# NerKor annotálási útmutató

Simon Eszter

2020. október 11.

## 1. Bevezetés

A számítógépes nyelvészet egy interdiszciplináris terület, amelynek célja az emberi nyelv szerkezetének gépi modellálása, valamint a természetes nyelvek számítógépes feldolgozása. Az információkinyerés a számítógépes nyelvészet egyik fontos alterülete; célja, hogy strukturálatlan szövegből automatikusan hozzájussunk a számunkra értékes információhoz. Mivel ezen információ nagy része tulajdonnevek formájában jelenik meg a szövegben, ezért a tulajdonnév-felismerésnek (named entity recognition, NER) kiemelt jelentősége van.

A NER során egy bemeneti tokensorozatban kell megnevezett entitást (named entity, NE) alkotó intervallumokat kijelölünk, ezeket véges sok kategóriába besorolva. Egy gépi tanuló algoritmus kiértékelése manuálisan annotált korpusszal való összevetés útján történik, és szokásosan maga az algoritmus is egy ilyen korpuszból tanulja meg a paramétereit automatikus módon. Ezért van szükség nagy méretű kézzel annotált korpuszokra. Ez a célja ennek a projektnek is, amelynek a tervezett kimenete a NerKor korpusz, egy egymillió tokenes kézzel tulajdonnév-annotált korpusz.

Egy NE a szövegnek egy olyan eleme, amely a világ valamely entitására unikusan referál – tulajdonnévvel, rövidítéssel, mozaikszóval vagy becenévvel, például:

(1) *Kosztolányi Dezső*

(2) *Szilas Menti Mezőgazdasági Termelőszövetkezet*

(3) *United Nations Educational, Scientific and Cultural Organization*

(4) *Déli-Shetland-szk.*

(5) *IBM*

(6) *Kiss János altábornagy utca*

(7) *Műegyetem*

(8) *The Coca-Cola Co.*

(9) *Kovács Pistike*

## 2. Az annotálás alapelvei

Fontos, az annotálás során végig szem előtt tartandó szabályok:

- Csak neveket annotálunk. Névnek nevezzük azt a kifejezést, ami unikusan, vagyis egyedi módon referál a világ valamely entitására. Tehát nem annotálunk olyan frázisokat, amelyek ugyan a világnak valamely egyedi részére utalnak, de nem teljesen egyértelmű módon. Például a *József Attila Gimnázium* mindenképpen annotálandó, de a szövegben szereplő az *a suli* frázis nem, hiába derül ki a szövegből, hogy melyik iskolára utal.
- A nevek nem kompozicionálisak. Mivel a nevek jelölete nem a név részeinek a jelöletéből áll össze, ezért a neveket nem bonthatjuk részekre az annotáláskor. Például a *Kossuth Lajos utca* egy névként jelölendő, hiába van benne egy személynév. Mindig a leghosszabb nevet (a legkülsőbbet) jelöljük a jelölhetők közül.
- Nem annotálunk egymást átfedő vagy egymásba ágyazott neveket. Vagyis minden annotációnak be kell fejeződnie, mielőtt egy másik elkezdődik.
- A neveket mindig az elsődleges referenciájuk alapján annotáljuk, nem az aktuális kontextusnak megfelelő értelmében, vagyis a metonimikusan viselkedő nevek is mindig az elsődleges címkéjüket fogják megkapni.
- Ha az azonosított NE ragozott formában szerepel a szövegben, a raggal együtt, a teljes alakot annotáljuk.
- A NE-k képzett alakjait nem jelöljük. Nem annotálandók tehát az olyanok, mint *magyarországi, fideszes, petőfieskedő.*
- A NE-hez nem tartozik hozzá az esetleg előtte álló névelő. Kivétel az az eset, amikor a határozott névelő része a névnek, például *The Hague, The Times.*

## 3. NE-típusok

A következő típusokat annotáljuk:

**PERSON:** Személynevek, becenevek, aliasok.

**ORGANIZATION:** Olyan csoportok nevei, amelyek valamilyen szervezett struktúrával rendelkeznek, mint például intézmények, vállalatok, kormányzati hivatalok, sportcsapatok, múzeumok, egyetemek.

**LOCATION:** Földrajzilag vagy politikailag definiált helyek nevei, úgymint városok, országok, hegyek, völgyek stb. Idetartoznak az emberalkotta építmények is, mint a repterek, utak, gyárak, épületek stb.

**MISC:** A felsorolt típusok egyikébe sem tartozó nevek.

Az útmutatóban a NE-ket [szögletes zárójelek] közé tesszük. A példáknál csak az olyan típusú NE-ket jelöljük, amelyikről éppen szó van. A példákban a személyneveket a PER, a szervezetneveket az ORG, a helyneveket a LOC, a be nem sorolhatókat pedig MISC rövidítésekkel jelöljük.

## 3.1. Személynevek (PERSON)

Személyekre utalhatnak teljes személynevek, becenevek, művésznevek, álnevek, aliasok. A kitalált személyek, úgymint mozihősök, mesefigurák, mitológiai alakok, illetve a szentek, bibliai alakok nevei is személynévként annotálandók, például:

(10) [$Ady_{PER}$] *írói álneve* [$Ida_{PER}$]

(11) *a legkisebb gyerek, aki gyakran játszik* [$Mikrobival_{PER}$]

(12) *zenéjével meglágyította* [$Hádész_{PER}$] *és* [$Perszephoné_{PER}$] *szívét*

A családnevek, az uralkodóházak nevei is személyekre, egészen pontosan személyek csoportjára referálnak, ezért azokat is személynévként kell megjelölni, például:

(13) *a* [$Széchényi_{PER}$] *család Nógrád megyéből származik*

(14) *a* [$Károlyiak_{PER}$] *Apáti nevű faluját felgyújtották*

## 3.2. Szervezetnevek (ORGANIZATION)

Azok a tulajdonnevek, amelyek egy szervezett struktúrával rendelkező csoportra referálnak, szervezetnévként annotálandók. A következők mind ilyenek:

- Cégek, vállalatok

  (15) *a* [$SERCO\ Kft._{ORG}$] *az eltelt évek során jelentős fejlődésen ment keresztül*

  (16) *1878-ban Grosvernor Lowry-val létrehozzák az* [$Edison\ Electric\ Light\ Co.\text{-}t_{ORG}$]

- Tőzsdék

  (17) *ingyenes tesztidőszak a* [$Budapesti\ Értéktőzsde_{ORG}$] *és a* [$Bécsi\ Tőzsde_{ORG}$] *kereskedési adataira*

- Multinacionális szervezetek

  (18) *az* [$Európai\ Unió_{ORG}$] *ezen a néven 1992-ben jött létre*

- Politikai pártok

(19) *bántalmazták a [Fidesz$_{ORG}$] egyik ajánlószelvényeket gyűjtő aktivistáját*

- Sportcsapatok

(20) *A [Budapest Black Knights$_{ORG}$] csapata fölényesen legyőzte a [Szolnok Soldiers$_{ORG}$] csapatát.*

- Katonai szervezetek

(21) *Az [Észak-atlanti Szerződés Szervezete$_{ORG}$] székhelye Brüsszelben van.*

Ezeken kívül vannak még azok a nevek, amelyek olyan épületekre vagy emberalkotta építményekre referálnak, amelyekre igaz az, hogy valamilyen szervezett struktúrával rendelkeznek, jellemzően aktorként szerepelnek az adott szövegkontextusban, és olyanokat tudnak csinálni, mint döntést hozni, árat emelni, nyilatkozni valamiről stb. A szövegben így viselkedő ilyenfajta NE-ket szervezetnévként kell annotálni. Idetartoznak például az alábbiak:

- Kórházak, egészségügyi intézmények

(22) *Fővárosi Önk. Péterfy Sándor utcai Kórház-Rendelőintézet*

(23) *Delej utcai Vérellátó Központ*

- Hotelek

(24) *Erzsébet Szálloda*

(25) *Four Seasons Hotels and Resorts*

- Színházak, múzeumok

(26) *Szépművészeti Múzeum*

(27) *Holdvilág Kamaraszínház*

- Egyetemek

(28) *Kossuth Lajos Tudományegyetem*

(29) *UCLA*

- Kormányzati hivatalok

(30) *Parlament*

(31) *Honvédelmi Minisztérium*

A szervezeteknek sajátjuk, hogy van székhelyük, és előfordul, hogy a szervezet nevét helymegjelölésként használjuk, például:

(32) *tűz ütött ki a* [*Kapos Hotelben*$_{ORG}$]

(33) *sztrájk az* [*SZFE-n*$_{ORG}$]

Mivel a neveket mindig az elsődleges referenciájuk alapján annotáljuk, ilyenkor is intézménynévként kell annotálni ezeket a neveket.

A közvetlenül a szervezetnév után álló közneveket, melyek hiányában a név nem ugyanarra referálna, a névvel együtt annotáljuk, például:

(34) *Az új menetrenddel úgy látszik a* [*Keleti pályaudvar*$_{ORG}$] *utastájékoztatása is megváltozott.*

Nem tartoznak viszont a szervezetnévhez a magyarázó, pontosító funkciójú elemek, illetve az alkalmi jelzők sem, például:

(35) *A* [*Botond*$_{ORG}$] *étterem mindennap 9-től 24 óráig várja vendégeit.*

Az általános intézményneveket, mint *rendőrség* vagy *kormány* nem annotáljuk, mert ezek nem unikusan jelölnek egy konkrét entitást.

Nem lehet továbbá szervezetnév melléknév.

## 3.3. Helynevek (LOCATION)

A helynévnek annotálandó entitások közé tartoznak többek között a kontinensek, az országok, a régiók, a városok, a települések, a repterek, az utak, a gyárak, az óceánok, a tengerek, a folyók, a szigetek, a tavak, a nemzeti parkok, a hegyek és a mitikus helyek, például:

(36) [*Franciaországot*$_{LOC}$] *kilenc ország határolja.*

(37) [*Szihalom*$_{LOC}$] *község* [*Heves megye*$_{LOC}$] [*Füzesabonyi kistérségében*$_{LOC}$].

(38) *A* [*Bükki Nemzeti Park*$_{LOC}$] *mintegy 95 százalékát erdő borítja.*

(39) [*Gatwick*$_{LOC}$] *délre,* [*Stansted*$_{LOC}$] *észak-keletre,* [*Luton*$_{LOC}$] *északnyugatra fekszik* [*Londontól*$_{LOC}$].

(40) *Platón dialógusaiban részletesen szól* [*Atlantisz*$_{LOC}$] *szigetéről.*

### 3.3.1. Összetett kifejezések

Az olyan összetett kifejezésekben, ahol földrajzi nevek vesszővel elválasztva szerepelnek, és a második név nagyobb helyre referál, tehát egyfajta pontosító funkciót tölt be, a neveket külön annotáljuk, például:

(41) [*Los Angeles*$_{LOC}$], [*Calfornia*$_{LOC}$]

(42) [*Budapest*$_{LOC}$], [*Magyarország*$_{LOC}$]

### 3.3.2. Köznévi tagok

Vannak olyan földrajzi nevek, melyek köznévi utótagot tartalmaznak. A közvetlenül a földrajzi név előtt vagy után álló köznévi frázisok, melyek hiányában a név nem ugyanarra referálna, a névvel együtt annotálandók, mint például az alábbiak:

(43) [*Váci utca*$_{LOC}$]

(44) [*Erzsébet híd*$_{LOC}$]

(45) [*Baranya megye*$_{LOC}$]

(46) [*Duna–Tisza köze*$_{LOC}$]

Nem tartoznak viszont a földrajzi névhez a magyarázó, pontosító funkciójú elemek, illetve az alkalmi jelzők sem, például:

(47) [*Kent*$_{LOC}$] *grófság*

(48) [*New York*$_{LOC}$] *állam*

(49) [*Gyöngyös*$_{LOC}$] *város*

(50) [*Mátra*$_{LOC}$] *hegység*

(51) [*Duna*$_{LOC}$] *folyó*

(52) *az olasz* [*Alpok*$_{LOC}$]

(53) *a lengyel* [*Magas-Tátra*$_{LOC}$]

(54) *a gyönyörű* [*Alpok*$_{LOC}$]

(55) *„Mit nekem te zordon* [*Kárpátoknak*$_{LOC}$]*...”*

## 3.4. Egyebek (MISC)

Ebbe a kategóriába kerülnek azok, amelyek NE-k, de a felsorolt kategóriák egyikébe sem illenek bele, mint a könyvcímek, újságnevek, konferencianevek, márkanevek, tőzsdeindexek nevei, programozási nyelvek, például:

(56) *A* [*Le Monde*$_{MISC}$] *francia napilap.*

(57) *fedezze fel a* [*Fiat*$_{MISC}$] *modelleket*

(58) *Érdekel a* [*Python*$_{MISC}$] *programozás?*