Annotálási útmutató a hunner projekthez

Simon Eszter

1. Bevezetés

Egy named entity (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*

Fontos, az annotálás során végig szem előtt tartandó szabályok:

- Csak neveket annotálunk[1]. 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

[1] Kivéve a ROLE és RANK kategóriákat.

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 *tag-for-meaning* elvét követjük. Vagyis egy kifejezést mindig az aktuális kontextusnak megfelelő értelmében annotálunk.

- Ha az azonosított NE ragozott formában szerepel a szövegben, a raggal együtt, a teljes alakot annotáljuk.

- A NE-k képzett alakjait nem jelöljük[2]. Nem annotálandók tehát az olyanok, mint: *fideszes, Orbán Viktor-i, gyurcsányozik, petőfieskedő.*

- A NE-hez nem tartozik hozzá az esetleg előtte álló névelő. Kivétel az az eset, amikor a határozott névelő része a névnek, pl. *The Hague, The Times.*

2. NE-típusok

A következő típusokat annotáljuk:

PERSON: Személynevek, becenevek, aliasok.
ROLE: Embereknek valamely szervezetnél betöltött szerepét, beosztását jelöli. Csak akkor jelöljük, ha ugyanabban a mondatban megtalálható a személy- és a szervezetnév is.
RANK: Személynevek közvetlen környezetében előforduló cím- és rangjelölő szavak.

[2]Kivéve a földrajzi nevek esetében az *-i/-beli* képzőkkel képzett mellékneveket.

ORGANIZATION: Olyan csoportok nevei, amelyek valamilyen szervezett struktúrával rendelkeznek, mint pl. intézmények, vállalatok, kormányzati hivatalok, sportcsapatok, múzeumok, egyetemek.
LOCATION: Földrajzilag vagy politikailag definiált helyek nevei, úgy mint városok, országok, hegyek, völgyek stb. Idetartoznak az emberalkotta építmények is, mint a repterek, utak, gyárak, épületek stb.
BRAND/PRODUCT: Márkanevek.
TITLE: Műcímek; egy könyv, vers, zenedarab, festmény stb. neve.
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 szerepeket a ROLE, a rangjelölőket a RANK, a szervezetneveket az ORG, a helyneveket a LOC, a márkaneveket a BRP, a műcímeket a TTL, a be nem sorolhatókat pedig MISC rövidítésekkel jelöljük.

2.1. Személynevek (PERSON)

Személyekre utalhatnak teljes személynevek, becenevek, művésznevek, álnevek, aliasok. A kitalált személyek, úgy mint mozihősök, mesefigurák, mitológiai alakok, ill. a szentek, bibliai alakok nevei is személynévként annotálandók.

A családnevek, az uralkodóházak nevei is személyekre, egészen pontosan személyek csoportjára referálnak, ezért azokat is személynévként kell megjelölni. Az angol családnevekhez jellemzően hozzátartozik a (nagybetűs) határozott névelő is, ilyenkor ez is a név része:

(10) *The Smiths*

2.2. Beosztások (ROLE)

A szerepek, beosztások egy szervezet és egy személy relációját jelölik, tehát egy személy szerepét az intézményi hierarchiában, nem szakmát, foglalkozást vagy tevékenységet. Akkor annotálandók, ha ugyanabban a mondatban megtalálható a szervezet és a személy is, amire/akire vonatkoznak.

(11) [Kis_{PER}], *a* [$Gittegylet_{ORG}$] [$elnöke_{ROLE}$] *megnyitotta a gyűlést.*

(12) [*Kis Ildikó*$_{PER}$], [*alkalmazott*$_{ROLE}$] *a* [*Műegyetemen*$_{ORG}$], *két gyermek anyja.*

Azokban az esetekben tehát, amikor a mondatban nem jelenik meg expliciten mind a személynév, mind a szervezetnév, nem jelöljük a beosztásokat. Például:

(13) [*Vasile Blaga*$_{PER}$] *belügyminisztertől tegnap azt kérdezték...*

(14) *Az esemény kiemelt támogatója az egészségügyi miniszter.*

(15) [*Gyurcsány Ferenc*$_{PER}$], *a magyar miniszterelnök*

2.3. Rangjelölők (RANK)

Idetartoznak azok a névelőzékek, amelyek a személy rangját, címét jelölik, mint például *Sir, Lord, gróf*; *PhD., Dr., Prof.*.

(16) [*Sir*$_{RANK}$] [*Paul McCartney*$_{PER}$] *új kórusművet írt*

(17) [*Lord*$_{RANK}$] [*Rothermere*$_{PER}$] *1927-es kampánya*

(18) *ükapja* [*gr.*$_{RANK}$] [*Széchenyi Ferenc*$_{PER}$] *volt*

Egymás mellett álló rangjelölőket külön-külön annotálunk, például:

(19) [*Prof.*$_{RANK}$] [*Dr.*$_{RANK}$] [*Fábián Eszter*$_{PER}$] *oktató kurzusai*

Rangjelölőként annotálandók az angolban szokásos névelőzékek, amelyek arról árulkodnak, hogy a személy házas-e (*Mr., Mrs.*), és ezek magyar megfelelői is (*úr, asszony, kisasszony*). Vannak viszont szorosan a névhez tartozó szuffixumok (*Jr., Sr.*) és ezek magyar megfelelői (*ifj., id.*) – ezek a személynévhez tartoznak. Tehát:

(20) [*Jock Ewing, Jr.*$_{PER}$], *a Dallas antikrisztusa*

(21) [*Ifjabb Hegedűs Lórántt*$_{PER}$] *még a vita elején arról beszélt...*

(22) *a főszereplő természetesen ez a bizonyos* [*Mr.*$_{RANK}$] [*Pornowsky*$_{PER}$]

(23) *nem most ismertük meg* [*Tarlós*$_{PER}$] [*urat*$_{RANK}$]

2.4. Szervezetnevek (ORGANIZATION)

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

- Cégek, vállalatok

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

  (25) *1878-ban Grosvernor Lowry-val létrehozzák az [Edison Electric Light Co.-t$_{ORG}$]*

- Tőzsdék

  (26) *Ingyenes tesztidőszak a [Budapesti Értéktőzsde$_{ORG}$] és a [Bécsi Tőzsde$_{ORG}$] kereskedési adataira*

- Multinacionális szervezetek

  (27) *az [Európai Unió$_{ORG}$] ezen a néven 1992-ben jött létre*

- Politikai pártok

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

- Sportcsapatok

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

- Katonai szervezetek

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

A szervezeteknek sajátjuk, hogy van székhelyük, és sűrűn előfordul, hogy a helyre a szervezetnévvel utalunk. Ilyenkor – a tag-for-meaning elvet követve – helynévként annotáljuk a szervezetnevet. Erről részletesebben lásd a 2.5. fejezetet.

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

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

  (32) *Delej utcai Vérellátó Központ*

- Hotelek

  (33) *Erzsébet Szálloda*

  (34) *Four Seasons Hotels and Resorts*

- Színházak, múzeumok

  (35) *Szépművészeti Múzeum*

  (36) *Holdvilág Kamaraszínház*

- Egyetemek

  (37) *Kossuth Lajos Tudományegyetem*

  (38) *UCLA*

- Kormányzati hivatalok

  (39) *Parlament*

  (40) *Honvédelmi Minisztérium*

Mivel a tag-for-meaning elvét követjük, ezeket csak akkor annotáljuk szervezetnévként, amikor a fent leírt módon viselkednek. Előfordulhat, hogy helyre referálnak; erről lásd a 2.5. fejezetet.

A közvetlenül a szervezetnév után álló deskriptív funkciójú közneveket a névvel együtt annotáljuk, például:

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

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

Az általános intézményneveket, mint *rendőrség* vagy *kormány* nem annotáljuk, mert ezek nem unikusan jelölnek egy konkrét entitást. Nem lehet továbbá szervezetnév melléknév.

2.5. 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:

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

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

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

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

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

2.5.1. 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 együtt annotáljuk, például:

(48) [*Los Angeles, California*$_{LOC}$]

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

2.5.2. Köznévi tagok

Vannak olyan földrajzi nevek, melyek köznévi utótagot tartalmaznak. A közvetlenül a földrajzi név előtt vagy után álló, magyarázó, pontosító funkciójú köznévi frázisok a névvel együtt annotálandók, mint például az alábbiak:

(50) [*Váci utca*$_{LOC}$]

(51) [*Erzsébet híd*$_{LOC}$]

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

(53) [*Duna–Tisza köze*$_{LOC}$]

(54) [*Kent grófság*$_{LOC}$]

(55) [*New York állam*$_{LOC}$]

(56) [*Gyöngyös város*$_{LOC}$]

(57) [*Mátra hegység*$_{LOC}$]

(58) [*Duna folyó*$_{LOC}$]

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

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

Nem tartoznak viszont a földrajzi névhez az alkalmi jelzők, mint:

(61) *a gyönyörű* [*Alpok*$_{LOC}$]

(62) *"Mit nekem te zordon* [*Kárpátoknak*$_{LOC}$]*..."*

Ha végképp nem tudod eldönteni, hogy egy köznévi tag a földrajzi név része-e, inkább annotáld a részeként.

2.5.3. Melléknevek

Földrajzi névként annotálandók az *-i/-beli* képzős melléknevek is, amennyiben azok földrajzi névből lettek képezve, és helyre referálnak. Például:

(63) *a* [*budapesti*$_{LOC}$] *események*

(64) *a* [*romániai*$_{LOC}$] [*Verespatakon*$_{LOC}$] *levő bánya*

2.6. Helyre referáló szervezetnevek, szervezetre referáló helynevek

A szervezeteknek nemcsak felépítési struktúrájuk van, hanem jellemzően székhelyük is, ezért sokszor előfordul a szövegben, hogy a helyre a szervezetnévvel utalunk. Ilyet tapasztalunk például a cégneveknél vagy a multinacionális szervezeteknél is, de sokkal jellemzőbb azokban az esetekben, amikor a név egy építményre utal. Fontos: ilyenkor az építménynév jellemzően nem aktorként szerepel, ahogy a szervezetneveknél (2.3. fejezet), hanem helyre referál a mondatban. A tag-for-meaning elvet követve az így viselkedő neveket a szövegben aktuális jelentésük alapján annotáljuk, de azt is jelöljük, hogy eredetileg intézménynevek (LOC:ORG). Például:

(65) *Az [Európai Uniónak$_{LOC:ORG}$] 20 országgal van közös szárazföldi határa.*

(66) *Az [Európai Unió$_{ORG}$] elfogadta a magyar konvergenciaprogramot.*

(67) *A [János kórházban$_{LOC:ORG}$] sok a macska.*

(68) *Sztrájkolnak a [János kórház$_{ORG}$] dolgozói.*

(69) *A [Nemzeti Múzeum$_{LOC:ORG}$] az 1848. március 15-ei események egyik fő helyszíne volt.*

(70) *Új kiállítást nyitott a [Nemzeti Múzeum$_{ORG}$].*

Hasonlóan a városok, országok és egyéb helyek neveit is használhatjuk az ott található intézmények nevei helyett. Mivel ilyenkor a helynév egy szervezetre referál, a tag-for-meaning elvnek megfelelően szervezetnévként kell annotálni, de itt is úgy, hogy jelöljük, hogy ez eredetileg helynév (ORG:LOC).

(71) *[Washington$_{ORG:LOC}$] [Moszkvával$_{ORG:LOC}$] tárgyal.*

(72) *[Washingtonban$_{LOC}$] található a [Fehér Ház$_{LOC}$].*

(73) *A [Fehér Ház$_{ORG:LOC}$] semmi információt nem ad ki.*

(74) *[Magyarország$_{ORG:LOC}$] miniszterelnöke [Németországba$_{LOC}$] látogatott.*

Hasonló logika alapján annotáljuk az olyan eseteket is, amikor egy helynévvel egy sportcsapatra utalunk:

(75) *A [Manchester$_{ORG:LOC}$] ma a [Münchennel$_{ORG:LOC}$] játszik.*

2.7. Márkanevek (BRAND/PRODUCT)

Idetartoznak a gyártmányoknak, termékeknek, készítményeknek márkanévként használt elnevezései:

(76) [*Hitachi GX3800*$_{BRP}$]

(77) [*Windows XP*$_{BRP}$]

(78) [*Coca-Cola*$_{BRP}$]

A márkanév után tájékoztató funkcióval álló köznévi kifejezés nem tartozik a névhez, nem annotáljuk a névvel:

(79) [*Lavazza*$_{BRP}$] *kávé*

(80) [*hp Laserjet*$_{BRP}$] *nyomtató*

2.8. Műcímek (TITLE)

Idetartoznak az ún. állandó címek, vagyis az újságok, hetilapok, folyóiratok címei, illetve az internetes hírportálok nevei.

(81) *mondta az* [*origo-nak*$_{TTL}$] *a főváros VI. kerületi önkormányzatának munkatársa*

(82) *A pénzügyminiszterek várhatóan módosítás nélkül elfogadják a javaslatot – írja a* [*Világgazdaság*$_{TTL}$].

Idetartoznak továbbá az ún. egyedi címek, vagyis az egyedi műalkotások: regények, versek, cikkek, esszék, festmények, szobrok, zeneművek, filmek, műsorok, színdarabok stb. címei.

(83) *David Bowie többek közt szerepelt még Martin Scorsese* [*Krisztus utolsó megkísértésében*$_{TTL}$] *is.*

(84) *Az* [*Ének a lyukas zászlóról*$_{TTL}$] *alcímű kötet percről percre követhető történetet kínál az érdeklődőknek.*

(85) *Az őszi szezon Kieselbach-árverésének unikális képe Moholy-Nagy László 1919-es* [*Önarcképe*$_{TTL}$].

2.9. Egyebek (MISC)

Ebbe a kategóriába kerülnek azok, amik az annotátor szerint NE-k, de a felsorolt kategóriák egyikébe sem illenek bele.