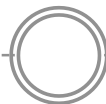




A nyelvtechnológia alapjai



6. Gyakorlat

Gyakorlati morfológia

- Az unifikáció olyan művelet, mely egyedül akkor nem hajtható végre, ha a tagokban egyazon jegy különböző értékekkel szerepel:
 - [NUMBER SG] U [NUMBER PL] = fail!
 - de:
 - [NUMBER SG] U [NUMBER SG] = [NUMBER SG]
 - [NUMBER SG] U [NUMBER []] = [NUMBER SG]
 - [NUMBER SG] U [PERSON 3] = [[NUMBER SG][PERSON 3]]
- HUMOR (High-speed Unification Morphology)

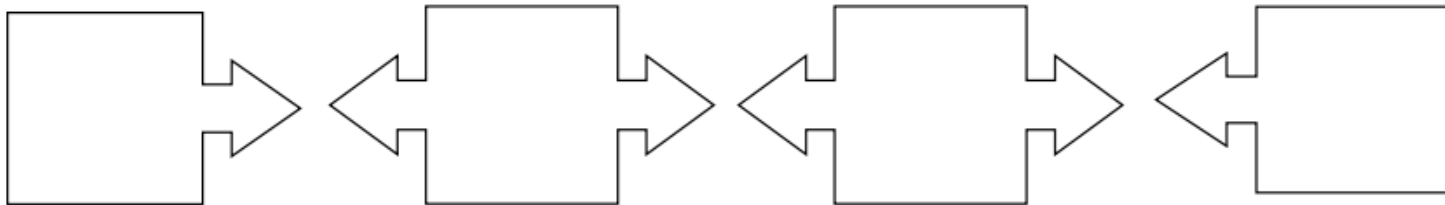


tő

nemterminális
toldalék

nemterminális
toldalék

terminális
toldalék



(relatív) tő / relatív toldalék

relatív tő / relatív toldalék

relatív tő / (relatív) toldalék



hó	[[+névszó +fn +szótári –elől –kerek –PL –PERS –ACC –ACCKötő +DAT +INS:V]
hav	[[+névszó +fn –szótári –elől –kerek +PL +PLkötő +PERS +ACC –DAT –INS]
képez	[[–névszó +szótári +elől –kerek –ÁS]
képz	[[–névszó –szótári +elől –kerek +ÁS]



ás	[–névszó –elől +ÁS] [+névszó +fn +szótári –elől –kerek +PL +PLkötő +ACC –ACCKötő +DAT +INS:S]
és	[–névszó +elől +ÁS] [+névszó +fn +szótári +elől –kerek +PL +PLkötő +ACC –ACCKötő +DAT +INS:S]
ak	[+névszó –elől –kerek +PL +PLkötő] [+névszó –elől –kerek –PL –PERS +ACC +ACCKötő +DAT +INS:K]
ek	[+névszó +elől –kerek +PL +PLkötő] [+névszó +elől –kerek –PL –PERS +ACC +ACCKötő +DAT +INS:K]
nak	[+névszó –elől +DAT] []
nek	[+névszó +elől +DAT] []

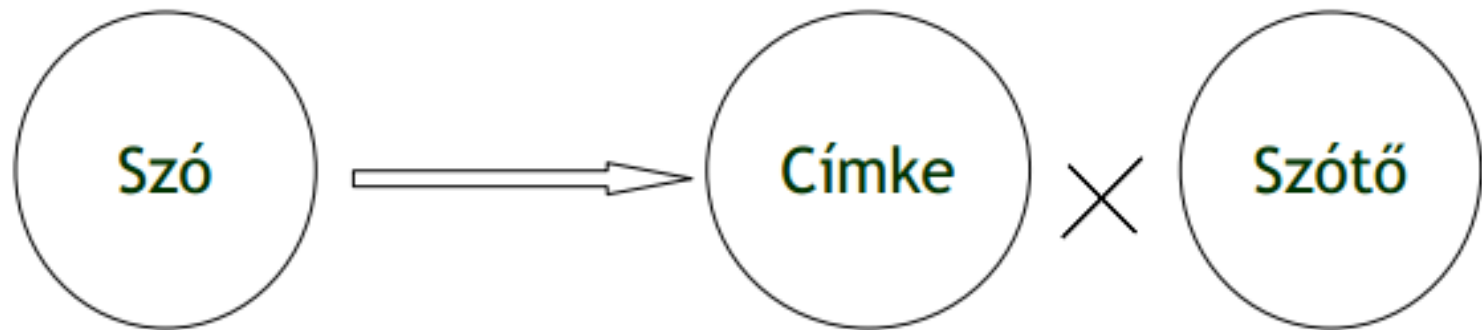


hó	[+névszó +fn +szótári -elől -kerek -PL -PERS -ACC -ACCKötő +DAT +INS:V]
*hav	[+névszó +fn <u>-szótári</u> -elől -kerek +PL +PLkötő +PERS +ACC +DAT -INS]
hó+nak	[+névszó +fn +szótári -elől -kerek -PL -PERS -ACC -ACCKötő +DAT +INS:V] [+névszó -elől +DAT]
*hav+nak	[+névszó +fn -szótári -elől -kerek +PL +PLkötő +PERS +ACC <u>-DAT</u>] [+névszó -elől <u>+DAT</u>]
hav+at	[+névszó +fn -szótári -elől -kerek +PL +PLkötő +PERS +ACC -DAT] [+névszó -elől +ACC]
*hó+vel	[+névszó +fn +szótári <u>-elől</u> -kerek -PL -PERS -ACC -ACCKötő +DAT +INS:V] [+névszó <u>+elől</u> +INS:V]



<http://corpus.nytud.hu/cgi-bin/e-szokincs/alaktan>
kereshetésűekéire

- Egyértelműsítő:



{tokenizált mondat, címkekészlet}
→ **szófaji egyértelműsítő** →
{címkézett (és szótövezett) mondat}

- Az utca végén levő legelőre mentem legeltetni.

Az = **az[DET]=Az**
az[NM]=Az
utca = **utca[FN]+[NOM]**
végén = **vég[FN]+é[POS]+n[SUP]**
vég[FN]+é[PSe3]+n[SUP]
levő = **lesz[IGE]=lev+ő[MIF]+[NOM]**
legelőre = **legelőre[HA]**
legelő[FN]+re[SUB]
legel[IGE]+ő[MIF]+re[SUB]
mentem = **ment[IGE]+em[Te1]**
ment[MN]+em[PSe1]
megy[IGE]=men+tem[Me1]
megy[IGE]=men+t[MIB]+em[PSe1]
legeltetni. = **legeltet[IGE]+ni[INF]+.[STOP]**
legel[IGE]+tet[MUV]+ni[INF]+.[STOP]



- hunpos (2007)
 - BME MOKK
 - TnT modell
- MagyarLánc (2010)
 - SZTE
 - Stanford POS-tagger + morfológiai elemző
- PurePOS (2013)
 - PPKE ITK
 - a TnT + integrált morfológiai elemzővel (Humor)



Demó



<http://corpus.nytud.hu/people/joker/demo/egy/>



kacsónak

✓ N

kacsóna + k

* N + PL

kacsón + ak

* N + PL

kacsó + nak

* V + PL3

kacsó + nak

✓ N + DAT

kacs + ó + nak

* V + PART + DAT

ka | csónak

* N | N

- Innentől a szóalak-gyakoriság dönt:

- kacsónak, *kacsónakot, *kacsónakra, ... → {}
- kacsó, kacsót, kacsóra, kacsói, ... → {kacsó}



Számítógépes morfológiai alkalmazások

A hibák és normák kezelése („proofing tools”)

- A helyesírási normák, a szabályok és a gépi helyesírásellenőrzés viszonya
- A helyesírási szabályzatok viszonya a gépi helyesírásellenőrzéshez
- Formai és jelentéstani szempontok gépi kezelése
- A gép a „nyelvhelyességi tanácsadó” szerepében
- A nehezen értelmezhető elemzések okozta problémák
- Szószinten javítható hibák: betűhibák és elütések, hibás egybeírások, hibás kötőjelezés, szóismétlések (!)
- Szószinten nem javítható hibák: hibás különírások, központoszási hibák, egyeztetési hibák, szórendhibák, vonatkozó névmások helytelen használata, stílushibák, helytelen szóhasználat, szóközök többszörözése vagy elhagyása, „értelmes” szóhibák

- Karakterhibák, elütések
- Valódi helyesírási hibák (általában szószintűek)
- Nyelvhelyességi hibák (általában nem szószintűek)
- Tipográfiai hibák (az átlag felhasználó nem is ismeri őket)
- A helyesírás-ellenőrzés lehetőségei és korlátai a szavak szintjén
- A szóellenőrzés és az ún. nyelvhelyesség-ellenőrzés viszonya
- A nyelvi programrendszer lehetséges hibái: a túlgenerálás (kör/kőr, alak/ház/tan + it/ít)

- Alapszótár ill. morfológia
- Ajánlómodul és adatbázisai
- Időleges sajátyszótárak
- Kiegészítő szótár
- Kizáró szótár
- A ragozó nyelvek esetén hasznos toldalékoló sajátyszótárak lehetőségei és nehézségei

Két string Levenshtein-távolsága

- A Levenshtein-távolság a Hamming-távolság egyfajta általánosítása.
- Nemcsak egyforma hosszú füzérekre, és nemcsak betűcsere van az átalakításban, hanem beszúrás és törlés is.
- $s = \text{alma}$, $t = \text{kala}$

		k	a	l	a
	0	1	2	3	4
a	1	1	1	2	3
l	2	2	2	1	2
m	3	3	3	2	2
a	4	4	3	3	2

ha $s[i] == t[j]$ akkor $\text{cost} = 0$; különben $\text{cost} = 1$;

$a = M[i-1, j] + 1$ (felső tag + 1 $\rightarrow$ törlés)

$b = M[i, j-1] + 1$ (baloldali tag + 1 $\rightarrow$ beszúrás)

$c = M[i-1, j-1] + \text{cost}$ (felső baloldali tag + cost – helyettesítés)

Helyettesítés	Törlés
Beszúrás	$M[i, j] = \min(a, b, c)$



1. Morfológiai elemzés

kérdésse → <nincs ilyen szó a magyarban>

2. Ajánlás

- törlés:

érdésse, krdésse, kérésse, kédésse, kérdése, kérdéss

- helycsere:

ékrdésse, krédésse, kérédsse, ..., kérdéses

- nyelvspecifikus csere:

kérdéssé, kérdésse, ...

3. Ellenőrzés a generátumok morfológiai elemzésével

kérdése, kérdéses, kérdésse, kérdéssé



kérdése

kérdés[FN]+e[PSe3]

főnévi

kérd[IGE]+és[IF]+e[PSe3]

főnévi

kérdéses

kérdéses[MN]

melléknévi

kérdés[FN]+es[SKEP]

melléknévi

kérd[IGE]+és[IF]+es[SKEP]

melléknévi

kérdesse

kérd[IGE]+es[MUV]+se[TPe3]

igei

kérdéssé

kérdés[FN]+sé[FAC]

főnévi

kérd[IGE]+és[IF]+sé[FAC]

főnévi



Alap

#VV

#VC

VV#

VV

VC_1C_2V

VC_iC_iV

VCc_1c_2V

Vc_1c_2CV

$Vc_{11}c_{12}c_{21}c_{22}V$

$Vc_1c_1c_2V$

Elválasztva

#VV

#VC

VV#

V-V

VC_1-C_2V

VC_i-C_iV

$VC-c_1c_2V$

Vc_1c_2-CV

$Vc_{11}c_{12}-c_{21}c_{22}V$

$Vc_1c_2-C_1c_2V$

Példa

autó

alaki

hazai

ba-uxit

er-kély

vet-tem

mor-zsa

asz-tal

tarisz-nya

össze / ösz-sze



Köszönöm a figyelmet



ध्यान देने के लिए आपका धन्यवाद
आप अपना ध्यान के लिए धन्यवाद
थैंक यू फॉर योर अटेंशन

Thank you for your attention
Thank you for your attention
Thank you for the attention

Merci de votre attention
Je vous remercie pour votre attention
Merci pour l'attention

感谢您的关注
谢谢你的关注
谢谢您的注意