Annotációs kérdések

Nyelvi adatok feldolgozása – 2019/20 tavasz
5. óra

Simon Eszter

MTA Nyelvtudományi Intézet

Tartalom

Annotációk összevetése

Bevezetés

ugyanarra a szövegre vonatkozó két annotáció összevetése:

1. az egyik erősebb → egy automatikus eszköz kimenetének egy gold standard annotációhoz való hasonlítása
2. egyenrangúak → két vagy több annotátor által készített kézi annotáció összehasonlítása

cél: az akár kézzel, akár géppel készült korpuszannotáció minőségének mérése

Azonosan címkézett elemek

szigorúan véve azonosan címkézett elemek azok, amelyeknek

1. ugyanazok a határaik, vagyis
    - ugyanott kezdődnek
    - ugyanott végződnek ÉS
2. ugyanaz a címkéjük

Pozitívok és negatívok

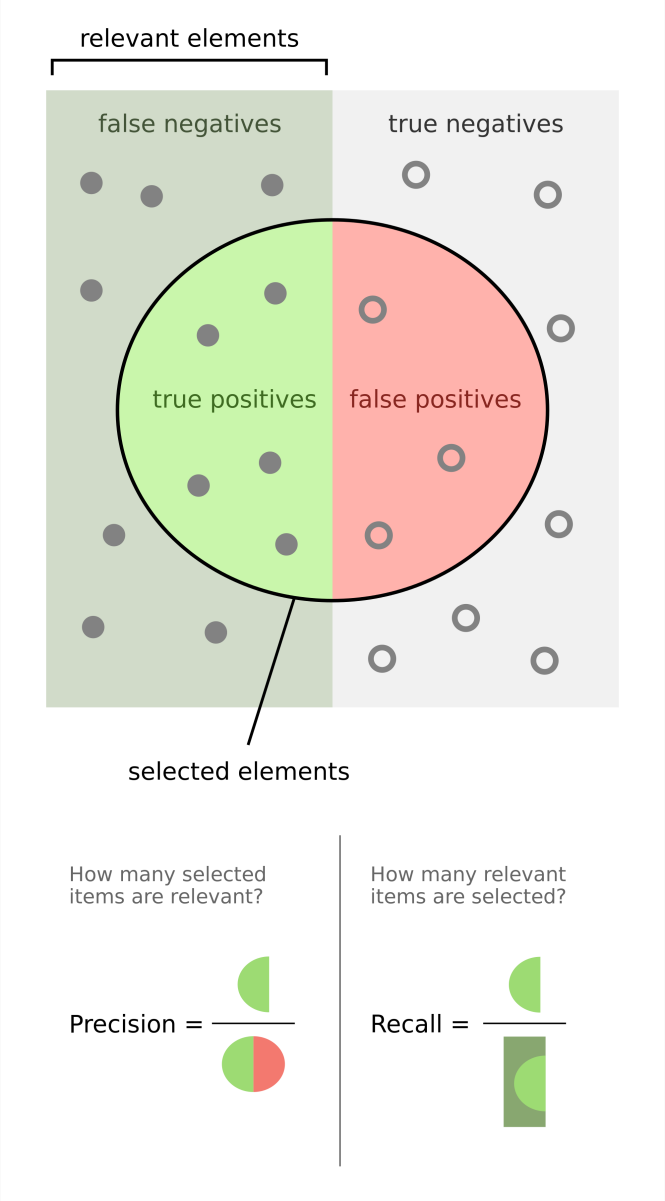
True Positive (TP): a rendszer helyesen felismerte a NE-t;

True Negative (TN): a rendszer helyesen bocsátott ki *O*-t, vagyis helyesen ismerte fel, hogy az adatpont nem NE;

False Positive (FP): a rendszer NE-nek jelölt egy adatpontot, ami nem az;

False Negative (FN): a rendszer nem ismert fel egy NE-t, pedig kellett volna.

Pontosság és fedés



Pontosság, fedés, F-mérték

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision + recall}$$

F-mérték

- a *pontosság* maximalizálása: minél kevesebb tévedés $\rightarrow$ szigorítás
- a *fedés* maximalizálása: minél több találat $\rightarrow$ megengedőbb rendszer

$\beta = 1$

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

Accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Az egyetértés mérési módjai 1.

mindegyik azon alapul, hogy az annotátorok egymástól függetlenül annotálnak (Artstein and Poesio, 2008)

Egyetértési arány (percentage of agreement):

$$\frac{A \cap B}{N}$$

Együttes valószínűség (joint probability of agreement):

$$\frac{2 * (A \cap B)}{A + B}$$

Az egyetértés mérési módjai 2.

a fenti módszerek nem veszik figyelembe, hogy az egyetértés történhet véletlenül is

- Cohen's κ (Cohen, 1960):
  - $\kappa = 1$, ha az annotátorok teljes mértékben egyetértenek
  - $\kappa = 0$, ha az annotátorok a véletlen egybeesésnél nem jobban értenek egyet
- Krippendorff's α (Krippendorff, 1980, 2004):
  - $\alpha = 1$, ha az annotátorok teljes mértékben egyetértenek
  - $\alpha = 0$, ha az elemek és a hozzájuk rendelt értékek között nincs semmi reláció, vagyis teljesen véletlen egybeesésről van szó
  - $\alpha < 0$, ha az egyet nem értés magasabb a véletlen egybeesésnél, vagyis szisztematikus egyet nem értésről van szó

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 és Nemeskey, 2012):

- $\kappa = 0,967$

Szeged NER korpusz
Szarvas et al., 2006:

- egyetértési arány: 99,6%

Metaforikus kifejezések felismerése

(Babarczy et al., 2010)
egyetértési arány:

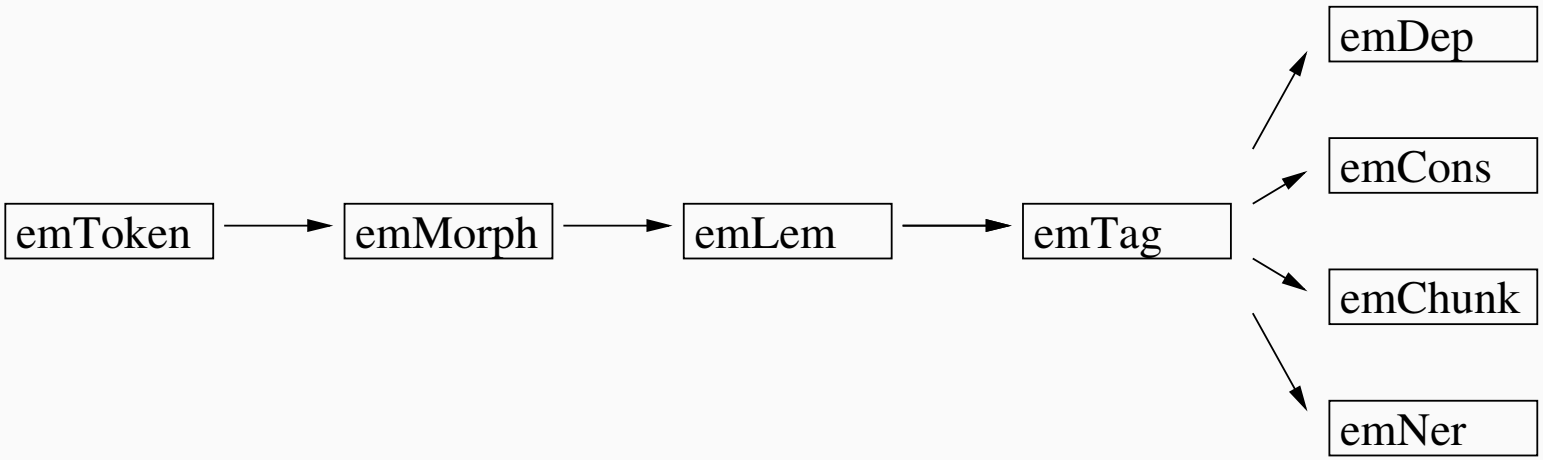
- 1. körben: 17%
- 2. körben: 48%

Annotációs szintek

Alapszintű szövegfeldolgozási szintek

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

Egymásra épülő annotációk



Elemzőláncok a magyarra

- e-magyar
  - *e-magyar.hu* és *emtsv*
  - az egyes modulok közötti átjárást az egységes formátum és az *xtsv* keretrendszer biztosítja
  - a modulok egymásra épülnek, de külön-külön is használhatók
  - az egyes elemzési lépéseknél ki-be lehet szállni a láncba
- magyarlánc
  - Java modulok
  - az egész egyben futtatható parancssorban és beépíthető nagyobb rendszerekbe is

Tokenizálás és mondatrabontás

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* 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.

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

Kimeneti formalizmusok 1.

MSD (Erjavec, 2004)

- pozícióalapú
- az első pozíció mindig a szófaji kategóriáé, a többi pedig további morfoszintaktikai infókat kódol
- *Vmis2s---y*: kijelentő módú, múlt idejű, egyes szám második személyű, tárgyas ragozású főige
- lemma külön
- nincs szegmentálás, nincs deriváció, nincsenek jelölve az allomorfok, csak morfoszintaktikai kódok vannak
- nem hierarchikus, és nem tükrözi a morfológiai jelöltséget
- sok nyelvre
- magyarlánc 2.0, Szeged Korpusz és Treebank 2.5

Kimeneti formalizmusok 2.

Universal Dependencies and Morphology

- univerzális szófajkódok fix halmaza és nyelvspecifikus elemekkel bővíthető feature–érték párok halmaza
- meg van adva, hogy milyen feature milyen értékeket vehet fel
- hierarchikus jegy–érték struktúra (Attribute–Value Structure, AVS)
- ez sem tükrözi a morfológiai jelöltséget
- lemma külön
- nincs szegmentálás, nincs deriváció, nincsenek jelölve az allomorfok, csak morfoszintaktikai kódok vannak
- *hozzád*:
  `Case=All|Number=Sing|Person=2|PronType=Prs`
- magyarlánc 3.0, Szeged UD Treebank

Kimeneti formalizmusok 3.

KR (Rebrus et al., 2012)

- hierarchikus: irányított körmentes gráf (fa)
- a gyökércsomópont a szófaj
- bináris morfoszintaktikai jegyek és ezek pozitív és negatív értékei
- lemma külön
- nincs szegmentálás, nincs deriváció, nincsenek jelölve az allomorfok, csak morfoszintaktikai kódok vannak
- *fotelben*: `fotel/NOUN<CAS<INE>>`, *fotelban*: `fotel/NOUN<CAS<INE>>`
- `hun*` eszközlánc

Kimeneti formalizmusok 4.

emMorph (Novák et al., 2017)

- van szegmentálás, jelölve vannak a derivációk, az allomorfok, van lemma, van morfoszintaktikai annotáció
- mint a glosszázás:

harmad napon halottaiból feltámada

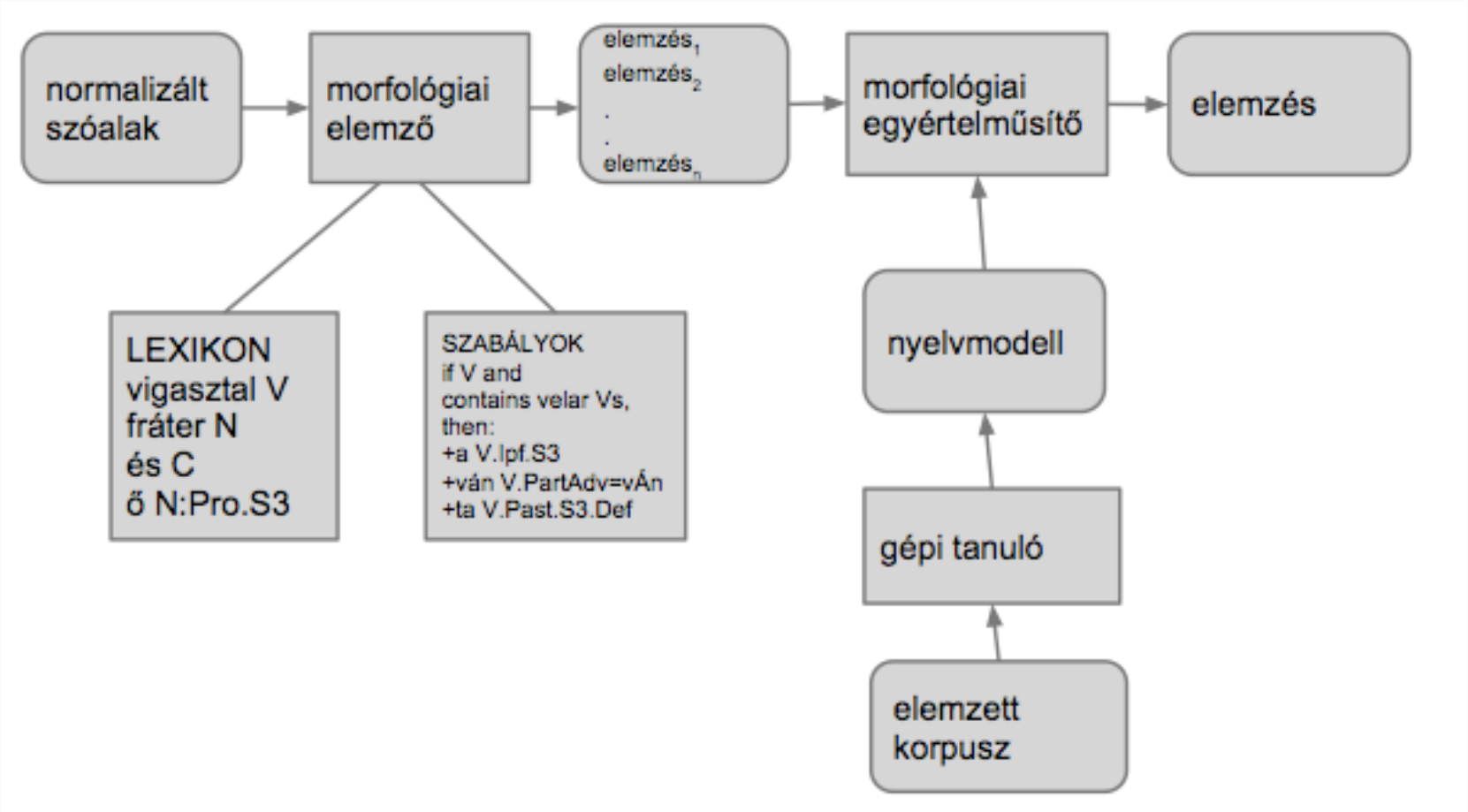
```
három[/Num]=harm + ad[_Frac/Num] + [Nom]
nap[/N] + on[Supe]
halott[/N] + ai[Pl.Poss.3Sg] + ból[Ela]
fel[/Prev] + támad[/V] + a[Pst.NDef.3Sg]
```

harmal	napon	halottay bool	felthamata
harmad	nap-on	halott-a-i-ból	fel-támad-a
third	day-SUP	dead-POSS-PL-ELA	up-rise-PST.3SG

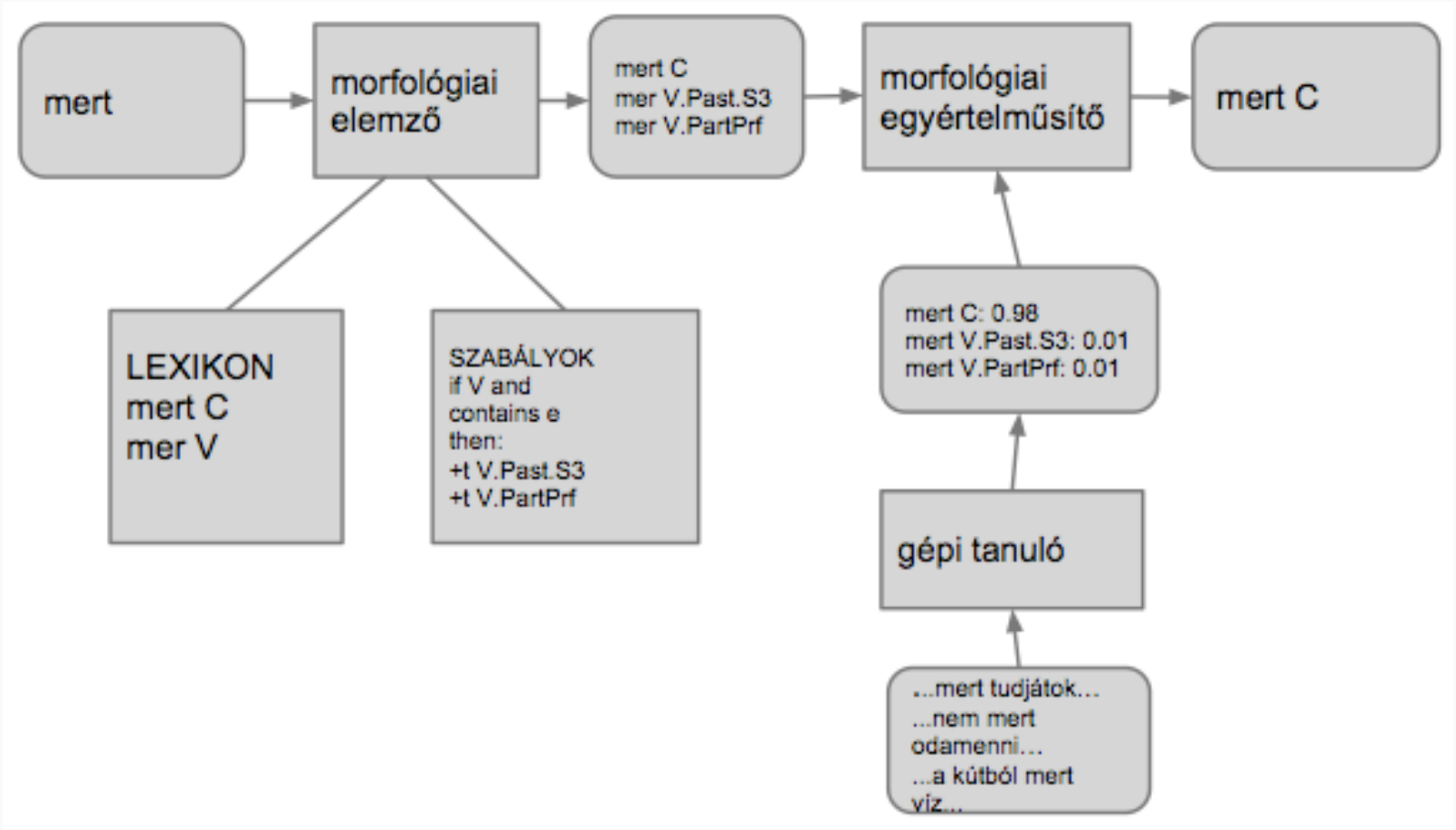
‘on the third day he is risen from the dead’ (Müncheni emlék 114v)

Morfológiai egyértelműsítés

Morfológiai egyértelműsítés 1.



Morfológiai egyértelműsítés 2.



Tulajdonnév-felismerés

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

Irodalom

Irodalom i

- Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, 34(4).
- 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.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46
- Erjavec, T. (2004). MULTEXT-East Morphosyntactic Specifications. Version 3.0. http://nl.ijs.si/ME/Vault/V3/msd/html/.

Irodalom ii

- Krippendorff, K. (1980). Content Analysis: An Introduction to Its Methodology. Sage, Beverly Hills, CA, first edition.
- Krippendorff, K. (2004). Content Analysis: An Introduction to Its Methodology. Sage, Thousand Oaks, CA, second edition.
- 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.
- Novák, A., Rebrus, P., and Ludányi, Zs. (2017). Az emMorph morfológiai elemző annotációs formalizmusa. In XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017), pages 70–78, Szeged.

Irodalom iii

- Rebrus, P., Kornai, A., and Varga, D.ú(2012). Egy általános célú morfológiai annotáció. Általános Nyelvészeti Tanulmányok, XXIV.:47–80.
- 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.