, żeby w książce „ulica głucha” jako NER była użyta małą literą;
*Ostateczna decyzja:* brak anotacji, ulica głucha jako ulica pusta, cicha

## PERSON

---

**1. Pan Bóg**, **Pan Jezus, Jezus Chrystus, Chrystus, Matka Boska, NMP, (Panna) Maryja**
całość jako persName (za instrukcją: „Do grupy tej, oznaczanej przez persName, należą nazwy indywidualnych osób i rodzin (…) fikcyjnych, legendarnych i pochodzących z wierzeń religijnych”)

Matka Boska - [Matka [Boska]$^{persName,\ relAdj}$]$^{persName}$

Matka Boska Fatimska - [Matka [Boska]$^{persName,\ relAdj}$[Fatimska]$^{settlement,\ relAdj}$]$^{persName}$

Anotujemy również w wykrzyknieniach: O Boże! Jezu Chryste!

**Szatan** – traktujemy jak wszystkie w/w nazwy postaci (chyba że użycie metaforyczne); w korpusie milionowym brak konsekwencji

Ksiądz każe się Jadźce cieszyć z wizyty **szatana**. To znaczy, że **szatan** jest w desperacji, przedstawił się, bo czuje, że ją traci. – ANOTUJEMY

Wtedy Jadźka pierwszy raz zobaczyła swojego **szatana** – NIE ANOTUJEMY

---

**2. Nazwy pospolite funkcjonujące jako nazwy własne**
Poetka, Nieznajomy, Pan Smutny – addname; w każdym przypadku decyduje kontekst

---

**3. Hasztagi**
#PosełAleksanderKwaśniewski
#Poseł[[Aleksander]$^{forename}$[Kwaśniewski]$^{surname}$]$^{persName}$
z oficjalnych hasztagów wyciągamy też ORGi (np. #MinisterstwoSportu)

---

**4. Nazwiska w nazwach chorób –** anotujemy

Choroba [Alzhaimera]$^{persName,\ persSurname}$

ALE! "dzieci z zespołem Downa" (anotujemy persName) vs "dzieci z Downem" (nie anotujemy – metonimia)

---

**5. Pseudonimy oraz imiona i nazwiska superbohaterów oraz postaci fikcyjnych** traktujemy tak jak imiona, nazwiska i ksywki postaci prawdziwych:

> Władający biegle kilkoma językami, fizycznie sprawny niczym [[Superman]$^{persName\_addname}$]$^{pe}$ i przystojniejszy od [[Jamesa]$^{forename}$[Bonda]$^{surnam}$]$^{persName}$.

Imiona i pseudonimy bohaterów zwierzęcych – anotujemy jako persName, jeśli są wyraźnie spersonifikowani.

(za instrukcją: nie anotujemy imion, nazw zwierząt nie posiadających cech ludzkich):

> Nie zabraknie jednak także starych przyjaciół, w tym przede wszystkim [[Osła]$^{persName\_addName}$]$^{persNa}$

---

**6. Nazwiska wieloczłonowe – rozdzielamy**

str. 143, przykład 9.20 sugeruje rozłączną anotację nazwisk złożonych:

> [[Janusz]$^{persName\_forename}$ [Korwin]$^{persName\_surname}$ - [Mikke]$^{persName\_surname}$]]$^{persN}$

(chociaż... Wołk-Karczewska – niekonsekwentna anotacja)

---

**7. Nicki na forach** - tylko addName i persName:

"magda: ...." "magda17 napisała:...." - nie wyciągamy stąd "Magdy" jako forename

podobnie "gosc_z_Myslowic" - nie wyciągamy Mysłowic

---

**8. Zdrobnienia imion:**

Za instrukcją: „forename – imię, ewentualnie złożone, zdrobniałe lub w liczbie mnogiej”.

> poprawne: forename (np. Ewka)
>
> odbiegające od normy: addname (np. Dżoni)

---

**9. Części składowe imion, nazwisk, psedonimów niefunkcjonujące niezależnie**

Za instrukcją: „Części składowe niemogące funkcjonować niezależnie jako nazwy własne, np. *van der, Junior* itp., nie maja odrębnego podtypu (…)”

> Jan Paweł II: [[Jan]$^{forename}$ [Paweł]$^{forename}$ II]$^{persName}$

Arabskie nazwiska - '**Al-**' jest włączane do nazwiska, nie działa jak 'von' czy 'der'

> Kasim Al-Sabti: [[Kasim]$^{forename}$ [Al-Sabti]$^{surname}$ ]$^{persName}$

# ORG

---

**1. PLACE_NAME a ORG**

**kraj/miasto/blok** anotujemy jako org tylko jeśli jest on/ono sprawcą czynności, nie nadużywamy tego - czyli jesli nie ma sprawczości w zdaniu to jest country, settlement lub bloc

SPRAWCZOŚĆ: Polska zabroniła, Warszawa nie przyjęła, UE ukarała (ORG)
BRAK SPRAWCZOŚCI: prezydent Polski, na terenie UE, wstąpić do NATO (COUNTRY, BLOC)

---

**2. GEOG a ORG**

ORG jeśli jest sprawczość, GEOG może być zaś wszystkim tym, co można wpisać w google maps

"nagrywała tę płytę 2 godziny w studio **MEGA**" - GEOG (lokalizacja)

"**Polska** nie może sobie w żadnym wypadku pozwolić na rozluźnienie polityki finansowej" – ORG (organizacja, użycie metonimiczne)

**TEATRY** -test "taksówki" - czyli GEOG zawsze w kontekście "taksówka podjechała pod...", ale również w kontekście "Wystawiono premierę "sztuki X" w teatrze X" („kilka razy w miesiącu gram w Teatrze Scena Prezentacje"). ORG tradycyjnie w bardziej sprawczym kontekście: "Teatr Mały natrafił na złotą żyłę (…)" )

---

**3. Organizacje państwowe**

**Sejm, Senat, Parlament** – anotujemy jako orgName jeśli zachodzi sprawczość („Podwyżki zostaną wypłacone dzięki ustawie, którą uchwalił Sejm")

**Rząd** (wariant stylistyczny Rady Ministrów w kontekście polskiej polityki), Episkopat (grupa ludzi, nie będąca organizacją) – nie anotujemy

**Wysoka Izba** – nie anotowana. Uznana za wariant stylistyczny nazwy własnej (za instrukcją pkt. 9 str. 137)

---

**4. Organizacja + skrót jej nazwy**

"[Międzynarodowa Organizacja Zdrowia]$^{orgName}$ [WHO] $^{orgName}$"

---

**5. ORG w ORG**

**np. Wydział Lekarski Akademii Medycznej w Krakowie**

[[Wydział Lekarski]$^{orgName}$[Akademii Medycznej w [Krakowie]$^{placeName.settlement}$ ] $^{orgName}$ ]$^{orgName}$

zagnieżdżony ORG może być oddzielnym ORGiem nawet jeśli on sam nie jest czystym NERem, który odnosi się do konkretnej organizacji (np. "Wydział Lekarski" - jest wiele wydziałów lekarskich w Polsce), ale zaznaczamy go wtedy, **kiedy ten ORG jest częścią większego ORG** (czyli Wydzial Lekarski jest częścia Akademii Medycznej).
Tym samym, nie zaznaczamy "Rady Gminy" w "Rady Gminy w Kęsowie", bo "Rada Gminy" nie jest odrębną częścią "Rady Gminy w Kęsowie".

---

**6.** anotujemy **ORGI, nawet jeśli nie są w swojej pełnej nazwie**, np.

**„**Stanął on wówczas przed [Sądem Okręgowym]$^{orgName}$” - jest wiele Sądów Okręgowych i nie wiadomo o który chodzi, ale wiemy, że chodzi o jeden konkretny sąd.

"[Powiatowy Urząd Pracy] $^{orgName}$ " - nie wiadomo w jakim powiacie jest ten urząd, ale jest wielka litera w nazwie instytucji

---

**7. Policja** (poza przypadkami, gdzie ewidentnie występuje jako formalna organizacja, oficjalny organ państwowy), Straż Pożarna, Straż Miejska - NIE ANOTUJEMY

**Straż Graniczna , Ochotnicza Straż Pożarna, Państwowa Straż Pożarna –** ANOTUJEMY

**Sanepid** – wielką literą ANOTUJEMY, małą literą NIE ANOTUJEMY (za oznaczeniami w milionowym korpusie)

---

**8. III RP** - tylko całość country lub org (kontekst),
**IV RP** – wyciągamy tylko RP (country lub org)

**PRL** - traktujemy jak Polska - czyli pierwsze znaczenie COUNTRY, chyba że jest sprawczość, anotujemy jako ORG (w stosunkach PRL (orgName) – RFN)
**peerelowski** – relAdj od głównego znaczenia czyli country

---

**9. Kluby Sportowe**

[[Mostostal]$^{org}$ [Kędzierzyn-Koźle]$^{settlement}$]$^{org}$

Za instrukcją: z uwagi na trudności interpretacyjne nie zostały określone kryteria pozwalające wybrać między strategią anotacji całościowej a zagnieżdżeniem nazwy geograficznej w tego typu jednostkach nazewniczych.

---

**10.** Nazwy programów, projektów, produktów NIE SĄ nazwami organizacji, w związku z tym NIE anotujemy:

Na **Onecie** czytałem
**Fundusz Ubezpieczeń Społecznych**
znaczący krok w realizacji ogłoszonych w ubiegłym roku planów **Ambicja 2012**

## GEOG

---

1. **wschodnie/północny/centralna** + GEOG/PLACE

anotujemy sam GEOG/PLACE, chyba że cała nazwa zwyczajowo funkcjonuje jako całość
*w całości: Europa Wschodnia, Irlandia Północna
*tylko geog: centralna Polska, wschodnie Podlasie

---

**2. Sale, aule itp.**

Według instrukcji kategoria geogName obejmuje nazwy obiektów geograficznych o cechach fizycznych wyróżniających je w terenie (…) nie podlegają podziałowi na podtypy. Stąd decyzja o nieanotowaniu części składowych budynków, jak sale, aule itd.

Odbyła się 30 stycznia br. w sali **Orchidea** i trwała do białego rana. / bez anotacji

---

**3. GEOGI zagnieżdżone**

zagnieżdżamy przymiotniki od-GEOGowe: [Ostrów [Mazowiecka] $^{\text{relAdj\_geog}}$] $^{\text{settlement}}$

NIE zagnieżdżamy rzeczowników: [Korea Południowa]$^{\text{country}}$, [Górny Śląsk]$^{\text{geog}}$ , [Irlandia Północna]$^{\text{country}}$

*wyjątek stanowią nazwy państw, gdzie zagnieżdżamy zwyczajową nazwę państwa w oficjalnej nazwie państwa: [Królestwo [Hiszpanii]$^{\text{country}}$]$^{\text{country}}$

---

**4. budynek, miejsce – GEOG:**

samochód podjechał pod Komendę Miejską w Radomiu

Pan X otworzył w 2000 roku resturację "Mamma Mia"

**ORGi tylko w uzasadnionych wypadkach (sprawczość):**

Komenda Miejska w Radomiu wydała nakaz aresztowania

"Mamma Mia" zatrudnia 5 pracowników.

---

**5. Oficjalne nazwy państw** – mimo że zwyczajowa i oficjalna nazwa jednostki terytorialnej (kraju) będą zawsze tożsame znaczeniowo (Królestwo Hiszpanii=Hiszpania), zdecydowaliśmy konsekwentnie wyciągać z nich:

a) nazwy zwyczajowe: [Królestwo [Hiszpanii]$^{\text{country}}$]$^{\text{country}}$

b) przymiotniki: [Federacja [Rosyjska]$^{\text{relAdj\_country}}$]$^{\text{count}}$, [Rzeczpospolita [Polska]$^{\text{relAdj\_country}}$]$^{\text{country}}$ [Republika [Czeska]$^{\text{relAdj\_country}}$]$^{\text{country}}$

---

**6. KIERUNKI ŚWIATA**

w podstawowym znaczeniu – nie anotujemy („poszedł na południe”, „słońce wstaje na wschodzie”)

w znaczeniu figuratywnym - wg instrukcji str. 151:

*przejawia się moda na Wschód*

(na życie w stylu, w jakim żyją ludzie określonej części świata) - **ORG**

*pojechał zarabiać na **Zachód***

(do zachodnich krajów Europy) – **GEOG**

*„(...) tysiąc sześćset kilometrów od bieguna północnego (...)”*

- anotujemy, trochę przylądek, trochę szczyt, trochę obiekt astronomiczny (wszystko wymienione w definicji GEOGa w instrukcji)

---

**7. Dopuszczamy anotacje identyczne zakresowo (a nawet z taką samą etykietą!)**

na [[Łazienkowskiej]$^{geog}$]$^{relAdj\_}$ **–** 1) geog, bo ulica + 2) relAdj_geog, bo derywat od Łazienek (geog)

## PLACE

---

**1. gmina + miasto** - anotujemy całość + anotujemy miasto,

gmina Krośniewice = [gmina [Krośniewice]$^{settlement}$ ]$^{region}$

za instrukcją:

" Kłopotliwe dla opisu zagnieżdżeń były też formy metonimiczne. W takich przykładach jak (9.101)–(9.102) Stupsk jest pierwotnie nazwą miejscowości, dlatego **skłanialiśmy się do oznaczania go jako nazwy zagnieżdżonej** (9.101)"

---

**2. parafia** – w instrukcji jest jako przykład placeName_district, ale może mieć znaczenia:

Msza odbyła się w parafii św. Kazimierza – **geog** (w budynku)

Zorganizowane przez parafię pw. Bartłomieja (...) – **org** (sprawczość)

Pochodzę z parafii rzekuń – **district** (dzielnica)

## RelADJ

---

**1.** Przymiotniki od państw i obszarów geopolitycznych, kulturowych odnosimy do najbardziej podstawowego znaczenia (patrz: 9.4.4. Niepewny status i pochodzenie przymiotników relacyjnych)

**polsko-rosyjskie** stosunki - relAdj od country

**arabski** - zawsze od country

---

**2.** skróty od nazw języków, np. **niem.** nie anotujemy

---

**3. polski** (jako przedmiot w szkole) – reladj, baza – country (jako odwołanie się do podstawowego znaczenia bazy)

---

**4.** Przymiotniki odnoszące się „geograficznie” jednak nie wiadomo dokładnie do czego – **relAdj od GEOG**

myśl **anglosaska**, język **łaciński**

(ALE! „**łacina”** NIE anotujemy, bo to rzeczownik; podobnie „język **Xhosa”** – nie anotujemy, bo to rzeczownikowa nazwa języka, a więc nie NER)

**zachodni** - ZAWSZE bedzie od relAdj_GEOG - o ile Zachód może być ORG lub GEOG (wg 151 w instrukcji), to przymiotnik od znaczenia pierwszego, czyli od GEOGa

(oczywiście wtedy kiedy jest od "Zachodu" a nie od "zachodu", typu "zachodni wiatr", bo wtedy w ogóle nie oznaczamy)

---

**5. sejmowy, senacki** - relAdj od ORG

---

**6. Nie anotujemy, jeśli derywacja jest zbyt długa (skomplikowana)**

uważany za symbol odnowy Kościoła w duchu **kluniackim**

„Duch kluniacki" od „Reforma kluniacka" od „Klasztor kluniacki" od „Klasztor w Cluny"

## PersDERIV

---

**1. warszawiacy** – w kontekście Legia Warszawa – od pierwszego znaczenia – czyli od SETTLEMENT

za instrukcją: [Niemcy]$^{persDeriv\_country}$ pokonali [Szwedów]$^{perdDeriv\_country}$ 3:2.

(nawet jeśli chodzi raczej o derywat od ORGa)

---

**2. Religie**

NIE ANOTUJEMY:

katolicki, katolicy, prawosławni, luteranie... - uznajemy, że to nazwy zbyt pospolite

franciszkański, karmelicki... - za instrukcją, nie anotujemy przymiotników, które nie wiadomo czy są od ORG (franciszkanie) czy od person (Franciszka)

---

**3. Narody, ludy, grupy etniczne**

Za instrukcją: „Istnieją tez narody, które nigdy nie wytworzyły własnego państwa i nie są związane z określonym terytorium, np. Romowie, których nazwy oznaczaliśmy jako derywacje osobowe bez określenia bazy derywacyjnej – por. (9.122). Podobne podejście przyjęto w stosunku do Żydów(…)"

· (Romowie, Arabowie, Żydzi) – persderiv od country

Ludy związane z konkretnym terytorium - Galowie, Germanie, Normanowie, Sumerowie, Frankowie, Sasi etc.

· Galowie (Galia) - persderiv od region

· Germanie (Germania) - persderiv od region

· Normanowie (Normandia) - persderiv od region