Annotációs útmutató
anafora és koreferenciaannotáláshoz

A feladat a gazdag nyelvi annotációval ellátott szövegekhez egy plusz elemzési réteg hozzáadása kézzel. A plusz elemzési réteg az anaforikus- és koreferenciakapcsolatokat tartalmazza. Az útmutató részletes leírást ad a jelenségről és ismerteti az annotációs felületet is.

1. Anafora és koreferencia

Az anafora- és a koreferenciaannotálás lehetővé teszi, hogy egy szövegen belül végigkövessük az abban szereplők sorsát, ezért hasznos adatot szolgáltat az információkinyerési feladatokban.

A névmások referenciájukat egy, a mondatban szereplő másik tartalmas elemtől kapják. Az anafora referenciáját a szövegelőzményben jelen lévő antecedensétől kapja, arra „utal vissza". A katafora fordítva működik, ekkor a névmást követi a szövegben az antecedense. Az anafora és antecedense közötti kapcsolatot grammatikai tényezők irányítják (pl. kötéselvek).

A koreferencia esetében az utaló elem nem egy, a szövegelőzményben szereplő szóra utal vissza, hanem közös referenciája van egy, a szövegelőzményben szereplő szóval (tehát ugyannarra a valóságbeli dologra hivatkoznak). A koreferens elemek közötti viszony alapaja lexikális jellegű.

Az anafora és a koreferencia jelenségét hasonlóképpen annotáljuk. Az anafora esetében megjelöljük, hogy melyik elem az antecedense, a koreferens elem esetében pedig megjelöljük, hogy melyik a hozzá legközelebb eső, őt megelőző elem, amivel közös koreferenciája van. A koreferencia annotálása láncokat fog eredményezni, amelynek első eleme egy bizonyos referenciájú tartalmas elem, a többi eleme pedig az összes, vele koreferens elem lineáris sorrendben.

Mind az anforikus-, mind a koreferenciakapcsolatok megjelenhetnek mondaton belül, de mondathatárokon is átívelhetnek.

1.1. Anafora

A következő névmások szerepelnek anaforikus kapcsolatokban utalóelemként:

- személyes
  ***Én** is beléptem a terembe. Már mindenki ismert **engem**.*
  Én** is beléptem a terembe. Már mindenki ismert **ZÉRÓ.OBJ.E/1.
  ZÉRÓ.SUBJ.E/1.** beléptem a terembe. Már mindenki ismert **ZÉRÓ.OBJ.E/1.
  ***Pista** elhozta a **ZÉRÓ.POSS.E/3.** kutyáját.*

- mutató
  *Pista elhozta **Annát** is, **az** meg ránk se bagózott.*
  ***Pista** belépett a terembe. **Ő** volt az utolsó.*

- kölcsönös
  A kutyák *körbeszagolták* ***egymást****.*
  Pista és Mari *körberajongták* ***egymást****.*

- visszaható
  Pista *rögtön bemutatta* ***magát****.*

- vonatkozó
  Pista *volt az,* ***aki*** *mindig elkésett.*
  Oda *megyek,* ***ahova*** *hívnak.*

- birtokos
  A kutya *nem látta, melyik gazda az* ***övé****.*

A személyes névmás (akár vonzatként, akár birtokos szerkezetben jelöletlen birtokosként) testetlen is lehet, ám ekkor is részt vesz anaforikus kapcsolatban, amint a fenti példákon látszik.

1.2. Koreferencia

Néhány példa a különböző típusó koreferenciakapcsolatokra:

- ***Erős Pista*** *megérkezett a terembe. Mindenki ismerte* ***Erős Pistát****.*
- ***Erős Pista*** *megérkezett a terembe. Mindenki ismerte* ***Pistát****.*
- *Elszökött* ***a kutyám****. Egyébként* ***nagyon okos eb****.*
- *Elszökött* ***a kutyám****. Egyébként* ***nagyon okos állat****.*
- *Elszökött* ***a kutyám****. Egyébként* ***nagyon okos uszkár****.*
- *Szép nagy* ***házat*** *vettünk. A kedvencem* ***az emeleti fürdőszoba****.*
- ***Erős Pista*** *megérkezett a terembe. Mindenki ismerte* ***azt a hülyét****.*
- ***Elszökött*** *a kutyám. Az egész hétvége elment* ***a szökés*** *miatt.*

2. Az annotáló felület

Az annotálást egy előkészített google táblázatban végezzük el. A táblázat a `tsv`-ként importált adatokat tartalmazza, az oszlopok tartalma a következő:

- **A – TOKEN ID**: a token mondatbeli sorszáma
- **B – FORM**: szóalak
- **C – LEMMA**: szótő
- **E – UPOS**: szófajcímke
- **F – FEATS**: morfológiai jegyek
- **G – HEAD**: függőségi elemzés: az anyacsomópont azonosítója

- **H – DEPREL**: függőségi elemzés: a függőségi kapcsolat típusa
- **J – REFREL**: anafora vagy koreferencia típusa
- **K – REFTYPE**: antecedens vagy előző koreferens elem azonosítója
- **L – javított szóalak**
- **M – javított tő**
- **N – javított szófaj**
- **O – javított morfológia**
- **P – tokenizálási hiba**
- **Q – függőségi hiba**

A mondatokat egy üres sor választja el egymástól, a mondat első kommentjében a teljes mondat olvasható, a második komment pedig a mondat dokumentumbeli sorszámát tartalmazza.

A token mondatbeli sorszáma egyesével növekvő szám, a zéró létige esetében az őt megelőző elem sorszámából képzett sorszám áll (.1), a zéró névmások esetében a zéró névmást megelőző elem sorszámából képzett sorszám áll a zéró elem grammatikai szerepével kiegészítve (.SUBJ, .OBJ vagy .POSS).

A szófajcímke és a morfológiai jegyek a Universal Dependencies címkekészletét használják. A címkekészlet feloldásához külön dokumentum[1] nyújt segítséget.

A függőségi elemzéshez tartozó két oszlop tartalma kevésbé releváns a jelenlegi annotálási feladat elvégzéséhez, mert az oszlopos megjelenítés nem teszi lehetővé a függőségi fa tanulmányozását. Bizonyos esetekben azonban segítséget nyújt az egyes szavak fában elfoglalt szerepe, hiszen az alanyok és egyéb vonzatok gyakran szerepelnek anaforikus és koreferenciakapcsolatokban.

Az annotálás során a REFREL és REFTYPE oszlopokba dolgozunk, ám előfordulhat, hogy más elemzési rétegben (tokenizálás, tövesítés, szófajcímke, morfológiai elemzés, függőségi elemzés) fedezünk fel hibát. A további hibákat az L-Q oszlopokban kell jelezni, illetve javítani, a hiba típusától függően. Ehhez az útmutatót lásd a 4. fejezetben!

3. Az annotálás módja

Mind az anaforikus-, mind a koreferenciakapcsolatok esetében külön figyelmet kell fordítani a mondathatáron átívelő kapcsolatokra. Ezért az annotálás első lépéseként célszerű elolvasni az egész szöveget, amihez a mondatok elemzés nélkül egy külön fájlban biztosítva vannak.

A névmási anafora esetében az anafora J (**REFREL**) oszlopába az antecedense mondatszámából (# sent_id) és az azonosítójából (**TOKEN ID**) képzett azonosítót kell írni. Emellett az anaforikus kapcsolat típusának rövidítését a K (**REFTYPE**) oszlopba. Az anaforikus kapcsolatok jelölését az 1. táblázat tartalmazza.

Három további típust is fel kell ismerni:

1. általános alany: ebben az esetben a névmás referenciája nem konkrét, de a szövegben további névmások mégis visszautalhatnak rá. Az első előfordulás esetén csak a K (**REFTYPE**) oszlopba kell beírni ezt a típust: **arb**.

[1] https://github.com/dlt-rilmta/panmorph/blob/master/panmorph_ud.pdf

névmás típusa	jelölés a táblázatban
személyes	prs
mutató	dem
kölcsönös	recip
visszaható	refl
vonatkozó	rel
birtokos	poss

1. táblázat. Az anaforikus kapcsolatok típusainak rövidítései.

2. beszélő: retorikai jellegű szövegek esetén előfordulhat, hogy egy névmás referense maga a beszélő (író). Jelölése az általános alanyhoz hasonlóan történik, típusa: **speak**.

3. hallgató: retorikai jellegű szövegek esetén előfordulhat, hogy egy névmás referense a hallgató (olvasó). Jelölése az általános alanyhoz és a beszélőhöz hasonlóan történik, típusa: **addr**.

A koreferenciakapcsolat esetében az adott szó J (**REFREL**) oszlopába a hozzá legközelebbi, őt megelőző, vele koreferenc szó mondatszámából (# sent_id) és az azonosítójából (**TOKEN ID**) képzett azonosítót kell írni. A koreferenciakapcsolat esetében a K (**REFTYPE**) oszlopban kell jelölni a koreferenciát. Kétféle koreferenciakapcsolatot jelölünk: a **holo** kapcsolattípust akkor válasszuk, ha a két elem között rész-egész kapcsolat áll fent, egyéb esetben **coref** a kapcsolat típusa.

Az anafora esetében a kapcsolat egy névmás és egy referenciális elem között vagy két névmás között áll fenn. A koreferencia esetében mindig tartalmas szóra utal vissza az utalóelem, nem névmásra!

4. A többi elemzési réteg hibái

Előfordulhat, hogy hibás elemzéssel találkozunk a korábban már ellenőrzött elemzési rétegekben.

Ha tokenizálási hibával találkozunk (két szó tévesen egybe van írva, egy szó tévesen külön van írva, nincs leválasztva az írásjel) vagy nincs beszúrva a megfelelő zéró elem (zéró létige, zéró névmások), akkor a táblázat P (tokenizálási hiba) oszlopában jelezni kell, hogy tokenizálási hiba lépett fel (tehát ezeket a hibákat csak jelezni kell, nem kell javítani).

Ha hibás szóalakkal, tővel, szófajcímkével vagy hibás morfológiai jegyekkel találkozunk, abban az esetben be kell írni a megfelelő címkét a megfelelő oszlopába (L javított szóalak, M javított tő, N javított szófaj, O javított morfológia oszlopokba) a javított értéket.

Ha hibás függőségi elemzéssel találkozunk, akkor a táblázat Q (függőségi hiba) oszlopában jelezni kell, hogy hibát találtunk a függőségi elemzésben (tehát ezeket a hibákat csak jelezni kell, nem kell javítani).

5. Munkaidő-nyilvántartó

A munka tempóját egy munkaidő-nyilvántartó táblázatban monitorozom, amelybe minden egyes annotált fájlhoz meg kell adni, hogy körülbelül hány percet vett igénybe az annotálása. A táblázatban a fájl melletti cellában jelezni kell, ha olyan probléma merült fel, amelyet nem lehet a fájlhoz tartozó felületen kijavítani vagy jelezni. Nem muszáj sorban haladni a fájlokkal, de egy fájlt érdemes egy lendülettel befejezni, ne legyen félbehagyott dokumentum a mappában.