

# Korpuszannotáció

Számítógépes nyelvészet – 2018 tavasz  
7. óra

---

Simon Eszter – Mittelholcz Iván

2018. április 11.

MTA Nyelvtudományi Intézet

## 1. Annotációs eszközök

## 2. Annotációk összevetése

Automatikus annotáció kiértékelése

Annotátorok közötti egyetértés

## Annotációs eszközök

---

## 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.

The screenshot displays the GATE 2.1\_02-beta build 1299 interface. The window title is "Gate 2.1\_02-beta build 1299". The menu bar includes "File", "Options", "Tools", and "Help". The left sidebar shows a tree view with "Applications" (arabic not trained), "Language Resources" (GATE document\_00095), "Processing Resources" (orthomatcher, arabic not trained grammar, arabic gaz, arabic tokeniser, reset), and "Data stores" (file:/share/nlp.18/diana/ga). The main window displays a document titled "GATE document\_00095" with the file path "file:/share/nlp.18/diana/gatecorpora/arabic/treebank/bbnfiles/test/processed/". The "Messages" tab is active, showing a list of annotations: "Text", "Annotations", "Annotation Sets", "Print", and "Annotations Editor". The text in the main window is Arabic, with various words highlighted in different colors (pink, blue, green, yellow, red) to indicate annotations. The right sidebar shows a list of "Default annotations" with checkboxes for "Cardinal", "Date", "Event", "Gpe", "Gpe\_desc", "Money", "Nationality", "Ordinal", "Org\_desc", "Organization", "Per\_desc", and "Person". The bottom status bar indicates "loaded in 2.677 seconds".

Gate 2.1\_02-beta build 1299

File Options Tools Help

te

Applications

- arabic not trained

Language Resources

- GATE document\_00095

Processing Resources

- orthomatcher
- arabic not trained grammar
- arabic gaz
- arabic tokeniser
- reset

Data stores

- file:/share/nlp.18/diana/ga

GATE document\_00095

file:/share/nlp.18/diana/gatecorpora/arabic/treebank/bbnfiles/test/processed/

Messages

Text Annotations Annotation Sets Print

Default annotations

- Key annotations
  - ☒ Cardinal
  - ☒ Date
  - ☒ Event
  - ☒ Gpe
  - ☒ Gpe\_desc
  - ☒ Money
  - ☒ Nationality
  - ☒ Ordinal
  - ☒ Org\_desc
  - ☒ Organization
  - ☒ Per\_desc
  - ☒ Person
- Original markups annotations

Annotations Editor Features Editor Initialisation Parameters

loaded in 2.677 seconds

## GATE képernyőlövések 2.

The screenshot displays the GATE (General Architecture for Text Engineering) software interface. The main window is titled "Document Editor" and shows a text document with a selected annotation. The annotation is a "Location" key, with a start position of 3067 and an end position of 3084. The text being annotated is "Mediterranean Sea".

The interface includes a menu bar (File, Options, Tools, Help) and a toolbar with various icons. The left sidebar shows a project tree with folders like "Applications", "Language Resources", "Processing Resources", and "Data stores". The main text area contains the following text:

This species reaches a maximum size of 445 cm total length and about 540 kg weight. The size range of fish taken by the commercial swordfish longliners is 120 to 190 cm body length in the northwestern Pacific; the average weight in the Mediterranean Sea ranges from 115 to 160 kg. Usually females are larger than males, and most swordfish over 140 kg are females. Adults grow over 230 kg (rarely) in the Mediterranean, up to 320 kg in the western Atlantic, and up to 537 kg in the southeast. The all-tackle-angling record for this species is a 536 kg fish caught off Iquique, Chile in 1953. There is little biological minimum size and age and some of the

The annotation table at the bottom of the text area shows the following data:

Type	Set	Start	End	Id	Features
Location	Key	3067	3084	850	(kind=water)

The right sidebar shows a "Key" section with a checked "Location" item and a "Original markups" section. Below this is a "Location" dropdown menu with a "kind" dropdown set to "water". At the bottom, there is a "1 Annotations (1 selected) Select:" field and a "New" button.

# GATE képernyőlövések 3.

☐ Manually annotate  
sentence:

☒ Show an existing tree

This is a partially annotated sentence.

The screenshot shows the GATE (General Architecture for Text Engineering) interface. At the top, there are two radio buttons: 'Manually annotate' (unselected) and 'Show an existing tree' (selected). Below the 'Show an existing tree' option, there is a text box containing the sentence 'This is a partially annotated sentence.' Below this, a tree diagram is displayed. The sentence is tokenized into words: 'This', 'is', 'a', 'partially', 'annotated', 'sentence.'. The word 'This' is annotated with 'Pron'. The word 'is' is annotated with 'V'. The word 'a' is annotated with 'DET'. The word 'partially' is annotated with 'ADV'. The word 'annotated' is annotated with 'V'. The word 'sentence.' is annotated with 'N'. A context menu is open over the word 'partially', showing a list of possible annotations: PropN, VTRANS, DET, ADJ, V, S (highlighted), NP, AdjP, PREP, VP, N, PP, ADV, and Conj. A mouse cursor is pointing at the 'S' option. At the bottom left, there is a button labeled 'Show Added Annotations'.

PropN  
VTRANS  
DET  
ADJ  
V  
S  
NP  
AdjP  
PREP  
VP  
N  
PP  
ADV  
Conj

Show Added Annotations

## Szövegekhez:

- brat
- MMAX2
- egyebek

## Hangzó anyagokhoz:

- ELAN
- Praat
- EXMARaLDA



## Annotációk összevetése

---

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

szigorúan véve azonosan címkézett elemek azok, amelyeknek

1. ugyanazok a határaik, vagyis

- ugyanott kezdődnek
- ugyanott végződnek ÉS

2. ugyanaz a címkéjük

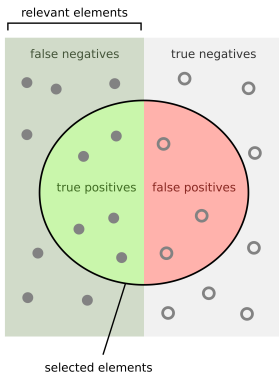
**True Positive (TP):** a rendszer helyesen felismerte a NE-t;

**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



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

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

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

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

- a *pontosság* maximalizálása: minél kevesebb tévedés → szigorítás
- a *fedés* maximalizálása: minél több találat → megengedőbb rendszer

$$\beta = 1$$

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

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



# 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  $\kappa$  (Cohen, 1960):
  - $\kappa = 1$ , ha az annotátorok teljes mértékben egyetértenek
  - $\kappa = 0$ , ha az annotátorok a véletlen egybeesésnél nem jobban értenek egyet
- Krippendorff's  $\alpha$  (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 érték magasabb a véletlen egybeesésnél, vagyis szisztematikus egyet nem értésről van szó

## Landis and Koch (1977)

$\kappa$	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

## 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.

- 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*.