# Annotációs szintek

## Számítógépes nyelvészet – 2018 tavasz
## 8. óra

Simon Eszter – Mittelholcz Iván

2018. április 18.

MTA Nyelvtudományi Intézet

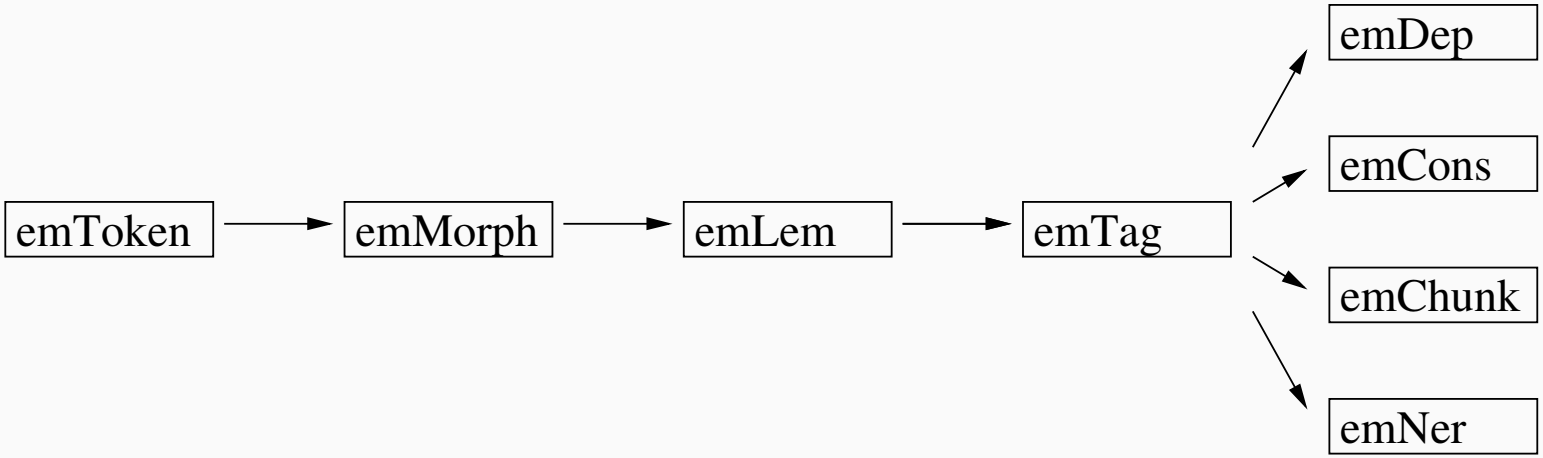
# Tartalom

# Bevezetés

# Alapszintű szövegfeldolgozási szintek

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

# Egymásra épülő annotációk



# Elemzőláncok a magyarra

- e-magyar
  - futtatható a GATE-keretrendszeren belül és használható online is
  - az egyes modulok közötti átjárást a GATE formátuma biztosítja
  - a modulok egymásra épülnek, de külön-külön is használhatók
- 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
- Szeged Korpusz és Treebank

## 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) (Trón, 2002)
- 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`
- Szeged 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

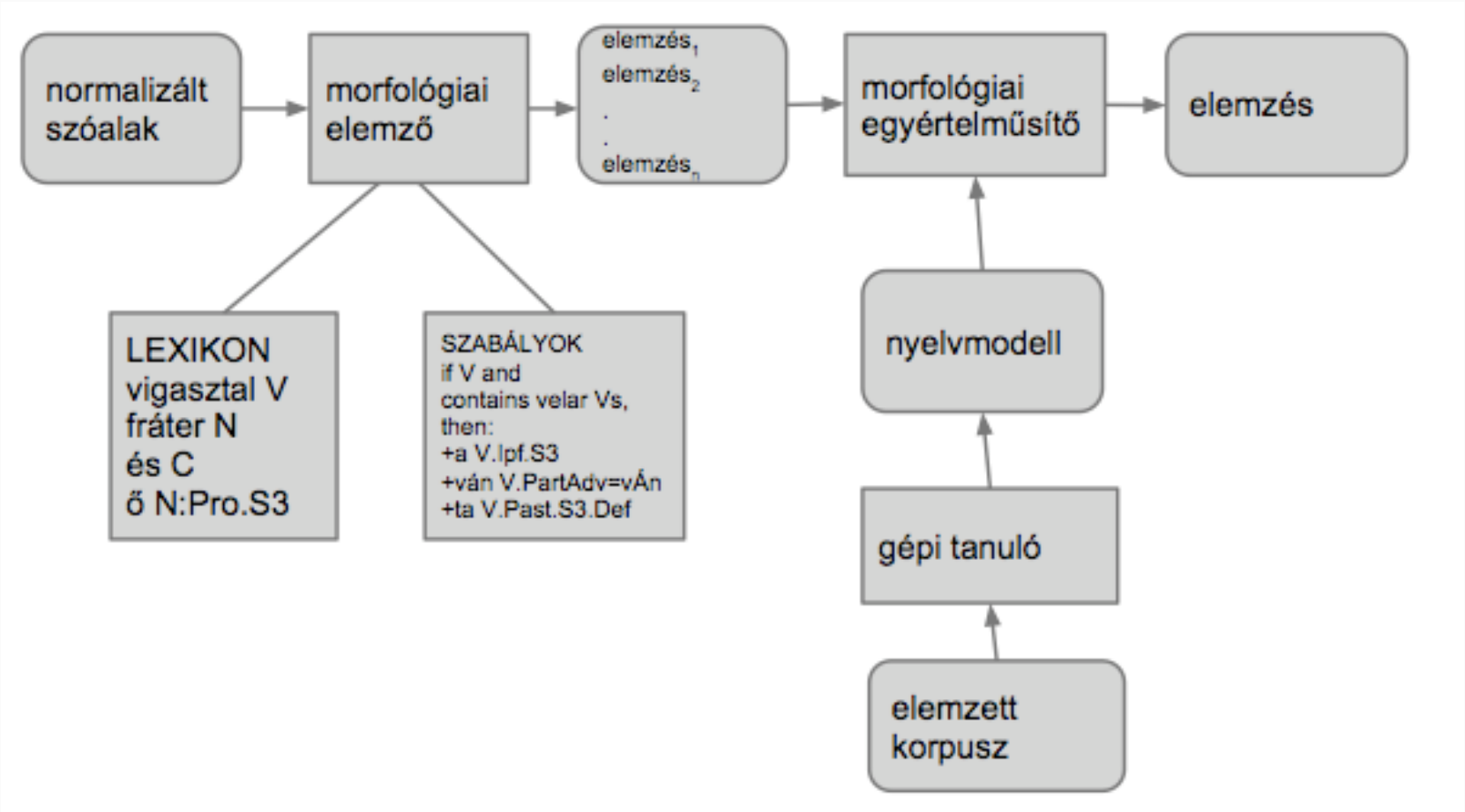
```
három[/Num]=harm + ad[_Frac/Num] + [Nom]
nap[/N] + on[Supe]
halott[/N] + ai[Pl.Poss.3Sg] + ból[Ela]
fel[/Prev] + támad[/V] + a[Pst.NDef.3Sg]
```

| | | | |
|---|---|---|---|
| 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 |

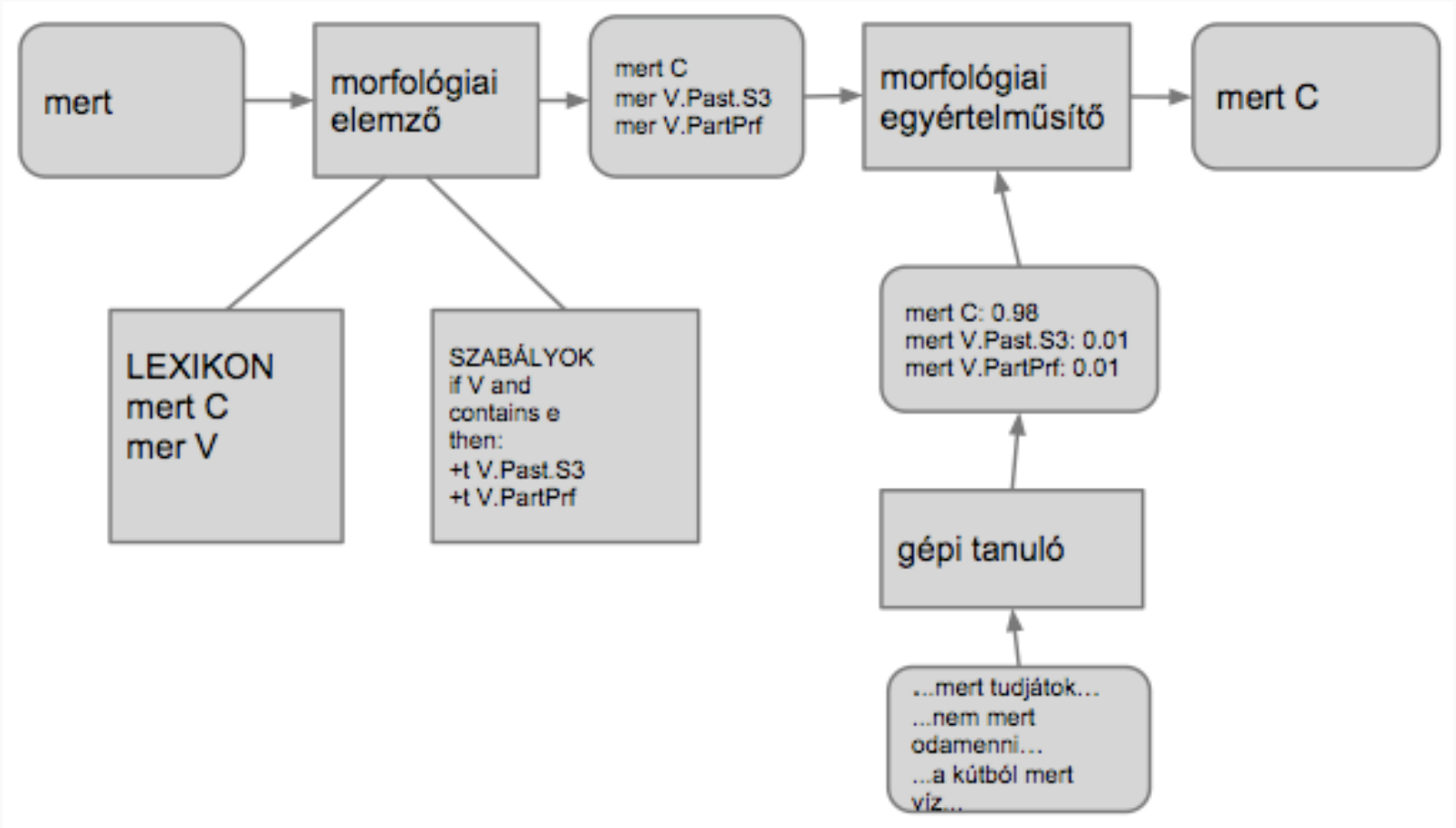
‘on the third day he is risen from the dead’ (Müncheni emlék 114v)

# Morfológiai egyértelműsítés

# Morfológiai egyértelműsítés 1.



# Morfológiai egyértelműsítés 2.




Nézzük meg az e-magyart!

# Szekvenciális címkézési feladatok

# HunTag(3)

- **hunner** (Varga & Simon, 2006) → **HunTag** (Recski & Varga, 2009) → Liblinear → **HunTag3** (Indig)
- felügyelt gépi tanuláson alapuló rendszer
- többféle szekvenciális elemzési feladatra alkalmas
- sztenderd CoNLL-formátum: tsv, BIE1
- Latin-2/UTF-8 bemenet
- bemenet: szavakra és mondatokra bontott szöveg, egyértelmű morfológiai annotációval ellátva
- kimenet: ugyanez + NE/NP címkézés
- GNU Lesser General Public License v3.0 licenc alatt elérhető: https://github.com/recski/HunTag/ és https://github.com/ppke-nlpg/HunTag3

# 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

# Szintaktikai elemzés

# 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

# References

Erjavec, T. (2004). *MULTEXT-East Morphosyntactic Specifications. Version 3.0*. http://nl.ijs.si/ME/Vault/V3/msd/html/.

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.

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

# Irodalom II

Trón, V. (2002). Attribútum–érték struktúrák. In Kálmán, L., Trón, V., and Varasdi, K., editors, *Lexikalista elméletek a nyelvészetben*, volume XIII. of *Segédkönyvek a nyelvészet tanulmányozásához*, pages 333–344. Tinta Könyvkiadó, Budapest.

# Házi feladat

# Házi feladat

- ketten-hárman összefogni,
- specifikálni egy annotálási alfeladatot,
- annotációs sémát és útmutatót gyártani hozzá,
- kiválasztani egy szöveget, tokenizálni és mondatra bontani az e-magyarral,
- ketten-hárman leannotálni, gold standardban megállapodni (NE dobjatok ki belőle mondatokat!),
- annotátorok közötti egyetértést számolni (akár több körben is),
- a szöveget áttolni az e-magyaron,
- a fent létrehozott gold standard annotációval összevetve kiértékelni az e-magyar teljesítményét az adott feladaton

# Az annotálási alfeladat

- legyen olyan, amit az e-magyar lefed,
- legyen az e-magyar kimenetével konvertibilis,
- ne legyen túl bonyolult,
- a gold standard szöveg tartalmazzon legalább 100 adatpontot

## Választható alfeladatok:

1. tulajdonnevek felismerése (PER, LOC, ORG, MISC)
2. maximális NP-k felismerése
3. tárgyesetű főnevet vonzó igék felismerése
4. ...

# Beküldés

## Mit kell elküldeni a végén?

- a bemenő szöveg
- annotációs séma és útmutató
- az annotátorok annotációit tartalmazó tsv fájl (ID TAB token TAB annotáció1 TAB annotáció2 TAB gold)
- az annotátorok közötti egyetértés számítása ((leírás vagy szkript) és eredmények)
- az e-magyar kimenete (tsv)
- kiértékelés ((leírás vagy szkript) és eredmények)

## Határidő

- ápr. 25.: a terv beküldése
- máj. 16.: a kész anyag beküldése