Nyelvi feldolgozó eszközök

Letölthető és/vagy beépíthető eszközök

e-magyar/emtsv

- https://github.com/dlt-rilmta/emtsv
- moduláris, bármelyik lépésnél ki-be lehet szállni, új modulokat könnyen lehet integrálni
- lassabb: dependenciáig REST API-val: 310 token/s
- python, docker
- elemzési szintek:
    - tokenizálás, mondatrabontás
    - morfológiai elemzés és egyértelműsítés
    - szintaktikai elemzés
        - függőségi elemzés
        - összetevős elemzés
    - tulajdonnév-felismerés
    - főnévi frázisok felismerése
    - hunspell
    - udpipe
    - szótáralapú kifejezésfelismerő
    - zérónévmás-beszúró
    - konverterek
    - kiértékelő
- nyílt forráskódú, szabadon felhasználható
- SOTA teljesítmény

stanza

- https://stanfordnlp.github.io/stanza/
- elemzési szintek:
    - tokenizálás, mondatrabontás
    - morfológiai elemzés és egyértelműsítés
    - szintaktikai elemzés
        - függőségi elemzés
    - tulajdonnév-felismerés
- nyílt forráskódú, szabadon használható
- kevésbé jó teljesítmény

udpipe

- http://ufal.mff.cuni.cz/udpipe
- elemzési szintek:
    - tokenizálás, mondatrabontás

        - morfológiai elemzés és egyértelműsítés
        - szintaktikai elemzés
            - függőségi elemzés
- még kevésbé jó teljesítmény
- még gyorsabb: dependenciáig: 3.300 token/s
- az egyes lépéseknél ki-be lehet szállni az egységes formátumnak köszönhetően, de új modulokat nem lehet integrálni
- nyílt forráskódú, szabadon használható

huspacy

- https://github.com/oroszgy/spacy-hungarian-models/
- elemzési szintek:
    - tokenizálás, mondatrabontás
    - morfológiai elemzés és egyértelműsítés
    - szintaktikai elemzés
        - függőségi elemzés
    - tulajdonnév-felismerés
    - szóvektorok
    - Brown-klaszterek
    - tokengyakoriságok
- a leggyorsabb: dependenciáig: 15.000 token/s
- nyílt forráskódú, szabadon használható
- python
- kevésbé jó teljesítmény
- bármelyik lépésnél ki-be lehet szállni, új modulokat könnyen lehet integrálni

magyarlánc

- https://rgai.sed.hu/node/100
- nem moduláris, csak egyben lehet futtatni az elejétől a végéig, új modulokat nem lehet integrálni
- gyorsabb: dependenciáig: 450
- java
- elemzési szintek:
    - tokenizálás, mondatrabontás
    - morfológiai elemzés és egyértelműsítés
    - szintaktikai elemzés
        - függőségi elemzés
        - összetevős elemzés
- letölthető, szabadon felhasználható
- SOTA teljesítmény

Futottak még

WebLicht

- https://weblicht.sfs.uni-tuebingen.de/weblicht/
- standard nyelvi feldolgozást végző eszközlánc grafikus felülettel
- magyarra a UDPipe van alácsövezve
- table view & visualization (mint az e-magyarnál: táblázat és ágrajz)
- van benne egy conll2tcf konverter: a TCF valami saját XML-like formátum, vagyis nekik is meg kellett oldaniuk a tsv2xml konverziót
- csak intézményi autentikáció után lehet belépni

BookNLP

- https://github.com/dbamman/book-nlp
- sztenderd nyelvfeldolgozó eszközlánc
- azért BookNLP, mert könyv hosszúságú szövegek feldolgozására fejlesztették
- csak angolra van