Korpuszannotáció, annotációs szintek

A számítógépes nyelvészet alapjai – 2022/23 tavasz
5. óra

Simon Eszter

2023. március 27.

Tartalom

Korpuszannotáció

A nyelvi annotáció

a sztenderd szövegfeldolgozó lépések a modern korpuszoknál nagyjából ugyanazok:

- szegmentálás (tokenizálás, mondatra bontás)
- morfológiai elemzés
- morfoszintaktikai egyértelműsítés

Mi kell az annotációhoz?

- annotációs séma
  - elméleti nyelvészeti alapok lefektetése (pl. mi a tulajdonnév?)
  - címkekészlet
  - az annotáció formátuma (inline vagy standoff)
- annotációs eszköz
- az annotátorok száma $\rightarrow$ annotátorok közötti egyetértés mérése
- annotációs útmutató
- az annotáció minőségének ellenőrzése

Annotációs útmutató

- az útmutatónak egyszerre kell kellően kidolgozottnak és egyszerűnek lennie, hogy az annotátorok számára követhető legyen → ha nem így van, akkor az annotátorok magas hibaszázalékkal fognak dolgozni
- tartalmaznia kell az annotációs feladat leírását, az annotálandó nyelvi elemek felsorolását és példákat arra, hogy mit kell és mit nem kell annotálni
- minél magasabb nyelvi szintre megyünk, minél több szemantika van benne, annál képlékenyebb a feladat → bizonyos nyelvi jelenségek nehezen megfoghatók/formalizálhatók
- ha az útmutató nem elég egzakt, akkor az annotátorok elkezdik követni az intuíciójukat → a nem teljesen egyértelmű esetekben ez problémákat okozhat

NER annotációs útmutatók

- MUC-7 Named Entity Task Definition (Chinchor, 1997)
- ACE (Automatic Content Extraction) English Annotation Guidelines for Entities (Linguistic Data Consortium, 2008)
- Hunner project proposal és útmutató
- NYTK-NerKor útmutatók

Az annotáció formátuma

inline (XML)

```
<s><w>Ez</w> <w>egy</w> <w>mondat</w> <c> .</c>
<s><w>Meg</w> <w>a</w> <w>második</w> <c> .</c>
```

standoff

Ez
egy
mondat
.

Meg
a
második
.

XML

EXtensible Markup Language
egyfajta jelölőnyelv (markup language) → vannak más hasonlók:
YAML, JSON, MD

Előnyei:

- mind ember, mind gép számára olvasható formátum
- támogatja a Unicode-ot
- szabványos és platformfüggetlen
- képes a legtöbb általános számítástudományi adatstruktúra ábrázolására

Hátrányai:

- szintaxisa elég bőbeszédű és részben redundáns
- nagyobb tárolási költség
- nincs lehetőség a dokumentum egyes részeinek közvetlen elérésére
- átfedő adatstruktúrák modellezése nehéz/lehetetlen

Standoff annotáció

- az eredeti dokumentumok sima szöveg fájlok maradnak
- az annotációk nem szövegközi tagek, hanem egy külső fájlban jelöljük úgy, hogy megadjuk, hogy az eredeti szöveg melyik karaktertartományára vonatkozik a címkézés, és hogy milyen címkét kap a szövegrészlet
- az annotálást teljesen különválasztjuk a használt feldolgozó eszközöktől, és közben minden információt megtartunk
- az átfedő és beágyazott annotáció is könnyen kezelhető

Beágyazott és átfedő annotáció

Beágyazott annotáció

<LOC><PERSON>Kossuth Lajos</PERSON>utca</LOC>

Átfedő annotáció

a Kossuth Lajos és a Petőfi Sándor utca sarkán

BIE1

[...] közölte Wolf László, az OTP Bank vezérigazgató-helyettese az MTI érdeklődésére.

közölte	O
Wolf	B-PER
László	E-PER
,	O
az	O
OTP	B-ORG
Bank	E-ORG
vezérigazgató-helyettese	O
az	O
MTI	1-ORG
érdeklődésére	O
.	O

BIE1

A szállásunk egy Balaton melletti kis üdülőfaluban, Zamárdiban volt.

A	B-NP
szállásunk	E-NP
egy	B-NP
Balaton	I-NP
melletti	I-NP
kis	I-NP
üdülőfaluban	I-NP
,	I-NP
Zamárdiban	E-NP
volt	O
.	O

A kézi annotáció minősége

Az annotáció minősége

- a kézzel annotált korpuszokat tanító- vagy kiértékelőanyagként használják felügyelt gépi tanulással működő eszközök számára
- felügyelt gépi tanulással működő rendszerek sikeressége a tanítóanyag minőségén múlik
- csak olyan feladatokat lehet felügyelt gépi tanulással megoldani, amelyeket az ember is képes elvégezni
- csak olyan nyelvi jelenségekhez tudunk kézi annotációt készíteni, amelyeket eléggé megértettünk ahhoz, hogy pontosan le tudjuk írni őket
- megbízható az annotáció, ha a jelenségek leírását több annotátor is hasonlóképpen megértette és ez alapján hasonlóképpen kódolják az egyes jelenségeket
- a feladatleírásnak tehát érthetőnek kell lennie az annotátorok számára, akik ideális esetben egyetértenek az egyes jelenségek címkézésében

Az annotátorok közötti egyetértés

- a cél a minél magasabb annotátorok közötti egyetértés
- minél egyszerűbben leírható nyelvi jelenség annotálásáról van szó, annál könnyebb magas annotátorok közötti egyetértést elérni, a nyelvi jelenség összetettségével az egyetértés mértéke is könnyen csökken
- mitől lehet alacsony?
  - a feladat megfogalmazása nem egyértelmű vagy nem teljes
  - az annotátoroknak túl sok kategóriát kell kezelniük
  - átláthatatlan felületen kell dolgozniuk

Az annotátorok közötti egyetértés

- az annotátorok (vagy kódolók), amikor kategóriákat rendelnek egyes elemekhez, szubjektív döntéseket hoznak
- ha az annotátorok egyetértenek az egyes elemekhez rendelt kategóriákban, akkor az adat megbízható, és ha a kódolók következetesen hasonló eredményt produkálnak, akkor hasonlóképpen értették meg a feladatot és az annotálási útmutatót, ezért a továbbiakban is hasonló eredményeket várhatunk tőlük
- *megfigyelt egyetértés*: azt mutatja meg, hogy az esetek hány százalékában értett egyet a két kódoló
- DE! nem elég, ha két kódoló egyetért, hiszen mindketten tévedhetnek is
- a címkék számának csökkentésével növekszik a megfigyelt egyetértés, ráadásul nem érzékeny az egyes címkék eltérő gyakoriságára
- megoldás: valószínűség-korrigált együtthatók, amelyek számolnak a véletlen eseményekkel is

Az annotátorok közötti egyetértés: koefficiensek

Különböző mérőszámok az egyetértésre:

- megfigyelt egyetértés
- S (Bennett, Alpert és Goldstein 1954): minden kategória ugyanolyan valószínű, a kategóriák között egyenletes eloszlást feltételez
- π (Scott, 1955): kategóriánként eltérő, de kódolók között megegyező eloszlás
- κ (Cohen, 1960): kategóriánként és kódolónként eltérő eloszlás, ez már kezeli az elfogultságot
- α (Krippendorff, 1980): nem csak az egyetértést vizsgálja, hanem az egyet nem értés különböző fokozatait

Az egyetértés erőssége

Landis and Koch (1977)

κ	strength of agreement
<0.00	poor
0.00 – 0.20	slight
0.21 – 0.40	fair
0.41 – 0.60	moderate
0.61 – 0.80	substantial
0.81 – 1.00	almost perfect

Példák

Tulajdonnév-felismerés

hunNERwiki korpusz (Simon and Nemeskey, 2012):

- $\kappa = 0,967$
- F-mérték: 92,94%

Szeged NER korpusz (Szarvas et al., 2006):

- egyetértési arány: 99,6%

Metaforikus kifejezések felismerése (Babarczy et al., 2010)

egyetértési arány:

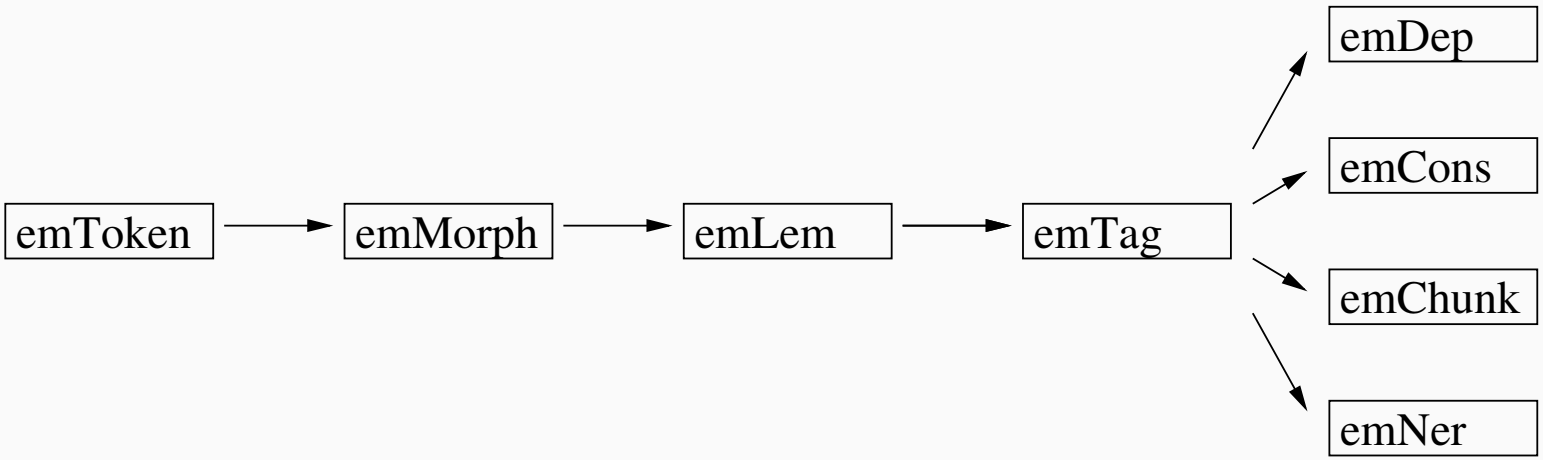
- 1. körben: 17%
- 2. körben: 48%

Szövegfeldolgozási és annotációs szintek

Alapszintű szövegfeldolgozási szintek

- mondatrabontás és tokenizálás
- morfológiai elemzés
- sekély szintaktikai elemzés
- mély szintaktikai elemzés
- tulajdonnév-felismerés
- ...

Egymásra épülő annotációk



Mondatszegmentálás

Mittelholcz (2017)

- 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* partikula).
- 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.

Morfológiai elemzés

tokenszintű elemzés → nem lát se előre, se hátra → no kontextus → többértelműség

kerekesszék

kerek/ADJ+esszé/NOUN<PLUR>

kerekes/ADJ+szék/NOUN

kerék/NOUN[ATTRIB]/ADJ+szék/NOUN

kerek/ADJ[ATTRIB]/ADJ+szék/NOUN

kerék/NOUN[ATTRIB]/ADJ+szék/NOUN

kerek/ADJ[ATTRIB]/ADJ+szék/NOUN

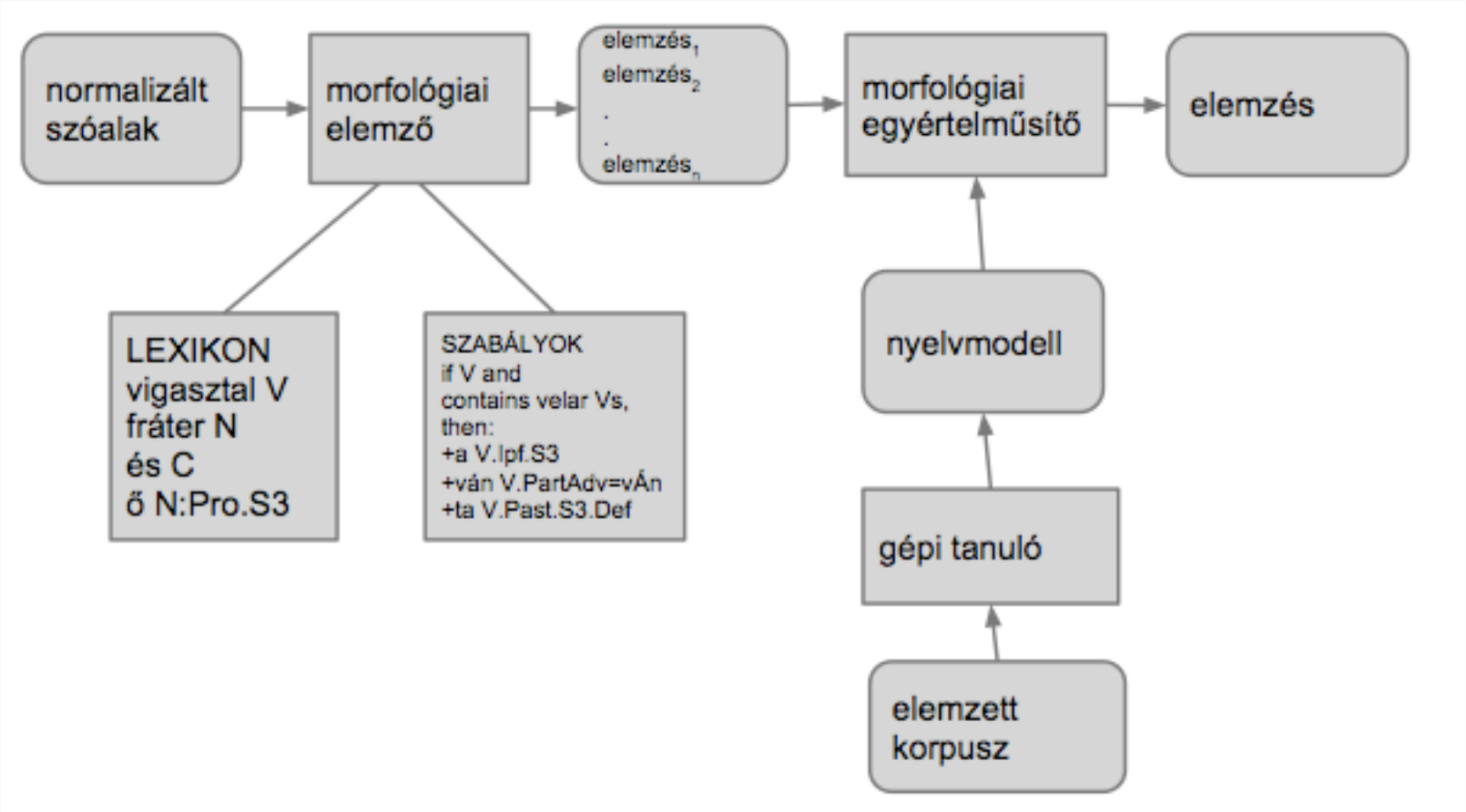
falucska

```
fa [/N] + luc[/N] + ska[/N] + [Nom]
fa[/N] + lucsok[/N]=lucsk + a[Poss.3Sg] + [Nom]
falu[/N] + cska[_Dim:cskA/N] + [Nom]
falucsok[/N]=falucsk + a[Poss.3Sg] + [Nom]
falucska[/N] + [Nom]
```

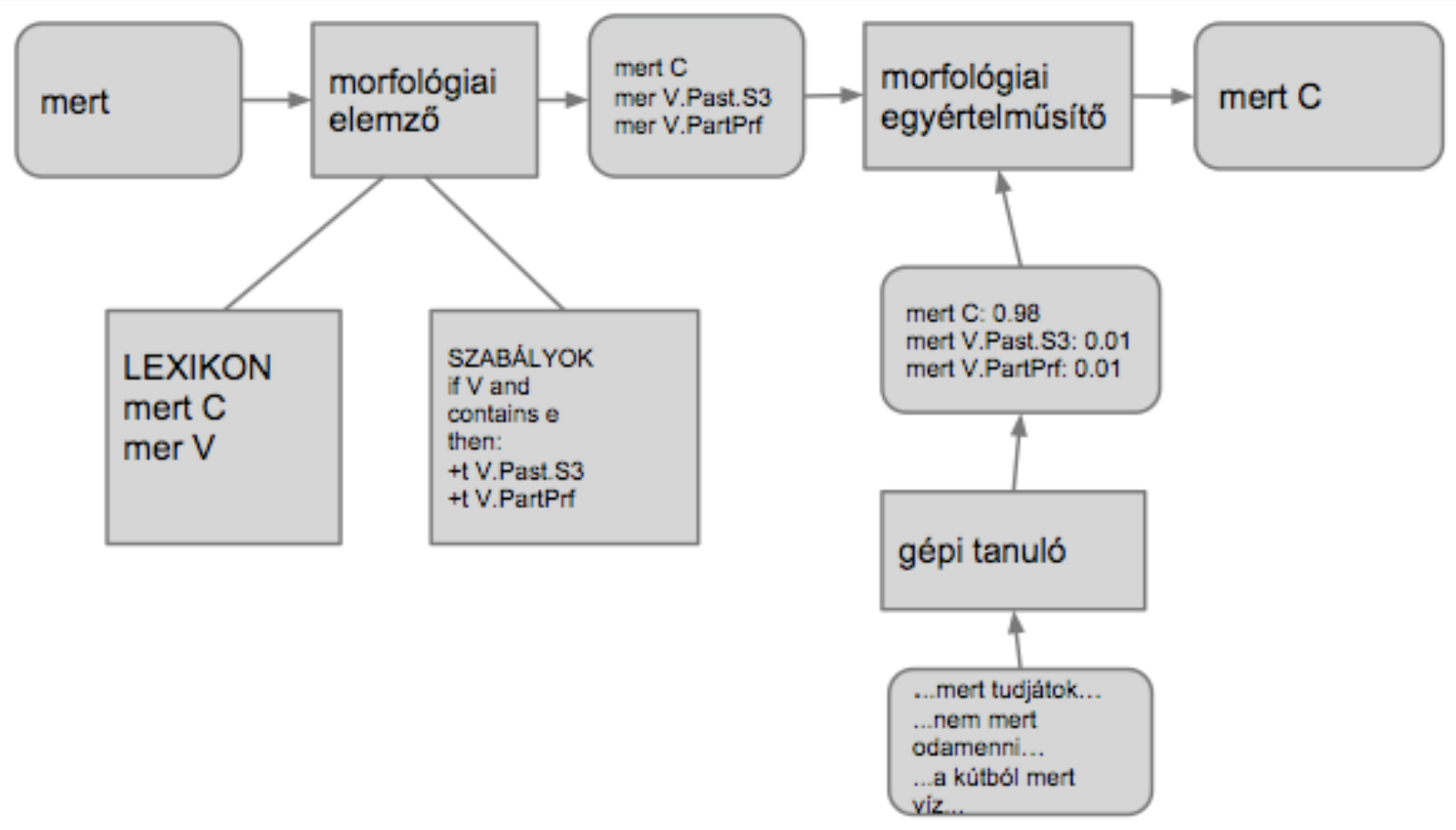
Mit tartalmazhat a kimenet?

- morfoszintaktikai információk
- jelentésre vonatkozó információk
- hangalakra vonatkozó információk (allomorfia)
- szófajkód
- lemma
- morfológiai szegmentumok

Morfológiai egyértelműsítés 1.



Morfológiai egyértelműsítés 2.



Nézzük meg az e-magyart!

Tulajdonnév-felismerés

Named Entity Recognition (NER)

2 lépésből áll:

1. a nevek lokalizálása strukturálatlan szövegben
2. a megtalált elemek besorolása előre definiált névosztályokba
   - *Person, Location, Organization, Date, Time, Money, Percent, Measure* (MUC)
   - *Person, Location, Organization, Miscellaneous* (CoNLL)

Kérdések, nehézségek

- a tulajdonnevek definiálása problémás
- egymásba ágyazott nevek és kompozicionalitás
- van-e a tulajdonnévnek jelentése?
- a tulajdonnevek a szintaxis szempontjából oszthatatlan nyelvi egységek
- nem lehet belülről módosítani őket
- a ragok mindig az NP-t alkotó tulajdonnév végére kerülnek
- a tulajdonnevek alaki sérthetetlenségének elve
- metonimikusan viselkedő tulajdonnevek
- eltérő annotációs sémák $\rightarrow$ még a statisztikai alapú rendszereket is nehéz átvinni egyik korpuszról a másikra, vagy egyik műfajról a másikra

Sekély szintaktikai elemzés

chunking
[Immár] [negyedik éve] [a Manchester United]
[a világ leggazdagabb csapata] [bevétel szerint].

1. minden frázis megtalálása egy mondatban
2. maximális NP-k megtalálása
3. alap NP-k megtalálása

Összetevős és függőségi elemzés

Összetevős elemzés

A mondatok összetevős szerkezeti elemzése azt tárja fel, hogy a szavak egymással kombinálódva milyen kifejezéseket alkotnak, illetve hogyan állnak össze egy mondattá.

Függőségi elemzés

A függőségi elemzés a mondatok szerkezeti egységei közötti függőségi viszonyokat (pl. alany, tárgy, jelző) tárja fel.

Összetevős és függőségi szintaktikai elemző

- kétféle elméleti keret szerint
- függőségi elemzés: Bohnet parser alapján
- összetevős elemzés: Berkeley parser alapján
- tanító adat: Szeged (Dependencia) Treebank
- bemenet: morfológiai egyértelműsítő kimenete
- kimenet: CoNLL formátum (függőségi elemzés), Berkeley kimeneti formátuma

Hogyan működik az elemző?

Irodalom

Hivatkozások

Babarczy, A., Bencze, I., Fekete, I., and Simon, E. (2010). The Automatic Identification of Conceptual Metaphors in Hungarian Texts: A Corpus-based Analysis. In Bel, N., Daille, B., and Vasiljevs, A., editors, *Proceedings of the LREC 2010 Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods*, pages 31–36, Malta.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Mittelholcz, I. (2017). *emToken*: Unicode-képes tokenizáló magyar nyelvre. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, pages 61–69, Szeged.

Irodalom ii

Simon, E. and Nemeskey, D. M. (2012). Automatically generated NE tagged corpora for English and Hungarian. In *Proceedings of the 4th Named Entity Workshop (NEWS) 2012*, pages 38–46, Jeju, Korea. Association for Computational Linguistics.

Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., and Csirik, J. (2006). A highly accurate Named Entity corpus for Hungarian. In *Electronic Proceedings of the 5th International Conference on Language Resources and Evaluation*.