Korpuszannotáció

Számítógépes nyelvészet – 2018 tavasz
7. óra

Simon Eszter – Mittelholcz Iván

2018. április 11.

MTA Nyelvtudományi Intézet

Tartalom

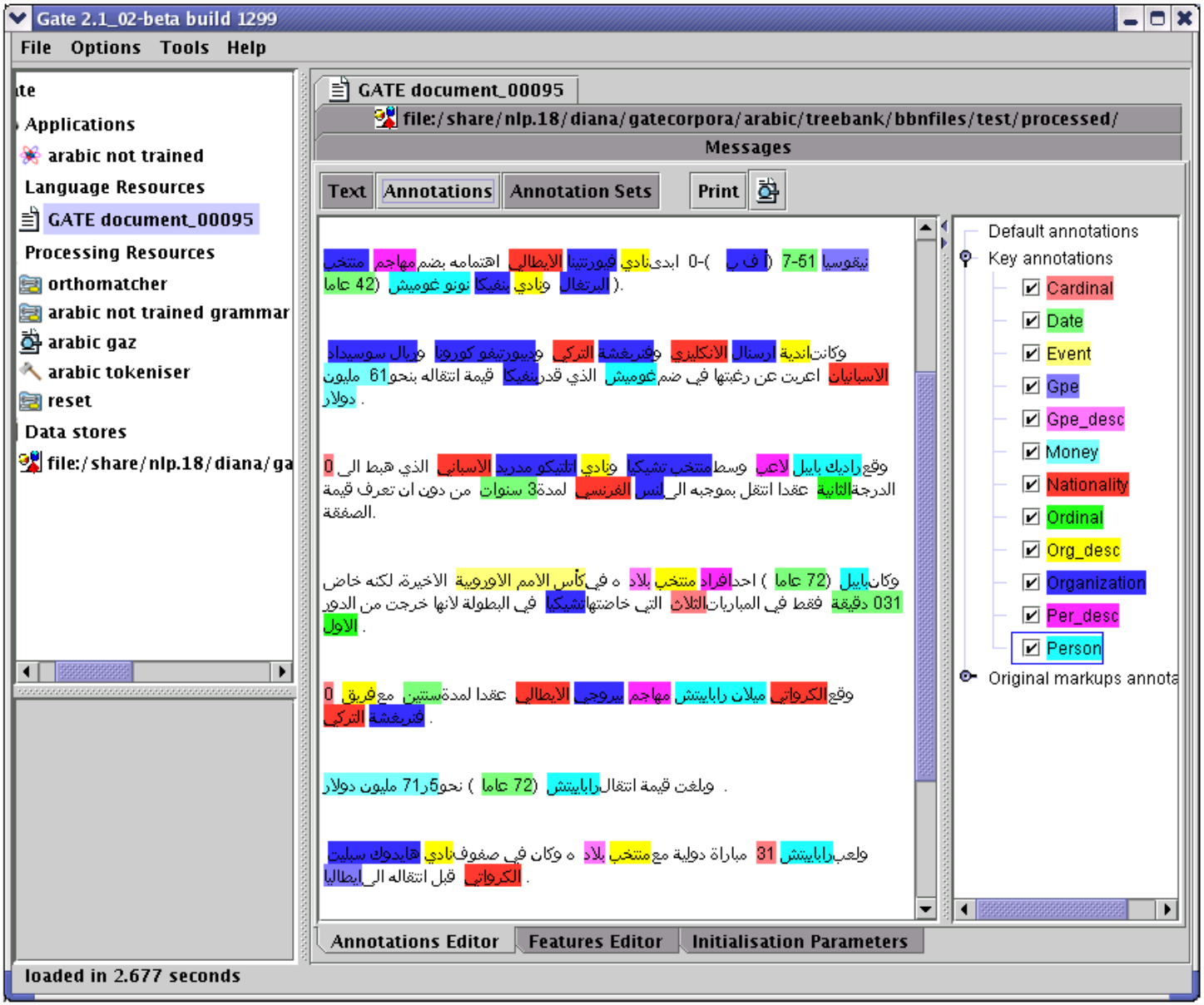
Annotációs eszközök

GATE

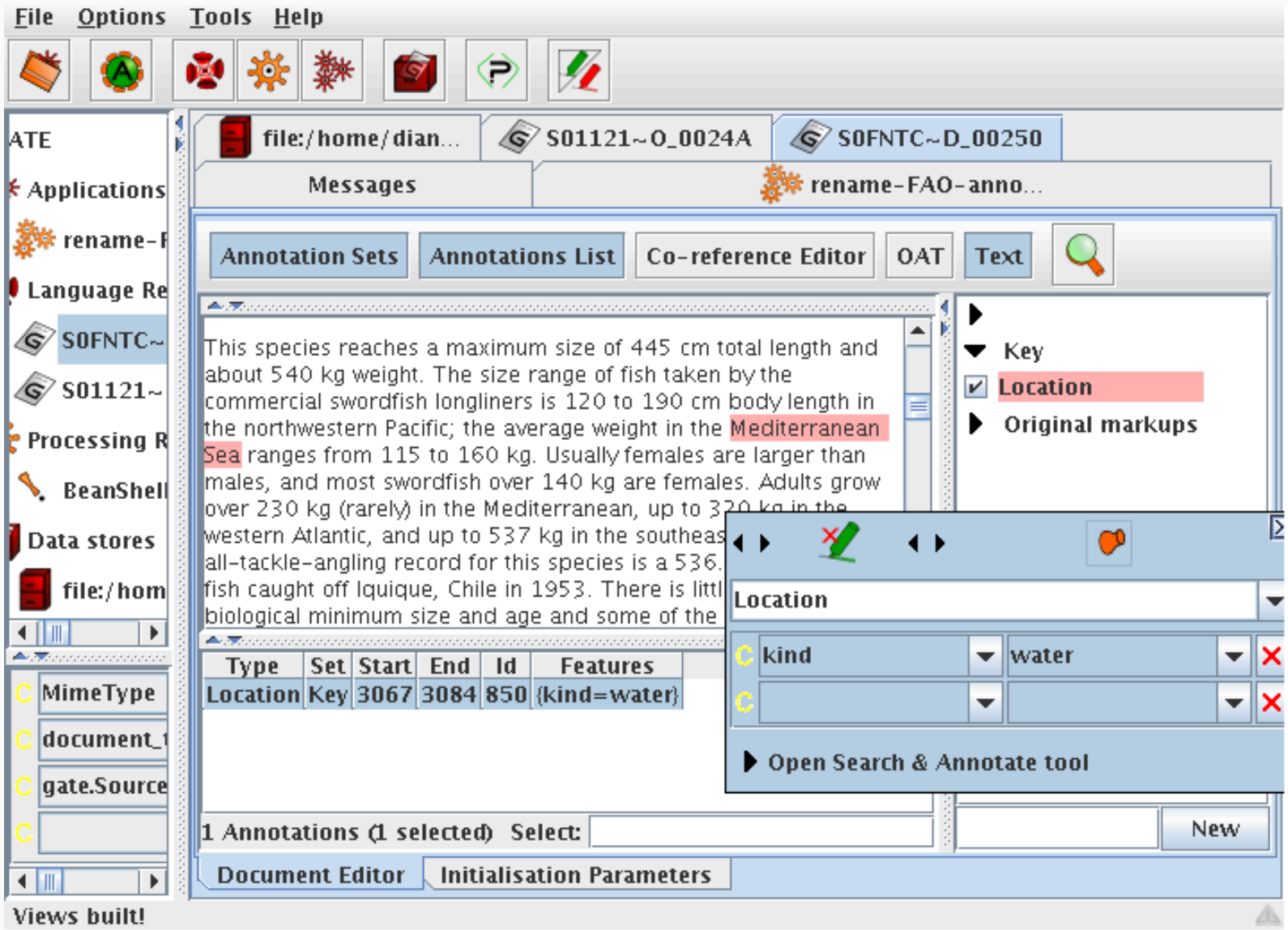
General Architecture for Text Engineering (GATE)

- LGPL 3.0
- GATE Developer: grafikus felület szövegfeldolgozó eszközökhöz
- GATE Teamware: kollaboratív annotációs eszköz
- Language Resources (LRs) és Processing Resources (PRs)
- standoff annotáció XML-ben :)

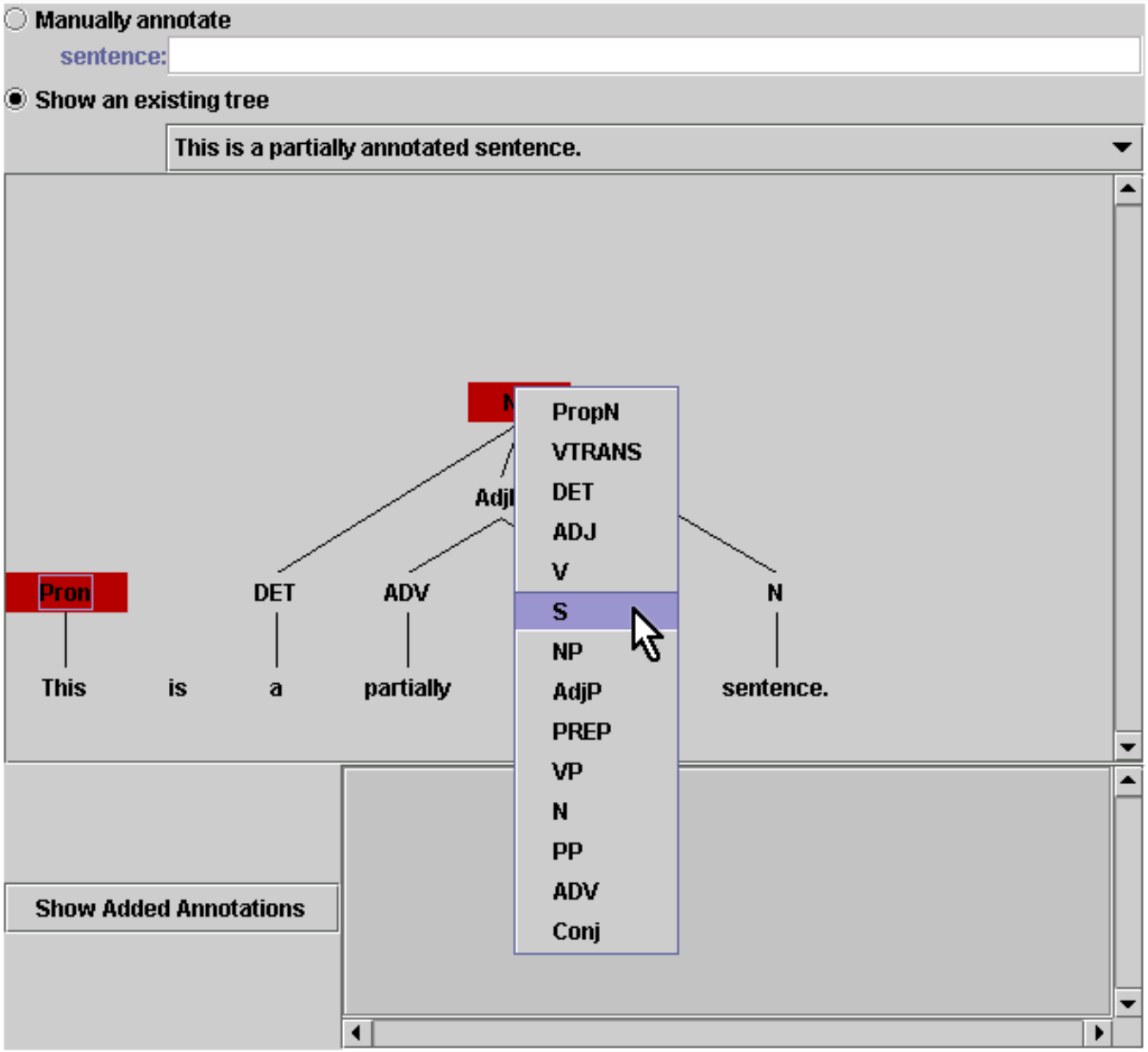
GATE képernyőlövések 1.



GATE képernyőlövések 2.



GATE képernyőlövések 3.



További annotációs eszközök

Szövegekhez:

- brat
- MMAX2
- egyebek

Hangzó anyagokhoz:

- ELAN
- Praat
- EXMARaLDA

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

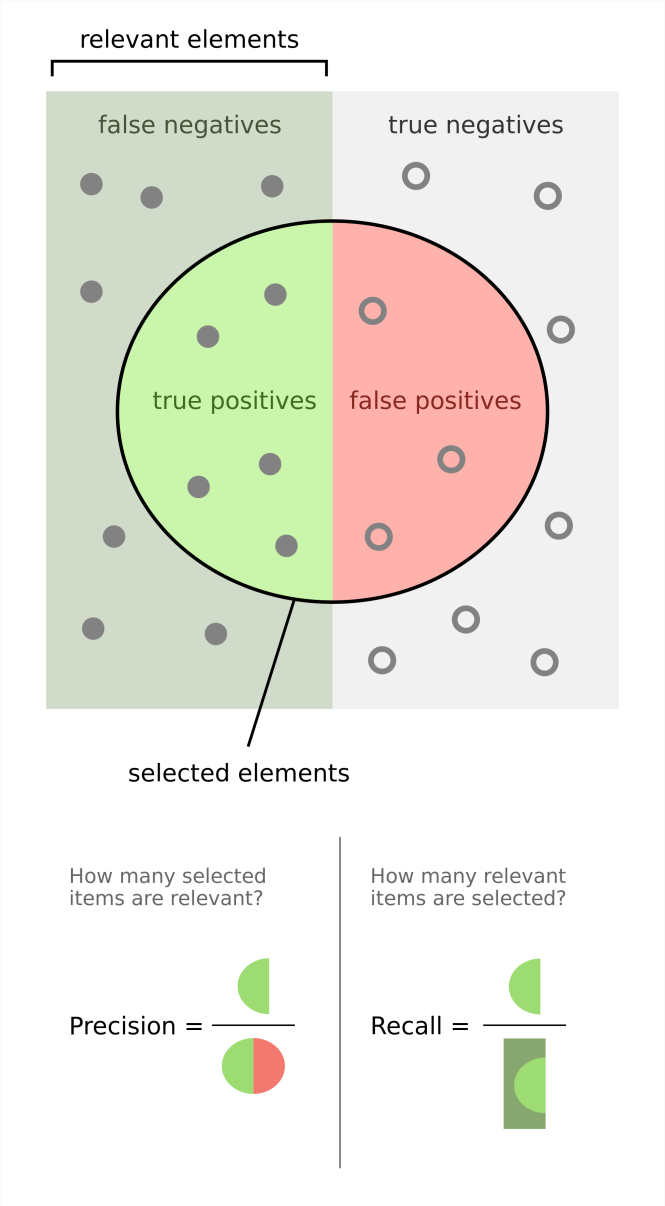
True Positive (TP): a rendszer helyesen felismerte a NE-t;

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):

$$\frac{A \cap B}{A + B - A \cap B}$$

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 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%

References

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.

Irodalom II

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Sage, Beverly Hills, CA, first edition edition.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, second edition edition.

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

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*.