

Morfológiai elemzés

A korpusz anyagának egy részén morfológiai elemzést is végeztünk. Ez összesen 62 szövegfájlt, azaz 77 224 tokent fedett le. Az elemzés célja az, hogy lehetővé váljon olyan grammatikai – főként morfoszintaktikai – jelenségek vizsgálata, amely elemzetlen szövegen nagyon időigényes és körülményes volna.

A morfológiai elemzés folyamata

A morfológiai elemzés során a szavakat morfémákra bontjuk, ezeket a morfémákat kategorizáljuk és felcímkézzük, majd megállapítjuk a szó lemmáját (szótári alakját). Ezt a folyamatot az *énekeket* szóalakkal szemléltetjük:

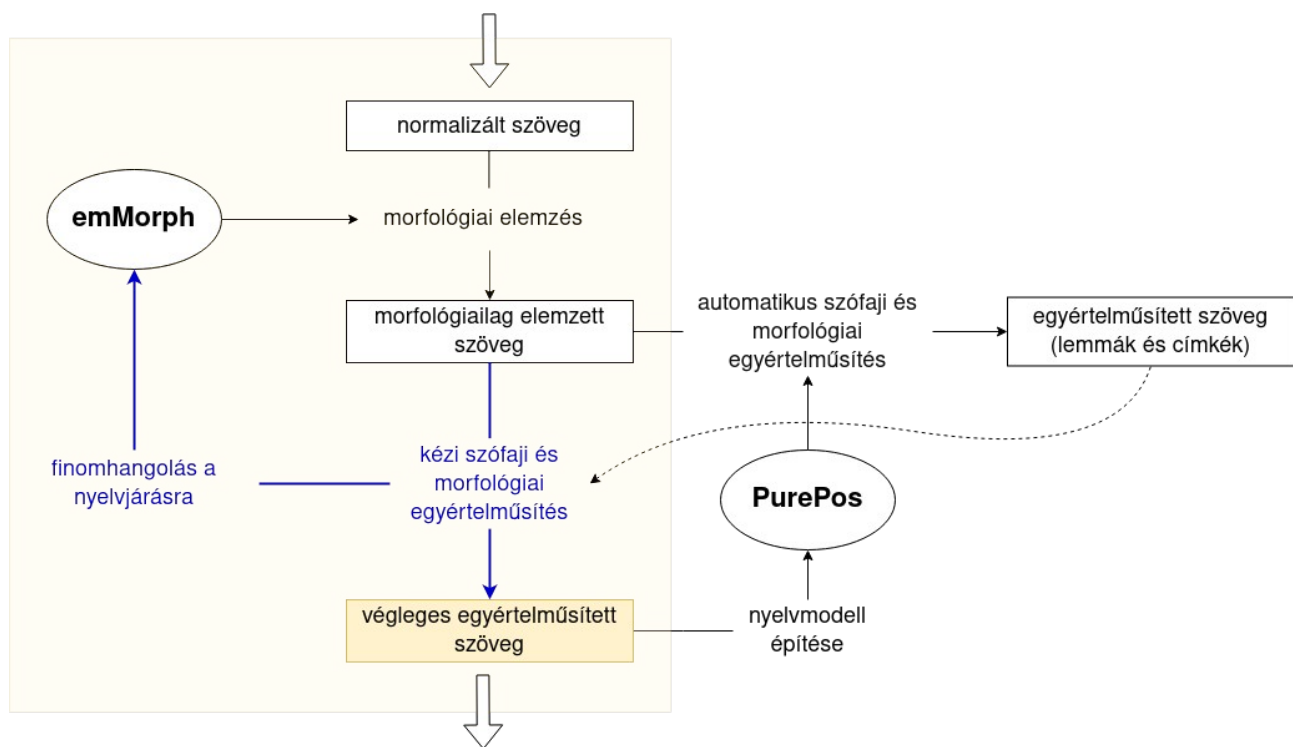
1. a szóalakban három morféma azonosítható, ezek határát pontokkal jelöljük: *ének.ek.et*
2. az egyes morfémák a következő címkéket kapják: *ének* → főnév (jelölése: N), *ek* → többes szám (Pl), *et* → tárgyeset (Acc); az előálló morfológiai elemzés tehát N.Pl.Acc
3. a szó lemmája az *ének* lesz: ez az a szó, amelyre keresve később a különféle toldalékok viselő alakok is megtalálhatók (pl. *énekünk*, *énekekkel*)

A fenti példában az adott szóalakhoz mindössze egy lehetséges elemzés tartozik, de a korpuszban jóval általánosabb az az eset, amikor egy szóalak többféle elemzést kaphat, és a szöveggörnyezettől függ, hogy melyik a helyes. Ekkor szófaji és morfológiai egyértelműsítés is szükséges. Az alábbi táblázat egy ilyen esetet szemléltet, a *mentek* szóalak háromféle lehetséges elemzését:

lemma	címke	jelentés
<i>megy</i>	V.P2	<i>megy</i> ige T/2. jelen idő (<i>ti mentek</i>)
<i>megy</i>	V.Past.P3	<i>megy</i> ige T/3. múlt idő (<i>ők mentek</i>)
<i>ment</i>	V.S1	<i>ment</i> ige E/1. jelen idő (<i>én mentek</i>)

Az automatikus morfológiai elemzéshez az emMorph elemző (Novák 2014; Novák et al. 2016) ó- és középmagyar nyelvállapotra finomhangolt változatát használjuk. Azért döntöttünk emellett, mert a moldvai magyar nyelvjárás sok szempontból ezekhez a korábbi nyelvállapotokhoz áll közel, így több olyan jelenséget is tudunk kezelni, amelyeket egy mai magyarra hangolt elemzővel nem ismernénk fel. Ilyen például a moldvai magyarban gyakori elbeszélő múlt (pl. *mondá, jöve*), vagy az összetett múlt idők (pl. *mondtam volt*). Természetesen szükség volt az elemző további, folyamatos finomhangolására is, hogy le tudja fedni a kizárólag moldvai magyarra jellemző jelenségeket. Ez elsősorban a tőtár bővítését jelentette nyelvjárási szavak hozzáadásával (pl. *csitil* ‘olvas’, *rozáriu* ‘rózsafüzér’, *csán* ‘csinál’).

Az automatikus morfológiai elemzést követően kezdetben kézzel történt a szófaji és morfológiai egyértelműsítés. Az átnézett szövegekből tanítókörpuszt készítettünk a PurePOS (Orosz – Novák 2012, 2013) számára, amely egy statisztikai alapú – rejtett Markov-modellre épülő – címkéző program. A PurePOS a tanítókörpuszból egy nyelvmodellt állít elő, amelynek segítségével az újabb moldvai magyar szövegeket automatikusan egyértelműsíteni tudja az adott kontextusban legvalószínűbb ‘lemma – morfoszintaktikai címke’ párok kiválasztásával. A PurePOS-sal egyértelműsített szövegeket is szükséges még kézzel átnézni és helyenként javítani, de ezzel az előfeldolgozási lépéssel jelentősen felgyorsítható a munka. A morfológiai elemzés teljes folyamatát az alábbi ábra szemlélteti:



A korpuszban használt morfológiai kódok listája

Morfológiai kód	Leírás	Példa
Abl	ablativus (-tól)	<i>anyukámtól</i>
Acc	accusativus (-t)	<i>imádságot</i>
Ade	adessivus (-nál)	<i>mámánál</i>
Adj	melléknév	<i>szent</i>
Adj col	színnév	<i>fehér</i>
_Adj_i	-i melléknévképző	<i>kicsibőjt</i>
Adj nat	népnév, nyelvnév	<i>magyar</i>
Adj Pro	melléknévi névmás	<i>olyan</i>
Adj Pro Int	melléknévi kérdő névmás	<i>milyen</i>
Adj Pro Rel	melléknévi vonatkozó névmás	<i>amilyeneket</i>
_Adj_s	-s melléknévképző	<i>magyaros</i>
Adj_Ű	-Ű melléknévképző	<i>többmódú</i>
Adv	határozószó	<i>még</i>
Adv (Adj)	melléknévi vonzatú határozószó	<i>túl</i>
Adv Pro	határozószói névmás	<i>mitőte</i>
Adv Pro Int	határozószói kérdő névmás	<i>miért</i>
Adv Pro Rel	határozószói vonatkozó névmás	<i>amikor</i>
All	allativus (-hoz)	<i>fehérepekhez</i>
APos	anaforikus birtokjel (-é)	<i>miénket</i>
C	kötőszó	<i>és</i>
Cau	causalis (-ért)	<i>valamiért</i>
Clit_is	az <i>is</i> elem (bármely funkciójában)	<i>is</i>
Cmp	középfok (felsőfokban is része az annotációnak,	<i>kicsibbek</i>

	a -bb végződés jele)	
Cond	feltételes mód	<i>imádkoznám</i>
Dat	dativus (-nAk), birtokos szerkezetekben is	<i>galuskának</i>
Def	definit (határozott tárgyas) alak	<i>kezdí</i>
Del	delativus (-rÓl)	<i>Kolozsvárról</i>
Det	determináns	<i>az</i>
Det Pro	névmási determináns	<i>ide</i>
Det Pro Int	kérdő névmási determináns	<i>mi</i>
Det Pro (PP)	névutó előtt álló névmási determináns (pl. <i>a mellett a kutya mellett</i>)	<i>a</i>
Det Pro Rel	vonatkozó névmási determináns	<i>amit</i>
Det Q	kvantordetermináns	<i>minden</i>
_Dim=cskA	-cskA kicsinyítőképző	<i>kölköcskék</i>
_Dim=kA	-kA kicsinyítőképző	<i>asztalkát</i>
Ela	elativus (-bÓl)	<i>levélből</i>
_Essmod	essivus modalis (-An, -Ul stb.)	<i>románul</i>
_Essmod=t	essivus modalis (-t)	<i>erőst</i>
_Essnum	essivus modalis (-An) számneveknél	<i>ötszázan</i>
Frac	törtszámnév (-Vd)	<i>negyed</i>
Ill	illativus (-bA)	<i>razbélbe</i>
Ine	inessivus (-bAn)	<i>Clujban</i>
Inf	infinitivus	<i>áldozni</i>
Inj/Utt	indulatszó/mondatszó	<i>ej</i>
Ins	instrumentalis (-vAl)	<i>holddal</i>
Ipf	elbeszélő múlt (-A)	<i>vala</i>
Mod	modális (-hAt)	<i>lehet</i>
_Mul	multiplikatív (-szOr)	<i>sokszor</i>
N	főnév	<i>parancsolat</i>
N mat	anyagnév	<i>arany</i>
N nat	népnév, nyelvnév	<i>amerikán</i>
N Pro	névszói csoport kategóriájú névmás	<i>én</i>
N Pro Int	névszói csoport kategóriájú kérdő névmás	<i>mit</i>
N Pro Rel	névszói csoport kategóriájú vonatkozó névmás	<i>mit (amit értelemben)</i>
Obj2	második személyű tárgy (-lAk)	<i>várlak</i>
Ord	sorszámnévképző (-Odik)	<i>harmadik</i>
P1	többes szám első személy	<i>imádkoztunk</i>
P1*	többes szám első személy, sztenderdtől eltérő használat	<i>felejsük (felejtjük értelemben)</i>
P2	többes szám második személy	<i>ütitek</i>
P3	többes szám harmadik személy	<i>vannak</i>
PartAdv	határozói igenév (-vA)	<i>lecsiptetve</i>
PartAdv=vÁn	határozói igenév (-vÁn)	<i>megvirradván</i>
PartPrs	folyamatos melléknév igenév (-Ó)	<i>álló</i>
Pass	passzív (szenvető) (-Atik, -tAtik)	<i>felnőtetett</i>
Past	múlt idő (-t)	<i>éltem</i>
Pl	többes szám (többes számú birtok is)	<i>menyekezőknél</i>
Pl=i	az i-ző birtokos alakokban	<i>kezit</i>

PP	névutó	<i>elé</i>
PxP1	többes szám 1. személyű birtokos	<i>időnkben</i>
PxP2	többes szám 2. személyű birtokos	<i>lábatokat</i>
PxP3	többes szám 3. személyű birtokos	<i>embereik</i>
PxS1	egyes szám 1. személyű birtokos	<i>tátám</i>
PxS2	egyes szám 2. személyű birtokos	<i>édesanyád</i>
PxS3	egyes szám 3. személyű birtokos	<i>ura</i>
Q	kvantor/számnév	<i>húsz</i>
Q Pro	számnévi névmás	<i>annyi</i>
Q Pro Int	számnévi kérdő névmás	<i>mennyit</i>
Q Pro Rel	számnévi vonatkozó névmás	<i>ennyit</i>
QPtl	kérdő partikula	<i>-e</i>
Q-Q	számnév-számnév	<i>egy-négy</i>
S1	egyes szám első személy	<i>tudok</i>
S2	egyes szám második személy	<i>ülsz</i>
S3	egyes szám harmadik személy	<i>jön</i>
S3*	egyes szám harmadik személy, sztenderdtől eltérő használat	<i>tudhassa</i>
SPUNCT	mondatzáró írásjel	<i>?</i>
Sub	sublativus (-rA)	<i>nyakadra</i>
Subj	kötőmód/felszólító mód	<i>aludjál</i>
Sup	superessivus (-On)	<i>patakon</i>
Supl	felsőfok (a leg- jele)	<i>legkisebb</i>
Tem	temporalis (-kor(on))	<i>mosdáskor</i>
Ter	terminativusi esetrag (-ig)	<i>reggelig</i>
Tmp_ante	időbeli megelőzés (-jA)	<i>esztendeje</i>
Tmp_inl	speciális időhatározói alak	<i>ősszel</i>
UNK	a hangfelvételtől nem jól kihallható szó	<i>@@</i>
V	ige	<i>tudok</i>
_V_itrB	befejezett melléknév igenév (-O?tt)	<i>csöngöltetett</i>
VPfx	igekötő	<i>meg</i>
WPUNCT	mondatközi írásjel	<i>,</i>

Felhasznált szakirodalom

- Novák Attila (2014): A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavík, pp. 1068–1073
- Novák Attila – Siklósi Borbála – Oravecz Csaba (2016): A New Integrated Open-source Morphological Analyzer for Hungarian. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, pp. 1315–1322.
- Orosz György – Novák Attila (2013): PurePos 2.0: a hybrid tool for morphological disambiguation. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*. Hissar, pp. 539–545.
- Orosz György – Novák Attila (2012): PurePos – an open source morphological disambiguator. In: *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*. Wrocław, pp. 53–63.