A Történeti magánéleti korpusz

Mi a TMK?

- Szövegek válogatásánál fő szempont: informális nyelvhasználat (megszorításokkal) ---> magánlevelek és tanúvallomások
- 15– 18. század, férfiak és nők, fiatalok és kevésbé fiatalok, iskolázottak és kevésbé iskolázottak az ország különböző tájairól, hogy történeti szociolingvisztikai vizsgálatokat is lehessen végezni
- Ugyanakkor semmilyen szempontból nem kiegyensúlyozott, de egy történeti korpusz aligha lehet az
- Eredeti szöveg (ahogy kiadták) + normalizált változat + morfológiai elemzés

A korpuszépítés folyamata

digitalizálás	A nyomtatásban megjelent szövegek szkennelése, szövegfelismertetés, javítás.
normalizálás	Az eredeti szövegváltozat átalakítása úgy, hogy a morfológiai elemző számára értelmezhető legyen.
morfológiai elemzés	Házon kívül történt a Humor magyar morfológiai elemző kibővített változatával, Novák Attila végezte.
egyértelműsítés	Elő-egyértelműsítés (szintén Novák Attila programja) után a felajánlott elemzési lehetőségek közül a megfelelő kiválasztása.

Normalizálás

- Átírás mai magyarra (a nyelvjárási hangtani jelenségek és a helyesírási esetlegességek eltűnnek)
- Cél: az eredeti morfológiai szerkezet pontos tükrözése úgy, hogy az átírt szöveg bemenetként szolgálhasson a számítógépes morfológiai elemzőnek.
- Teljesen kézi, de minden normalizált szöveget három körben ellenőriztünk, hogy minél kevesebb hiba maradjon.
- A vitás eseteket megbeszéltük, a választott megoldást rögzítettük.

Bemenet és kimenet

{\!lat!Ad·6tum}¶
{\!lat!Ad·6tum}¶
Hallotta·a·fatens·Balogné·Aszszonyom·szájából,¶
Hallotta·a·fatens·Balogné·asszonyom·szájából,¶
hogy·a·majorne·egy·patyolatot·kért,¶
hogy·a·majorné·egy·patyolatot·kért,¶
hogy¶
hogy,¶
<ha·neki·egy·patyolatot·ád>¶
<ha·neki·egy·patyolatot·ad,>¶
el\·alszik·a·te·Urad.¶
elalszik·a·te·urad.¶

[2] Bosz. 12. Bereg megye, ?Gelénes, 1702. ,> ,> (NE) - 971726

~~!lat!Ad-6tum~~	Hallotta	a	fatens	Balogné	Aszszonyom	szájából,
~~!lat!Ad-6tum~~	Hallotta	a	fatens	Balogné	asszonyom	szájából,
	hall	a	fatens	Balog+né	asszony	száj
	V.Past.S3.Def	Det	N	N	N.PxS1	N.PxS3.Ela

hogy	a	majorne	egy	[patyolatot]	kért,
hogy	a	majorné	egy	**patyolatot**	kért,
hogy	a	major+né	egy	**patyolat**	kér
C	Det	N	Det	**N.Acc**	V.Past.S3

hogy	ha	neki	egy	[patyolatot]	ád
hogy,	<ha	neki	egy	**patyolatot**	ad,>
hogy	ha	ő	egy	**patyolat**	ad
C	C	N|Pro.Dat.S3	Q	**N.Acc**	V.S3

el alszik	a	te	Urad.
elalszik	a	te	urad.
el|+alszik	a	te	úr
VPfx.V.S3	Det	N|Pro.S2	N.PxS2

A kettő között

- Humor morfológiai elemző módosított változata: a tőtárhoz 5000+ lemma, a toldaléktárba 50 új toldalék (és ezeknek az allomorfjai); az elemző nyelvtanát is módosítani kellett
- Nagy kihívás pl.: névmási paradigmák – sokkal több elem, szabálytalanságok, paradigmák bizonyos elemei alulreprezentáltak
- http://www.morphologic.hu/urali/rm/index.php
- Címkerendszer: https://tmk.nytud.hu/cimkejegyzek.php

Egyértelműsítés: néha ijesztő --

-- néha vicces


árulónak,
árulónak,
áru+ló[N.Dat]


almákbul,
almákból,
al+mák
N.Ela


börökre
bőrökre
bőr+ökör[N.PxS3]


bánatom
bánatom
bán+atom
N

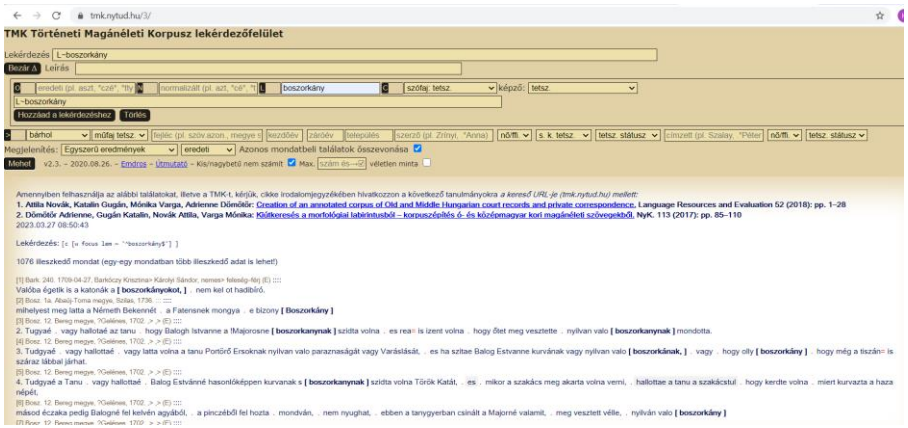

fogtál
fogtál,
fog+tál[N]


félelmében
Félelmében
fél+elme
N.Ine

Adatgyűjtés régen és ma

- Korpusz: 600e karakternyi nyomtatott anyag
- Konkordancia: kb. 12000 cédula, két év munka

- Korpusz: 8,6 M karakter; magyar nyelvű rész: 7,68M karakter, 1112000 elemzett szövegszó
- Konkordancia: 4-5 másodperc



A keresésről:

TMK.NYTUD.HU