

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 Asszonyom szájából,¶

Hallotta a fatens Balogné asszonyom szájából,¶

hogya majorne egy patyolatot kért,¶

hogya majorné egy patyolatot kért,¶

hogya¶

hogya,¶

<ha neki egy patyolatot ad>¶

<ha neki egy patyolatot ad,>¶

el alszik a te Urad.¶

elalszik a te urad.¶

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

Hall!Ad·6tum	Hallotta	a	fatens	Balogné	Asszonyom	szájából,
Hall!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

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

hogya	ha	neki	egy	patyolatot	ad
hogya,	<ha	neki	egy	patyolatot	ad,>
hogya	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 --

és és és[C]	azon azon azon[Det Pro]	adossághát adósságát adósság[N.PxS3.Acc]	hogya hogya hogya[C]	kérettette, kérettette, kér[V.Fact.Fact.Past.S3.Def]
az nap aznap az+nap[N.Tmp_inl]	szolgálójának szolgálójának, szolgáló[N.PxS3.Dat]			kéret[V.Fact.Past.S3.Def] kér[V.Fact.Pass.Past.S3.Def] kér[V.Fact.Fact.Past.S3.Def]
a mely <amely a+mely[N Pro Rel]	által által által[PP]	a a a[Det]	pénz pénz pénz[N]	kéret[V.Pass.Past.S3.Def] kér[V.Pass.Fact.Past.S3.Def] kér[V.Pass.Pass.Past.S3.Def]
lábaj lábai láb[N.PxS3.Pl]	és és és[C]	egyébb egyéb egyéb[Adj Pro]	tagjai tagjai tag[N.PxS3.Pl]	kéret[V.Fact.PartPrf_Subj=tA.PxS3] kér[V.Fact._Nact=tA.PxS3] kér[V.Fact.Pass._Nact=tA.PxS3]
az az az[Det]	Asszonnak asszonyinak asszony[N.Dat]	ellenben ellenben ellenben_szemben[Adv (Abl		kéret[V.Pass._Nact=tA.PxS3] kér[V.Fact.PartPrf=Att.PxS3] kér[V.Fact.Pass.PartPrf_Subj=tA.PxS3]
és és és[C]	fél fél fél[Q]	keze, keze kéz[N.PxS3]	és és fél[Q]	kéret[V.PartPrf=Att.PxS3] kér[V.Pass.Fact._Nact=tA.PxS3] kér[V.Fact.Fact.PartPrf_Subj=tA.PxS3]
lábaj lábai láb[N.PxS3.Pl]	és és és[C]	láb láb láb[N.PxS3]		kér[V.Fact.Fact.PartPrf_Subj=tA.PxS3] kér[V.Fact.Fact._Nact=tA.PxS3] kér[V.Fact.Fact.PartPrf_Subj=tA.PxS3]
hogya hogya hogya[C]	mind mind mind[Adv]	ez e e[Det Pro]	napigh napig nap[N.Ter]	kér[V.Pass.Fact.PartPrf_Subj=tA.PxS3] kér[V.Fact.Fact._Nact=tA.PxS3] kér[V.Pass.PartPrf=Att.PxS3]
is is is[Clit_is]				kér[V.Pass.Pass._Nact=tA.PxS3] kér[V.Pass.Pass.PartPrf_Subj=tA.PxS3]
!!at!Authenticate-modificavit-in-ee: !!at!Authenticate-modificavit-in-ee:				
				sebesültek sebesültek, is+ebesül[V.PartPrf.Pl]
				tt tt ct.Pass.Past.S3]
				harmad napjára, harmadnapjára, harmad+nap[N.PxS3.Sub]
				ass.Past.S3]

-- néha vicces

árulónak,
árulónak,
áru+ló[N.Dat]

almákból,
almákból,
al+mák
N.Ela

börökre
börökre
bőr+ökör[N.PxS3]

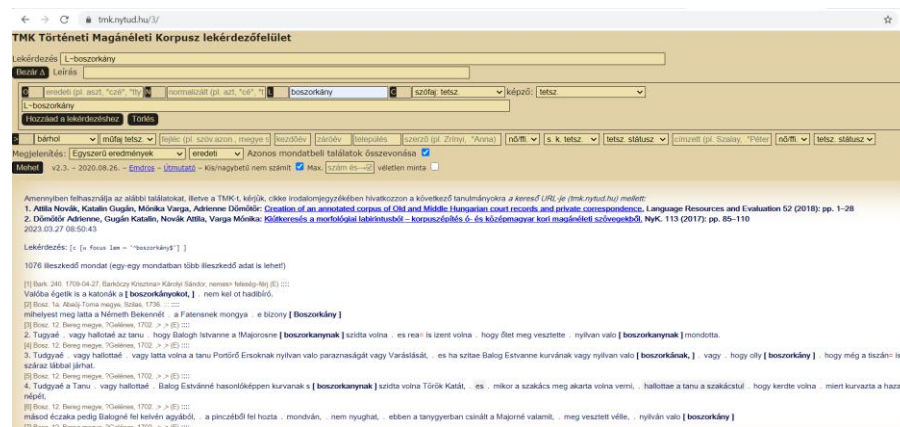
bánatom
bánatom
bán+atom
N

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