# 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:

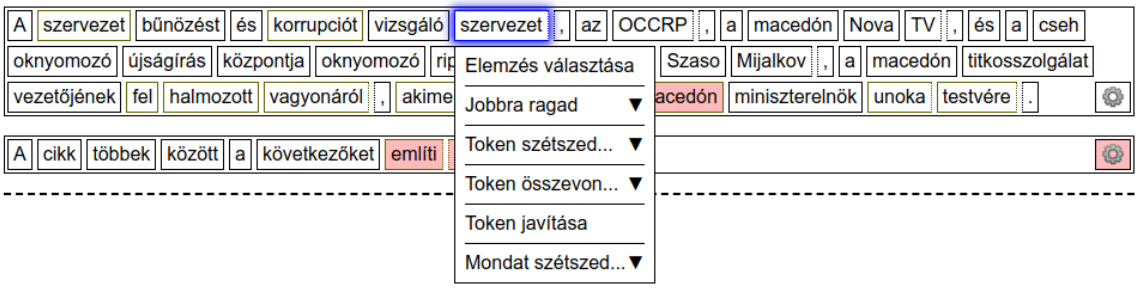
https://www.java.com/en/download/

(Letöltés után telepítsük!)

Fontos, hogy a DocBuilder.zip-et olyan mappába töltsük le (és csomagoljuk ki), aminek a nevében nincs ékezet vagy szóköz! A DocBuilder.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/)

A “Megnyit” opciónál válasszuk ki az annotálandó xml fájlt! (Gyakorló példa: level3.xml)

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. (Az írásjel előtti szó jobbra ragad, maga az írásjel balra, a többi szó nem ragad.)

- 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.

  A tokenek módosítása (szétszedés, összevonás, javítás) esetén mindig kérjünk új elemzést!

- 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**

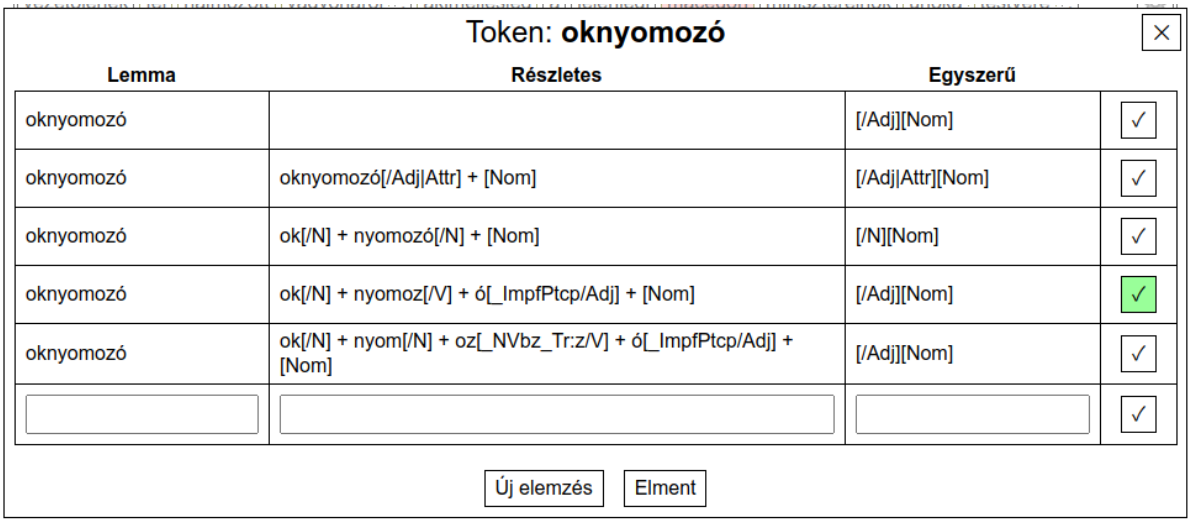
Külön tokennek számít:

- mondatvégi és mondatközi írásjel
- idézőjel
- zárójel, kivéve szón belül, pl. alak(ok)
- -e kérdő partikula (kötőjellel egyben maradva)

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*

**4. Morfológiai elemzés**




A morfológiai elemzésnél minden szóhoz 3 kitöltendő mező tartozik:

- Lemma (szótő)

- 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 nem tartozik részletes elemzés.

- 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/en/textmodules/emmorph_codelist