Annotálási útmutató

Tokenizálás, morfológia

1. A feladat

Az annotáció első fázisában a tokenizálást és a morfológiai elemzést végezzük el. A tokenizálás a szöveg elemzendő egységekre (tokenekre) bontását jelenti, ebben a lépésben tehát a mondat- és szóhatárokat állapítjuk meg. A morfológiai elemzés során az egyes szóalakok elemeit (morfémáit) és a hozzájuk tartozó morfoszintaktikai jegyeket adjuk meg.
Az annotálást egy erre a célra készített felületen végezzük, ahova a szövegfájlokat már előfeldolgozott formában töltjük be.

2. A felület

A felület megnyitásához Java környezet szükséges, ami letölthető innen:

https://www.java.com/en/download/

(Letöltés után telepítsük!)

Fontos, hogy a DocBuildert olyan mappába töltsük le, aminek a nevében nincs ékezet vagy szóköz! A DocBuilder2.jar indítása után egy böngésző ablakban nyílik meg a felület. (Ha a program nem nyitja meg magától a böngészőt, akkor nyissuk meg külön, és írjuk a címsorba: http://localhost:8000/ vagy: http://127.0.0.1:8000 Ez tipikusan Linux és Mac rendszereknél fordul elő.)

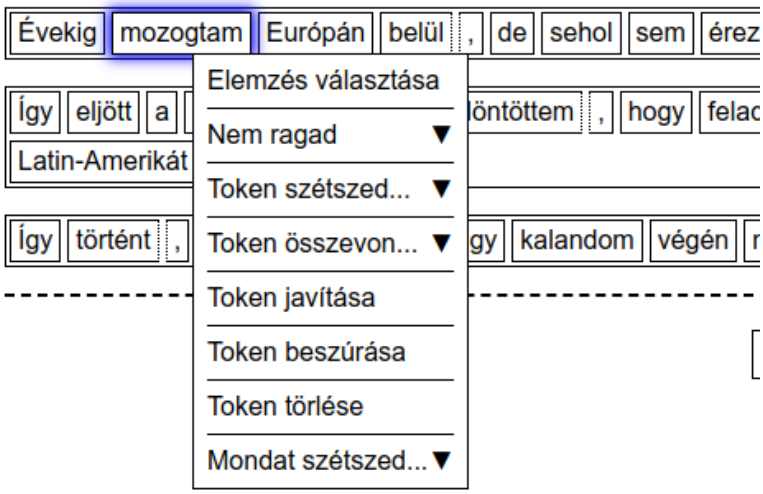
A “Megnyit” opciónál válasszuk ki az annotálandó xml fájlt!

A felső menüsorban lehet váltani a táblázatos és a normál (szöveges) nézet között. A táblázatos nézet célja egyrészt az áttekintés, másrészt a gyorsmentés. Az annotálást érdemes mindig a táblázatos nézetben kezdeni. A táblázatban az e-magyar által javasolt elemzések jelennek meg. Ezeket a sor végén a pipára kattintva tudjuk egyszerűen elmenteni, ha helyesek. A könnyű szavak megtalálását segíti az is, hogy kékkel vannak színezve azok a szavak, amikhez az e-magyar csak egyféle elemzést rendelt. (Megjegyzés ehhez: ritkán, de előfordul, hogy az e-magyar csak egy elemzést ad, de az nem jó. Ne mentsd automatikusan a kék szavakat!)

A könnyű szavak gyorsmentése után érdemes visszaváltani normál szöveges módba, és ott megnézni a nehezebb szavakat. A táblázatos nézetben elmentett szavak a normál nézetben is fehérek lesznek, és fordítva.

Az annotálás befejezésével a táblázatos nézetben lehet egyszerűen átnézni a munkátokat.

A szavakra kattintva a következő funkciók érhetők el:



- Elemzés választása

  A felugró ablakban megjelennek az e-magyar által javasolt elemzések, ezek közül a pipára kattintva (azt zöldítve) kell kiválasztani a helyeset. Ha nincs helyes elemzés, akkor az utolsó üres sorban kézzel kell azt megadni. Választásunkat az “Elment” gombra kattintva mentsük el! Az “új elemzés” gombra akkor van szükség, ha nincsenek a szóhoz javasolt elemzések (mert javítottuk a tokenizálást), ekkor a gombra kattintva elküldhetjük az újonnan keletkezett tokent az e-magyarnak elemzésre. (Ehhez a funkcióhoz internet kapcsolat szükséges!)

- Ragadás

  Ragadó tokenek (pl. írásjelek) esetén itt kell megadni, hogy melyik szomszédos szóhoz ragad a token. (Azaz eredetileg melyik szóval volt egybeírva, pl. írásjel előtti szó jobbra ragad, maga az írásjel balra.)

- Token szétszedése

  Helytelen egybeírás esetén szétszedhetjük a tokent két tokenre.

- Token összevonása

  Helytelen különírás esetén összevonhatjuk az egybe tartozó tokeneket.

- Token javítása

  A hibás tokeneket itt lehet javítani. Irodalmi szöveg esetén csak a nyilvánvalóan sajtóhibákat javítjuk, a többi szövegtípusnál minden helyesírási és gépelési hibát, kivéve, ha az - vélhetően - szándékos (pl. stílusparódia). Nem javítjuk a beszélt nyelvi alakokat (pl. *asszem, nemtom,* stb.). Ezeknél a részletes elemzésben igyekszünk megadni a normalizált alakot. (l. “Nem sztenderd szóalakok részletes elemzése”)

- Token beszúrása, törlése

Tokent beszúrni az aktuális token elé és után is lehet.

A tokenek módosítása (szétszedés, összevonás, javítás) vagy új token beszúrása esetén a program automatikusan lekéri az új elemzést az e-magyartól. A manuális új elemzésre akkor van szükség, ha ez az internetkapcsolat hiánya miatt elmaradt.

- Mondat szétszedése / Mondat összevonása (a doboz sarkánál található fogaskerékre kattintva)

  Egy dobozba csak egy mondat tartozhat. Ha a mondatokra szedés hibás, azt is javítani kell.

3. Tokenizálás

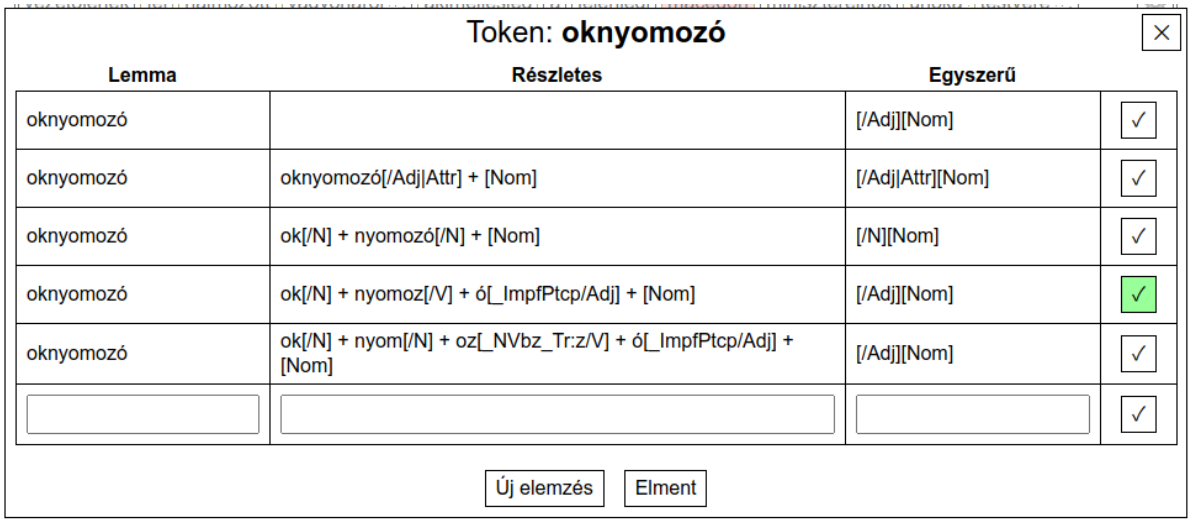
Külön tokennek számít:

- mondatvégi és mondatközi írásjel
- idézőjel
- zárójel, kivéve szón belül, pl. alak(ok)
- -e kérdő partikula (kötőjellel egyben maradva)
- Gondolatjel, és minden egyéb írásjel, ami eredetileg sem tapadt a szomszédos szavakhoz

Nem számít külön tokennek:

- emotikont alkotó írásjelek
- rövidítések, dátumok pontjai
- koordinált elő- vagy utótagot jelző kötőjel, pl. *elő- vagy utótag*
- szóközi kötőjel, perjel, pl. *Budapest-Szeged*, *és/vagy*
- összetett írásjelek, pl. *?!, ...*

4. Morfológiai elemzés



A morfológiai elemzésnél minden szóhoz 3 kitöltendő mező tartozik:

- Lemma (szótő)

  Itt nem megyünk vissza az “abszolút” tőig, azaz nem választjuk le a képzőket. A fő szempont, amit tartsunk szem előtt az, hogy a szótőnek és az egyszerű elemzésnek (POS-tag) egyértelműen meg kell határoznia a szóalakot.

- Részletes

  Itt adjuk meg a szóelemeket és a hozzájuk tartozó morfológiai jegyeket. A részletes elemzésnél a fő kérdés az, hogy meddig elemezzünk egy-egy szót. Az általános elv az, hogy addig, ameddig a szó jelentése szerint ennek értelme van. A fenti *oknyomozó* példában az ok[/N] + nyomoz[/V] + ó[_ImpfPtcp/Adj] + [Nom] a jó választás, mert az *ok+nyomozó* összetétel transzparens, továbbá a *nyomozó* ebben az esetben egy igéből képzett melléknév. A *nyomoz* igét azonban már nem indokolt visszavezetni a *nyom*-ra, mert itt már jelentésváltozás történt, a *nyomoz*-t már nem tekintjük a *nyom*-ból képzett igének (bár etimológiailag nyilván köze van hozzá).
  Az írásjelekhez és idegen szavakhoz nem tartozik részletes elemzés. Más esetben azonban nem lehet üres ez a mező! Az e-magyar néha ad olyan kimeneteket, ahol a részletes elemzés nincs megadva (l. oknyomozó példa első sorát), ezt soha ne válasszuk, ha nincs más lehetőség, kézzel kell megadni a részletes elemzést.

  Az allomorfokat az eredeti morfémával együtt adjuk meg tő[/szófaj]=tőallomorf alakban:

A zéró allomorfokat üres szögletes zárójellel ([]) jelöljük.

- Egyszerű

  Az egyszerű mezőbe a (végső) szófajcímke és a szó morfológiai jegyei kerülnek, de itt már egy szóalakként kezeljük a szót, nem tüntetjük fel az elemeit.

Az e-magyar kódlistája leírással és példákkal itt található:

https://e-magyar.hu/hu/textmodules/emmorph_codelist

A kódlistából hiányzik a -ka/-ke kicsinyítő képző kódja. Ezt a -cska/-cske mintájára adjuk meg kézzel: [_Dim:kA/Adj]

Szintén hiányzik az -l denominális igeképző, ennek kódja a -z denominális igeképző mintájára:

- [_NVbl_Tr:/V] -- ha tranzitív (pl. update-el vmit)
- [_NVbl_Ntr:l/V] -- ha intranzitív (pl. blogol)

5. Általános kérdések

- **Nem elemzendő szavak**

  A nem elemzendő tokeneknél a lemma és az egyszerű elemzés mezőbe `[None]` címke kerül, a részletes elemzés üresen marad.

  Nem elemzendő:

  - címek szavai, kivéve az utolsót (l. “Tulajdonnevek elemzése”)
  - felismerhetetlen vagy elemezhetetlen (pl. nyelvjárási) szóalak, pl.: ***Osztig*** *mög akartam mutatni a tőkis világnak*
  - felsorolást jelző betűk, pl. *a), b), c),* stb.

- **Képzett szavak részletes elemzése**

  A kódlistában szereplő képzőket általában jelöljük, kivéve ha a szóalak jelentése eltér a szóelemek kompozicionális jelentésétől, vagy annál absztraktabb pl.:

  házasság -- nem *házas* + *ság* (kivéve, ha valóban a *házas* tulajdonsággal való rendelkezést jelenti)
  tudás -- nem *tud* + *ás* (kivéve, ha nem általános jelentésben használjuk, hanem konkrétan valaminek a tudásáról beszélünk)

  Akkor sem elemezzük külön az -ó/-ő melléknévképzőt, ha a szó főnévként szerepel, pl. *olvasó, fürdő(hely)*

- **Összetett szavak részletes elemzése**

  Az összetett szavak elemeit mindig jelöljük, ha a kapcsolat köztük látható, és levezethető belőle a szó jelentése: pl. *kávéfőző, kertajtó, kutyavásár*

  Egy szóalaknak tekintjük azokat az összetételeket, ahol az összetett szó jelentése eltér az elemekből nyelvtanilag értelmezhető szó szerinti jelentéstől: pl. *lóhalálában, tagbaszakadt*

- **Nem sztenderd szóalakok részletes elemzése**

  Az eredeti tokent ne javítsuk ki, de szótőnek a normalizált alakot adjuk meg, és ennek megfelelően elemezzünk!

  Pl.: (Egri csillagok 1. fejezet)

  - Derdő - mondja a fiúnak - , **uttyunk** .

- **Nem produktív képzők**

  A nem produktív képzőket nem elemezzük, az ilyenekkel képzett szavakat tőszavaknak tekintjük. (Pl. a *futkos* alakot már nem elemezzük tovább a *fut* tőig, mert nem produktív képzővel létrehozott alak.)

- **Alkalmi szófajváltás**

  Az úgynevezett alkalmi szófajváltást nem jelöljük a morfológiai annotációban. Ez a szintaxis körébe tartozik.

  Pl.:
  *Abban az időben nem volt ritkaság a **török** az utakon.*
  → A *török* elemzése [/Adj|nat][Nom], függetlenül attól, hogy itt alanyi szerepben áll.

  Ez alól kivételt képez az az eset, amikor metanyelvi funkcióban használunk egy egyébként nem főnévi szóalakot (pl. *megálljt parancsol)*, ebben az esetben főnévnek címkézzük azt.

  Nem tartoznak ide a már lexikálisan is főnevesült képzett szavak, mint pl. a foglalkozásnevek (*kutató*).

- **Tulajdonnevek**

  A tulajdonnevek lemmája mindig nagybetűs.

  - **összetett földrajzi nevek**

    Jelöljük az összetételi tagokat, ha:
    - az egyik elem köznév, pl.: *Földközi-tenger, János-hegy*
    - mindkét elem értelmezhető önállóan, pl.: *Dél-Amerika*

    Nem kezeljük összetett szóként, ha nincs köznévi tagja, és az elemek nem értelmezhetők önállóan, pl.: *Új-Zéland*

  - **címek**

    A címeket speciális főnévi csoportoknak tekintjük, ezért a címek utolsó szavát minden esetben főnévi ([/N]) címkével jelöljük (hozzátéve az esetleges esetragot). A cím többi szava [None], azaz nem elemzett címkét kap.

    Pl.: *Nézd[None] ,[None] ki[None] van[None] itt[None] ![/N] + [Nom]*
    Jojo[None] + nyuszi[/N] + ban[Ine]

- **Idegen szavak**

  A hosszabb idegen nyelvű szövegrészletek szavai nem kapnak részletes elemzést, az egyszerű elemzésük [/X].

  Pl.: *Ordnung[/X] muss[/X] sein[/X]*

A magyar szövegbe illeszkedő idegen szavaknál viszont lehetőség szerint megadjuk a szófajt, és az elemzést is (ha magyarul van ragozva). Az idegen tulajdonneveket is a magyar tulajdonnevekhez hasonlóan elemezzük.

Pl.: *taco[/N]=tacó + s[_Adjz:s/Adj] + [Nom]*

Az idegen nyelvű kifejezések rövidítésére az [/X|Abbr] címkét használjuk.

Pl.: *Mr. [/X|Abbr] Smith [/N]*

A kötőjellel illesztett ragoknál tőallomorfnak elemezzük a kötőjeles tövet:

House-ba → `House[/N]=House- + ba[Ill]`

- **Számok**

A számok számjegyeit (az e-magyarral ellentétben) nem elemezzük külön tagokként.

1[/Num|Digit] + 2[/Num|Digit] + 3[/Num|Digit] helyett:

`123[/Num|Digit]` a helyes elemzés

Ez igaz a törtszámokra is, egy tokenként, [/Num|Digit] címkével elemezzük őket:

`3,6[/Num|Digit]`

- **Elváló igekötők**

Ha az igekötő elválik az igéjétől, az igét, Pethő és mtsai (2022) javaslata alapján igekötős igeként elemezzük. Az igekötő megjelenik a lemmában és, zéró morféma formájában, a részletes elemzésben is:
*20 perc alatt **értünk** oda*

értünk

odaér	oda[/Prev]=[] + ér[/V] + tünk[Pst.NDef.1Pl]	[/V][Pst.NDef.1Pl]	✓

6. Specifikus kérdések (Gy. I. K.)

- **Létigék és kopulák**

A létigék és kopulák töve *van* vagy *lesz*, az alakjuktól függően:

van-, vol-, vagy-* → *van
le-, lesz-* → *lesz

A ***lehet*** szóalak lehetséges elemzései:

- **lesz[/V]=le + het[_Mod/V] + [Prs.NDef.3Sg]**
  ha létige vagy kopula

- **lehet[/V] + [Prs.NDef.3Sg]**
  ha modális segédige

A ***volna*** szóalak elemzése:

- **van[/V]=vol + na[Cond.NDef.3Sg]**
  minden esetben

A ***levő/lévő*** elemzése:

lesz[/V]=lev+ő[_ImpfPtcp/Adj]=ő+[Nom]

A ***való*** elemzése:

van[/V]=val+ó[_ImpfPtcp/Adj]=ó+[Nom]

- **[Det|Pro] vs. [N|Pro]**

Az *ez/az* mutató névmásokat [Det|Pro]-nak címkézzük a következő szerkezetekben:

***ez** a hely, **az** az ablak*

Minden egyéb esetben az *ez/az* mutató névmások címkéje [N|Pro].
A *valami* névmást [Det|Pro]-nak címkézzük, ha közvetlenül főnév előtt áll, pl.:

*mint **valami** máglya*

- **Vonatkozó névmások**

Az ***akkor*** szóalak lehetséges elemzései:

- **akkor[/Cnj]**
  feltételes mondatban, ha nem időre vonatkozik:
  *Ha felkészülsz, **akkor** nem lesz gond.*

- **az[/N|Pro]=ak + kor[Temp]**
  ha főnévre utal:
  *Szeretem a karácsonyt, mert **akkor** együtt van a család.*

- **akkor[/Adv]**
  ha partikula:

***Akkor** ideje indulni!*

- **akkor[/Adv|Pro]**
  egyébként (ha vonatkozó névmás)

Az ***amikor*** szóalak lehetséges elemzései:

- **ami[/N|Pro|Rel] + kor[Temp]**
  ha főnévre utal:
  *Ez volt az a nap, **amikor**...*

- **amikor[/Adv|Pro|Rel]**
  egyébként (ha vonatkozó névmás)

A többi vonatkozó névmásnál is ezek mintájára járunk el. Némely ragozott alaknál lehetséges még a [/Det|Pro]-s elemzés, pl.:

***attól** a naptól kezdve...*
az[/Det|Pro]=at + tól[Abl]

Pethő Gergely, Sass Bálint, Kalivoda Ágnes, Simon László, Lipp Veronika. Igekötő-kapcsolás. In: Berend Gábor, Gosztolya Gábor, Vincze Veronika (szerk.): *MSZNY 2022, XVIII. Magyar Számítógépes Nyelvészeti Konferencia,* Szegedi Tudományegyetem, Informatikai Intézet, Szeged, 77-91