Annotációs szintek

Számítógépes nyelvészet – 2018 tavasz
8. óra

Simon Eszter – Mittelholcz Iván

MTA Nyelvtudományi Intézet

Tartalom

Tokenizálás

Mondatszegmentálás

- Minden mondat.
- Mondathatárok azonosítása.
- Pontos problémák:
  - Rövidítések (*du. 5-kor*).
  - Római számok (*V. László*).
  - Sorszámok (*10. éve, hogy …*).
- Egyéb nehézségek:
  - Idézeten belüli mondatok.
  - Zárójelen belüli mondatok.

Tokenizálás

- Detokenizálhatóság és elválasztás (és az *-e* paritkula).
- Szóalkotó karakterek, szónemalkotó karakterek, és amik köztük vannak:
  - Zárójelek, idézőjelek, aposztrófok kezelése.
  - Rövidítések végén lévő pont vs. mondatvégi pont.
- Számok (space-szel tagolt számok, mértékegységek, képletek, dátumok).
- Informatikai kifejezések (URL, elérési út, emailcím).
- Smiley-k és emoji-k.

Irodalom I

Mittelholcz Iván (2017): *emToken: Unicode – képes tokenizáló magyar nyelvre.* In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). Szeged (2017)