Annotációs útmutató az emMorph címkék javításához

A feladat: az `e-magyar` kimeneteként kapott tokenizálás, tő és morfológiai elemzés kézi ellenőrzése, javítása.

1. A felület

Az ellenőrzést – és az esetleges javítást – egy google táblázatban végezzük el. A táblázat a `tsv`-ként importált adatokat tartalmazza, az oszlopok tartalma a következő:

- **A token**
- **C tő**
- **D részletes címke**
- **E tag** (emMorph címke)
- **F helyes** (a helyes elemzés megjelölésére szolgáló oszlop)
- **G javított tő** (az esetlegesen javított szótő)
- **H tokenizálás** (a tokenizálás javítására szolgáló oszlop)
- **I javított token** (a token javítására szolgáló oszlop)
- **J megjegyzés** (megjegyzés)

A `tsv` fájl úgy áll elő, hogy a szöveg minden egyes tokenjéhez az `e-magyar` emTag[1] modulja által egyértelműsített elemzése alá felsorakozik az adott token összes szóbajöhető elemzése, amelyeket az `e-magyar` morfológiai elemző és tövesítő modulja, az emMorph+emLem[2] ad. Az annotátor feladata eldönteni, hogy az emTag kimenete helyes-e, ha nem, akkor a lehetséges elemzések közül melyik helyes, valamint ha egyik se helyes, akkor be kell vinnie a helyes elemzést és tövet. A J oszlopba egyéb megjegyzéseket lehet írni.

A táblázat az ellenőrzés és javítás megkönnyítésére formázásokat tartalmaz. A tokenek kék háttérrel jelennek meg és a mondatokat egy üres sor választja el egymástól. A táblázat rejtett oszlopokat is tartalmaz, ezeknek az oszlopoknak a tartalma nem szükséges az annotáláshoz. Az annotátor csak az F, G, H, I és J oszlopokat szerkesztheti.

2. A feladat

Az annotátornak az F oszlopban kell megjelölnie a helyesnek ítélt elemzést. A helyes elemzést az F oszlopba írt X jelöli – a dokumentum szerkesztése előtt minden token esetében az emTag kimeneteként kapott, az adott tokenhez tartozó első sorban. Ha az annotátor úgy ítéli, hogy nem jó az emTag kimenete – akár a tő, akár a morfológiai címke esetében –, akkor az F oszlopból kitörli az X-et, majd az alatta felsorolt többi elemzés közül kiválasztott helyes elemzés mellé írja az X-et ugyanebben az oszlopban. Az X-szel megjelölt elemzés sora sárga kiemeléssel jelenik meg. Ha a tokenhez tartozó első sorban felkiáltójel szerepel az F oszlopban, akkor hiányzik a megfelelő részletes elemzés a tokenhez. Ekkor az annotátornak muszáj kiválasztania egy másik elemzést. Erre a rózsaszín kiemelés hívja fel a figyelmet. A javítás végére minden tokenhez maximum egy kiemelt sor tartozzon és nem maradhat felkiáltójel az F oszlopban (tehát tokenenként egy darab sárgával kiemelt sor legyen és ne legyen a dokumentumban rózsaszínnel kiemelt sor)!

[1] `https://e-magyar.hu/hu/textmodules/emtag`

[2] `https://e-magyar.hu/hu/textmodules/emmorph`

2.1. Kézi elemzés megadása

Ha nem talál megfelelő elemzést felsorolva, akkor a token első sorába, az F oszlopba – tehát az eredetileg X-et tartalmazó cellába – kézzel beírja a jólformált, helyes morfológiai címkét. Ebben az esetben egy sor sem kap sárga kiemelést.

2.1.1. A helyes elemzés kiválasztása

A helyes elemzés kiválasztásához vagy kézi megadásához az emMorph morfológiai kódlistája az `e-magyar` honlapján elérhető[3].

Az elemzés ellenőrzéséhez az emMorph önmagában is használható[4], melynek módja a következő: az url-ben kell megadni az elemzendő szót, pl. `https://emmorph.herokuapp.com/dstem/kismajmokkal`. Ekkor megjelenik a szó összes lehetséges elemzése a tővel, a morfológiai kóddal és egy részletesebb elemzéssel együtt.

2.2. Helyes tő kézi megadása

Ha az eredeti elemzéshez tartozó tő is helytelen, és nincs olyan tő+elemzés pár, amely közül mindkettő helyes lenne, akkor a megfelelő elemzés kézi megadása mellett a helyes tövet is meg kell adni a G oszlopban. Amennyiben az eredeti szóalak helytelen helyesírású vagy archaikus alak, a sztenderd helyesírású tövet kell megadni.

2.3. Helyes szóalak kézi megadása

Előfordulhat, hogy az elemzetlen bemenet eleve hibás szóalakot tartalmazott, legyen szó akár elütésről, akár helyesírási hibáról. Ebben az esetben a szóalakot is ki kell javítani. A helyes szóalakot az I oszlopban kell megadni. Ha szükséges, a tő és a morfológiai elemzés kézi javítását is meg kell tenni.

[3] `https://e-magyar.hu/hu/textmodules/emmorph_codelist`

[4] `https://emmorph.herokuapp.com/dstem/`

2.4. Tokenizálási hibák javítása

Az `e-magyar` tokenizáló modulja, az emToken[5] is eredményezhet hibás kimenetet, így a hibás tokenizálást is javítani kell. Előfordulhat, hogy helytelenül jelöli ki a mondathatárokat, illetve a tokenek határait.

2.4.1. Mondatok összetokenizálása

A mondatok közötti határokat egy üres sor jelzi. Ha a tokenizáló egy mondatot tévesen két mondatra bontott szét, akkor az annotátornak jelölnie kell a hibát. A mondathatárt jelző fölösleges üres sor H oszlopába bele kell írni: *mondat össze*!

2.4.2. Mondat széttokenizálása

Ha a tokenizáló két mondatot tévesen egy mondatként tokenizált, akkor egy üres sort kell beilleszteni a megfelelő helyre. Ennek a beillesztett üres sornak a H oszlopába bele kell írni: *mondat szét*!

2.4.3. Token összetokenizálása

Ha egy szó tévesen két vagy több tokennek van jelölve, akkor az összetokenizálást a következőképpen kell megtenni: az összetokenizálandó szó összes sorába a H oszlopba bele kell írni, hogy *token össze*, az összetokenizálandó szó első sorába az I oszlopba kell beírni a megfelelő szóalakot, az F oszlopba a hozzá tartozó helyes morfológiai címkét, a G oszlopba pedig a szótövet. Az összetokenizált token az eredeti sorba fog esni, tehát előfordulhat, hogy az emMorph által eredményezett címke helyes, ekkor maradhat az X az F oszlopban.

Összetokenizálásra akkor is szükség lehet, ha maga az elemzetlen bemenet elírást vagy helyesírási hibát tartalmazott és emiatt volt két tokennek jelölve egy szó.

2.4.4. Token széttokenizálása

Ha két vagy több token tévesen egy tokennek van jelölve, akkor a széttokenizálást a következőképpen kell megtenni: annyi új sort kell beszúrni a

[5] `https://e-magyar.hu/hu/textmodules/emtoken`

hibás token sora alá, ahány plusz tokenre szét kell tokenizálni azt. A beszúrt új sorokba a H oszlopba bele kell írni, hogy *token szét*, az I oszlopba kell beírni a megfelelően szétválasztott tokeneket, az F oszlopba az azokhoz tartozó helyes morfológiai címkét, a G oszlopba pedig a szótövet. A szétválasztott token első része az eredeti sorba fog esni, tehát előfordulhat, hogy az emMorph által eredményezett címke helyes, ekkor maradhat az X az F oszlopban.

Széttokenizálásra akkor is szükség lehet, ha maga az elemzetlen bemenet elírást vagy helyesírási hibát tartalmazott és emiatt volt egy tokennek jelölve két szó.

2.5. Token törlése vagy beszúrása

Hibásan fölösleges tokent törölni úgy lehet, hogy a H oszlopba beírjuk: *token töröl*. Hiányzó tokent beszúrni egy új sor beillesztésével lehet, ahol az F oszlopba kell írni a megfelelő címkét, az I oszlopba kell beírni a megfelelő szóalakot, az F oszlopba a hozzá tartozó helyes morfológiai címkét, a G oszlopba pedig a szótövet.

2.6. Egyéb megjegyzések

Bármilyen felmerülő kérdést vagy megjegyzést a J oszlopba kell írni.