Számítógépes morfológia

Nyelv és informatika – Pécs, 2022/23 tavasz
3. óra

Simon Eszter – Vadász Noémi

2023. február 18.

Tartalom

A szókincs modellezése

Morfológiai elemzés

1. a szóalakot elemi morfémáira bontja
2. meghatározza a morfémák lexikális alakját
3. meghatározza a morfémák morfoszintaktikai tulajdonságait (esetleg egyéb nyelvtani tulajdonságokat)

pl. *többértelműségeket*

- `sok[/Num]=tö+bb[_Comp/Num]=bb+értelem[/N]=értelm+ű[_Adjz:Ú/Adj]=ű+ség[_Nz_Abstr/N]=ség+ek[Pl]=ek+et[Acc]=et`
- `többértelmű[/Adj]=többértelmű+ség[_Nz_Abstr/N]=ség+ek[Pl]=ek+et[Acc]=et`
- `többértelműség[/N]=többértelműség+ek[Pl]=ek+et[Acc]=et`

A szókincs modellezése

Az összes lehetséges szóalak felsorolása helyett:

- a morfémák szerepelnek a szótárban
- szabályokkal írjuk le a szóalakok felépítését

$\rightarrow$ formális nyelvtan

Nehézség: a jelentésfüggő morfológiai jelenségek kezelése

- **túlgenerálás** a produktív szabályok által
- **zártság**

Szótárak és szabályok

morfoszintaktikai szabályok: az egyes morfémák hogyan (milyen sorrendben és milyen feltételek mellett) következhetnek egymás után egy szóalakban

- a morfémák szótárban (morfématárban, lexikonban) vannak
- metainformációkkal, különféle osztályokba rendezve
- a tőmorfémák külön lexikonban, szófajkódokkal
- affixumok is külön (a szóalakban pre- vagy szuffixumok)
- külön lexikonba mehetnek az előtagok, a nem feltétlenül szóvégi szuffixumok (képzők)

Különböző szabálymegadási modellek

- kétszintes morfológiák
- folytatási osztályok
  - minden morféma mellett jelöli, hogy milyen morféma követheti (pl. minden tőhöz, hogy milyen toldalék)
  - *labda [főnév] (+ t [tárgyrag], + val [eszközhatározó rag], + nak [birtokosrag], …); + k [többesszám jele] (+ at [tárgyrag], + nak [birtokosrag], + val [eszközhatározó rag], + ból [helyhatározó rag] stb.)*
  - a morfémák osztályai az egyes folytatási osztályok
- unifikációs modellek
  - minden morféma (mindkét oldalán) összetett adatstruktúra a morféma morfoszintaktikai és morfofonológiai tulajdonságaival
  - nem a kapcsolódó morfémákat, hanem a morfémát megelőző és követő morfémák jellemzőit (morfoszintaktikai és fonológiai, hangképzési tulajdonságok) tárolja
  - az illeszkedési pontoknál meg kell vizsgálni (unifikáció), hogy a jegyszerkezetek (feature structure) passzolnak-e

Szótár a számítógépen

- a szótárban elfogadható méretűnek kell lennie
- a benne való keresésnek elég gyorsnak kell lennie

$\rightarrow$ nem elég egy lineárisan kereshető lista

megoldás: állapotátmenetes modell (állapotgép, automata) ahol a természetes nyelv teljes – véges, de nem korlátos, sőt nem is zárt – szókincsét egy véges halmazzal közelítjük.

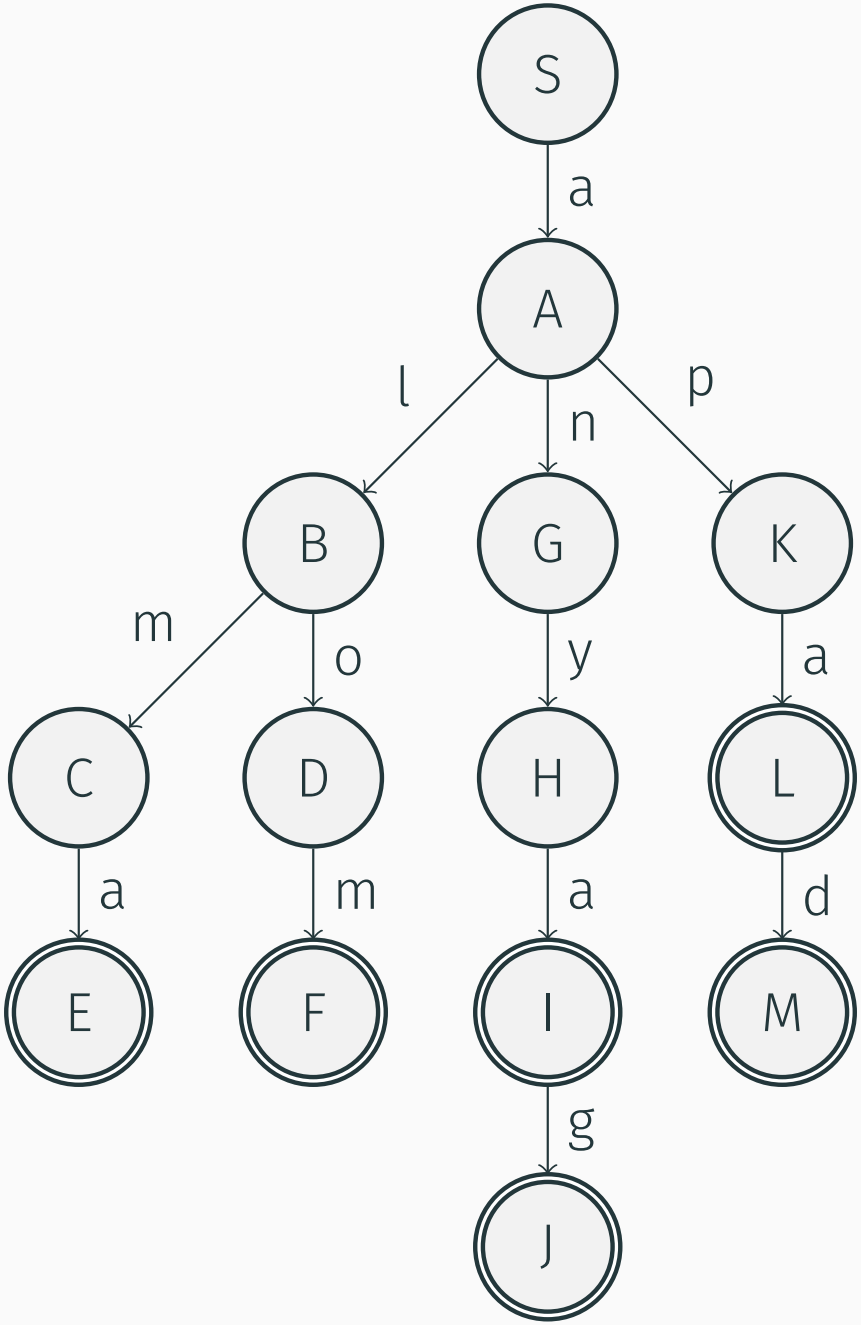
A véges automata (finite state automaton, FSA)

- a legegyszerűbb állapotátmenet-modell
- képes azonosítani (elfogadni) bizonyos karaktersorozatokat
- bemenet: az azonosítandó karaktersorozat
- a karaktersorozat elemeinek hatására különböző állapotokba kerül (állapotátmenet)
- elfogadta a karaktersorozatot, ha a végén elfogadó állapotban van
- az elfogadott összes karaktersorozat halmaza az automata által felismerhető *nyelv*

1. véges sok állapota van
2. a következő állapotot csak az előző állapot és a bemeneten kapott utolsó karakter figyelembe vételével határozza meg
3. az állapotátmenet során más műveletet nem hajt végre

kiterjesztett véges automata (extended finite state machine, EFSM): az utóbbi két feltétel közül valamelyik nem teljesül

Az állapotátmenet-gráf (state transition graph)

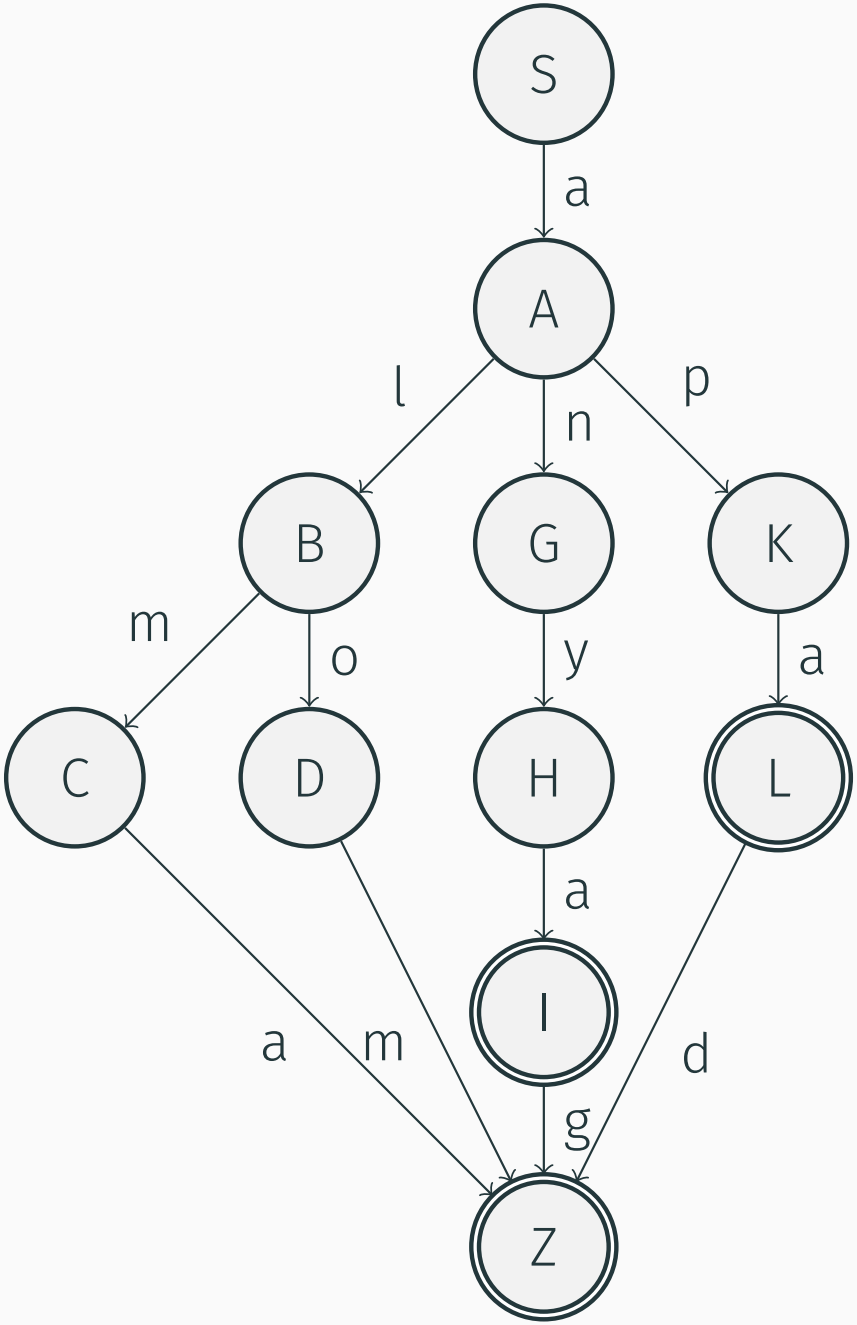


Teljesítmény

pl. ötbetűs szavak egy százezer elemű szótárban

- lépések száma
  - egyszerű lista
    - annyi elemet olvasunk el, ahányadik eleme a szó a szótárnak
    - ha a szó nincs benne a szótárban, akkor végig kell olvasnunk
  - automata
    - az egy szó megkereséséhez szükséges műveletek száma a szó betűinek számával arányos
    - ha a szó nincs benne a szótárban, akkor annyi elemi művelet kell, amennyi a szótárban tárolt leghosszabb szó betűinek száma
- adatigény
  - egyszerű lista: 24 betű
  - automata: 13 betű

További tömörítési lehetőségek



Kétszintes morfológia

- 1983: Koskenniemi
- az egyes morfémák felszíni és lexikális alakjai közötti transzformáció
- véges automata
- a felszínen olvasott szóalakok és az azokat felépítő morfémák lexikális alakja között közvetlen kapcsolatot definiál
- az egyszerű véges automata csak a szintaktikai elemzés legelső lépését képes elvégezni egy toldalékolt szóval: megállapítja, hogy része a modellezett nyelvnek, vagy sem
- az automata kibővítésével az automata kimenete gazdagítható: képes megadni
  - a morfémák alap lexikális alakját (értelm $\rightarrow$ értelem)
  - a morfémák morfoszintaktikai kódját (-ja $\rightarrow$ [birtokos jelző])
- megfordítható (elemzés és generálás)
- minden szabályt véges automata (reguláris nyelvtan) segítségével határoznak meg

Kétszintes morfológia (Koskenniemi)

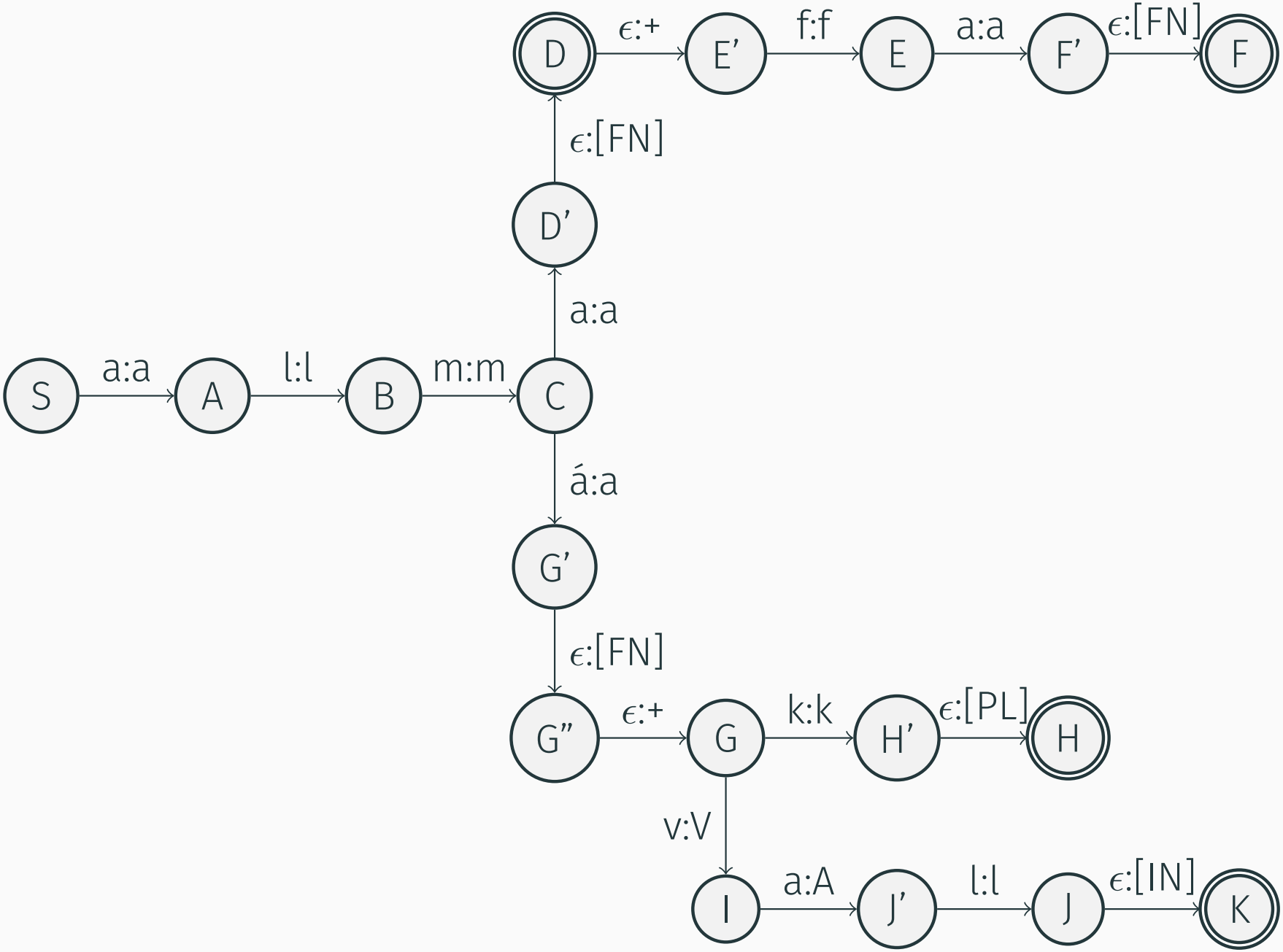
Mit kell tudnia egy morfológiai elemzőnek?

- a bemeneti szóalak felismerése mellett
- a tőmorfémák lexikális alakját és
- az egyes morfémák morfoszintaktikai kódjait is szolgáltatja
- tehát nem csak **felismer**, hanem **transzformál**

Kétszintes morfológia

- minden, teljes hasonuláson és hangrend-illesztésen átesett toldalékmorféma *lefordítva* ugyanúgy jelenik meg
- a fordító az absztrakt szimbólumok alkalmazásával jelzi, hogy a felszínen nem feltétlenül ugyanazok a szimbólumok szerepeltek
    1. szint: a bemeneti szalagon fogadott szimbólumsorozat (surface level)
    2. szint: a kimeneti szalagon megjelenő szimbólumsorozat (lexical level)
- a két szint között transzformáció
- az FST működése megfordítható (generálás)

FST (finite state transducer)



Nyelvi jelenségek a számítógépes morfológiai elemzésben

Egy morfológiai elemző (Humor)

- egy szóalak lehetséges elemzéseit morfsorozatokként adja meg
- minden morfhoz felszíni és mögöttes alak
- strukturált információ vagy szerkezetnélküli címke
- belső összetevős szerkezet nélküli lapos morfsorozatok
- reguláris szónyelvtan: determinisztikus és epszilonmentes véges állapotú automata:
  - gyorsabb, mint egy környezetfüggő nyelvtanon alapuló elemző
  - elkerülhető sok irreleváns szerkezeti többértelműség elő állítása
- az elemző olyan morfokat keres a szótárában, amelyeknek a felszíni alakja illeszkedik a megadott szó még elemzetlen részére
- a lexikon morfsorozatokat is tartalmaz
- kétféle ellenőrzés:
  - lokális kompatibilitás-ellenőrzés az egymás mellett álló morfok között: morfofonológiai (pl. mgh-harmónia) és lokálisan ellenőrizhető morfotaktikai feltételek (pl. névszói toldalékok csak névszótövekhez)
  - az elemzést alkotó morfémák megfelelnek-e az adott nyelv morfológiai konstrukcióit leíró szónyelvtannak

Szóalaktani adatbázis, a nyelvész feladatai

- a nyelv morfémakategória-készletének leírása
- a tő- és toldalékalternciók leírása (reg.kifekkel)
- a morfológiai tulajdonságok feltérképezése
- szomszédos morfok közötti szelekciós megszorítások definiálása (két tulajdonsághalmaz morfonként, egy balról, egy jobbról)
- a morfémák és allomorfok tulajdonságai közötti implikációs viszonyok megadása
- a tő- és toldaléklexikonok előállítása (lexikai alakjuk, a kategóriájuk és a megjó- solhatatlan vagy rendhagyó tulajdonságaik és elvárásaik; öröklési mechanizmus)
- a szónyelvtan leírása
- külön toldaléknyelvtan leírása (gyorsítás az agglutináló nyelvek esetében, a szónyelvtan egyszerűsödik)

A kutya szó reprezentációja

```
lemma : 'kutya[FN]'
  root: 'kutya'
  allomorf: 'kutya'
   mcat: 'S_FN'
   rp: -Vs -nyi -sÁg –tAlAn =_s =_t =i =jA =vAl VHB
       Vfin cat_N cmp2 sfxable mcat_stem'
   rr: '!FVL'
   lp: 'Cini comp2 k_ini'
   lr: '!cat_vrb'
  allomf:'kutyá'
   mcat: 'S_FN'
   rp: '-Vs -nyi -sÁg –tAlAn =_s =_t =i =jA =vAl
VHB
       Vfin cat_N cmp2 sfxable mcat_stem'
   rr: 'FVL'
   lp: 'Cini comp2 k_ini'
   lr: '!cat_vrb'
```

Lexikai többértelműség

- gazdag morfológia → a lehetséges alakok száma magas → morfológiai elemzés kell
- produktív szóalaktani jelenségek (ragozás, képzés, szóösszetétel) kezelése
- a lexikai többértelműség gyakori
  - homonímia
  - a különböző paradigmák véletlenszerű vagy rendszerszerű átfedései
  - a paradigmán belüli rendszeres átfedések

Lexikai többértelműségek

Átfedések az igei paradigmán belül:

- *vettem, mostam* (én valamit vagy én azt)
- *vennétek, mosnátok* (ti valamit vagy ti azt)
- *vennék* (én valamit vagy ők azt)
- *eszik* (ő vagy ők azt)
- *néztek* (ti most vagy ők akkor
- *festette* (ő azt akkor vagy ő azt valakivel akkor)

Átfedések a névszói paradigmán belül: *gyerekével* (az ő gyerekével vagy a gyerek valamijével)

Egyedi tőhomonímák: *vár, várnak, nyúl, nyúlnak* (tőhomonímák)

A paradigmák egyes tagjai esnek egybe: *mentek, csend*

Fontos szerepet játszik a valószínűség!

Valószínűség

– *Van egy nyers gyémántom.*
– *Na és van csiszoltad is?*

valószínűtlen, kizárható esetek:

- a sokmorfémás elemzés kompozicionálisan kiszámítható jelentése abszurd
- a kompozicionális jelentés nem képtelenség, de lexikalizált jelentésű tő van benne (az anyanyelvi beszélőben fel sem merül, hogy felbontsa morfémákra)

Többértelműségek a képzett szavak körében

-s	*harcosak, barackosak* (melléknév)
	harcosok, barackosok (főnév)
-ó	*abban bizakodó* (melléknévi igenév)
	bizakodóak (melléknév), *ablakmosó* (főnév)
	ablakmosó gép (jelzői helyzetben)
-z(ik)	*ftp-zik, (le)ftp-z: ftp-zett/ftp-zett le*
-(t)at(ik)	*kihirdettet, kihirdettetik: kihirdettetett*

- *nyávogós* 1. *nyávog+ó+s* 2. *nyávog+ós*
- *katonáskodik* 1. *katoná+skodik* 2. ?*katoná+s+kodik*
- *elmagyarosodik* 1. *magyar+osodik* 2. ?*magyar+os+odik*

panmorph: morfológiai címkekészletek

panmorph

- összegyűjtöttük és közzétettük a magyarra alkalmazott morfológiai annotációs sémákkal és címkekészletekkel kapcsolatos elérhető információkat
- konvertereket írunk a címkekészletek között
- https://github.com/nytud/panmorph

Morfológiai címkekészlet:

- **informativitás**: pontosság és teljesség
- **adekvátság**: nyelvészetileg megalapozott kategóriák
- **egyszerűség**: kézi és automatikus feldolgozhatóság

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
- magyarlanc 2.0 elemzőlánc kimenete

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) (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
- magyarlanc 3.0 elemzőlánc kimenete

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)

panmorph konverterek

Közvetlen leképezés az egyik címkekészletről a másikra:

- *emmorph2msd*
- *emmorph2conll*
- *emmorph2ud*

Miért az emMorph címkét konvertáljuk?

- az emMorph címkekészlet a legfinomabb, legrészletesebb
- az egyik bevett elemzőlánc, az e-magyar bocsátja ki, ezért jól beilleszthetők a konverterek a szövegfeldolgozási folyamatba

Konverzió címkekészletek között

- szerencsés, ha egy-az-egyhez megfelelés áll fenn a bemenet és a kimenet között
- sok esetben kellett aleseteket és kivételeket kezelni
- néha a lemmára vagy a token felszíni alakjára is támaszkodni kellett
- zárt szóosztályok (pl. kötőszavak, névmások) esetén felsorolhatóak az alesetek
- igekötők, tulajdonnevek kezelése eltér
- bizonyos címkék soha nem jelennek meg a kimenetben, noha a kimeneti készletek tartalmaznak címkéket a jelenségekre: -nAk ragos névszók, segédigék

Irodalom

Hivatkozások

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

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.