Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

# Magyar ismeretlenszó-elemző program fejlesztése

Novák Attila[1,2], Nagy Viktor[2] és Oravecz Csaba[2]

[1] MorphoLogic Kft., Budapest
[2] MTA Nyelvtudományi Intézet, Budapest
{novak,nagyv,oravecz}@nytud.hu

**Kivonat** Nagy korpuszok számítógépes feldolgozása során elkerülhetetlenül belecütközünk abba a problémába, hogy a szövegekben szereplő szóalakok igen jelentős részét nem tudja a rendelkezésre álló morfológiai elemzőprogram elemezni, mert hiányzik az adatbázisából a szó töve. Ugyanakkor ezeknek az elemezhetetlen szóalakoknak a nagy része tartalmaz toldalékokat, ezért valamilyen módon ezeket is elemezni kell. Ennek a problémának a kezelésére olyan hibrid eljárást lehet alkalmazni, amely szimbolikus parciális morfológiai elemzőből és egy olyan statisztikai alapú eszközből áll, amely az első lépésben a szimbolikus ismeretlenszó-elemző által előállított hipotézisteret a kívánt mértékűre szűkíti.
**Kulcsszavak**: ismeretlenszó-elemzés, morfológiai elemzés, eloszlások hasonlósága, statisztikai egyértelműsítés

## 1. Bevezető

Jellemzően agglutinatív nyelvek számítógépes feldolgozása során a nyelvben előforduló lehetséges szóalakok igen magas száma miatt morfológiai elemzés alkalmazása gyakorlatilag megkerülhetetlen. Ez a lépés nem helyettesíthető egyszerű szótárból történő lekérdezéssel [1], hiszen egy ilyen szótárnak szinte az összes lehetséges szóalakot tartalmaznia kellene, ez pedig technológiailag kezelhetetlen lenne. Kézenfekvő megoldás egy morfológiai elemző eszközt alkalmazni, amely egy tőtárra támaszkodva képes az inflexiós, a produktív derivációs és szóösszetételi jelenségeket kezelni, valamint adott szóalakokhoz a tövüket hozzárendelni.

Nagy korpuszok számítógépes feldolgozása során viszont kikerülhetetlen az a probléma, hogy a szövegekben szereplő szóalakok igen jelentős részét nem tudja a rendelkezésre álló morfológiai elemzőprogram elemezni, mert hiányzik az adatbázisából a szó töve. Ugyanakkor ezeknek az elemezhetetlen szóalakoknak a nagy része tartalmaz toldalékokat, ezért valamilyen módon ezeket is elemezni kellene. Az ismeretlen szavak elemzését általában valamilyen sztochasztikus tanuló eljárásból származó modellel próbálják megoldani, amelyet tanító korpuszon fejlesztenek ki. Ez a modell aztán még kiegészíthető külső információval is, mint például a szókezdő nagybetű megléte [2]. Ezek az eljárások azonban, még akkor is, amikor igen nagy mennyiségű annotálatlan tanító anyagot képesek használni [3], nehézkesen alkalmazhatók magyar nyelvre, elsősorban a sokféle és hosszú toldalékszekvenciákból adódó „kevés adat" (*sparse data*) probléma miatt.

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

A dolgozat ezért egy már létező morfológiai elemzőn alapuló szimbolikus ismeretlenszó-elemző eljárást mutat be, amelyet nagy korpuszból nyert statisztikai információt használó modell egészít ki, melynek segítségével a szimbolikus ismeretlenszó-elemző által generált hipotézistér hatékonyan szűkíthető. A dolgozat a következőképpen épül fel. A 2. rész rövid leírást ad a morfológiai elemzőről. A 3. rész a szimbolikus ismeretlenszó-elemzőt tárgyalja, míg a 4. részben bemutatjuk a tesztelés illetve modellépítés során használt adatokat. Az 5. rész a különböző paraméterekkel futtatott teszteket írja le, és bemutatja, hogyan lehetséges a szimbolikus ismeretlenszó-elemző által generált elemzések számát nagy korpuszból nyert szóalak és szufixum statisztika segítségével csökkenteni. Rövid összefoglalás zárja a dolgozatot a 6. részben.

## 2. A morfológiai elemző

Bár a morfológiai elemzés kulcsfontosságú a természetes nyelvfeldolgozásban, különösen agglutinatív jellegű nyelvek esetében, a morfológiai elemzők mint különálló nyelvfeldolgozó eszközök kevés figyelmet kaptak az irodalomban, és legtöbbjük kereskedelmi termék[1]. Az általunk alkalmazott HuMOR („High speed Unification MORphology") morfológiai elemző szintén egy kereskedelmi cég terméke [6]. Az elemző klasszikus „egyed-és-elrendezés" stílusú elemzést [7] végez. A bemenő szóalakot morfémák sorozatára bontja, ahol mindegyikhez (i) egy *felszíni alakot* (amely a bemenő sztringben megjelenik), (ii) egy *lexikai alakot* (a morféma szótári alakját), illetve (iii) egy *kategória címkét* (amely tartalmazhat strukturált információt vagy lehet egy homogén címke) rendel. Az elemző alapvetően véges állapotú automataként ábrázolt reguláris szónyelvtannal rendelkezik, így egyszerű morfémalistaként adja meg a lehetséges elemzéseket, vagyis nem rendel belső szerkezetet az elemzett szóalakhoz. Ennek következményeként sokkal hatékonyabban tud működni, mint egy környezetfüggetlen elemző, mivel elkerüli a CF elemző által szükségképpen produkált szerkezeti többértelműségeket.

A bemeneti szóalakhoz az elemző az alábbi módon keresi a lehetséges elemzéseket. A lexikonban olyan morfémákat keres, amelyek felszíni alakja illeszkedik a bemenő szóalak kezdetére (illetve később a még elemzetlen részére). Az elemző lexikonja egyszerű morfémákat és morfémasorozatokat is tartalmaz, melyek rendhagyó tövek és toldalékszekvenciák eltárolt elemzései; ezek így egy lépésben azonosíthatók, növelve ezzel az elemzés hatékonyságát. A karakter alapú illeszkedési feltételen kívül az elemző minden lépésben kétfajta ellenőrzést végez, melynek segítségével a keresési tér korai szakaszban szűkíthető.

Egyrészt az éppen vizsgált morfémának a megelőzővel való lokális kompatibilitása kerül ellenőrzésre, másrészt arról kell megbizonyosodni, hogy a morféma olyan kategóriájú, amely együtt az eddig ellenőrzött résszel az adott nyelvben

[1] Érdemes megjegyezni azonban, hogy az utóbbi időben a komplex morfológiájú nyelvekre fejlesztett annotált nyelvi erőforrások megjelenése miatt egyre több figyelem jut erre a területre is (vö. a Xerox eszközöket [4], illetve egy független implementációért lásd pl. [5]).

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

lehetséges szókonstrukciót alkot. A globális szószerkezet-ellenőrzés azokra a morfémákra hajtódik végre, amelyekre a lokális ellenőrzés sikeres volt. A lehetséges szószerkezetek a HuMOR elemzőben kiterjesztett véges állapotú automataként reprezentáltak. Az elemző kimenetét az 1. ábra illusztrálja.

```
analyser>lehetőségekben
lehetőség[S_FN]+ek[I_PL]+ben[I_INE]
lehető[S_MN]+ség[D=FN_PROP]+ek[I_PL]+ben[I_INE]
lehet[S_IGE]+ő[D=MN_MIF]+ség[D=FN_PROP]+ek[I_PL]+ben[I_INE]
lesz[S_IGE]=le+hető[D=MN_HATO]+ség[D=FN_PROP]+ek[I_PL]+ben[I_INE]
```

FN = főnév MIF = melléknévi igenév S_ = tő
MN = melléknév PROP = MN→FN képző I_ = rag
INE= inesszívusz HATO = modális képző D_ = képző
PL = többes szám

**1. ábra.** A morfológiai elemző kimenete.

A morfémákat ' ' jel választja el, reprezentációjuk a lexikai alakkal kezdődik, melyet a kategória követ. Ha a felszíni alak különbözik a lexikaitól, az előbbi '=' jel után szerepel, egyébként nincs megadva. A kategóriacímkét a morféma morfológiai kategóriáját meghatározó prefixum előzi meg. Képzők esetén a képzett szó szófaját is megadja az elemzés.

Nyelvfeldolgozó feladatokban az elemző által szolgáltatott nagyon részletes elemzésre általában nincs szükség. Ezért egy szótövesítő eljárást kell alkalmazni, amely az adott szóalak tövét és inflexiós toldalékait azonosítja, oly módon, hogy az összetett szavak tagjai és a derivációs toldalékok a tő részeként szerepelnek és nem jelennek meg független elemként az elemzésben. Az 1. ábrában található elemzéseket a lemmatizáló egy elemzéssé vonja össze, ahogy azt a 2. ábra mutatja.

```
lemmatiser>lehetőségekben
lehetőség[FN][PL][INE]
```

**2. ábra.** A lemmatizáló kimenete.

## 3. A szimbolikus ismeretlenszó-elemző

Semmilyen robusztus és széles lefedettséget biztosító nyelvfeldolgozó eszközlánc nem tud hatékonyan működni olyan eljárás használata nélkül, amely a rendszer tudásbázisa által nem ismert nyelvi jelenet képes kezelni. Tipikus sztochasztikus szóalakszintű annotáló eszközök, pl. egyértelműsítők, jellemzően valamilyen

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

beépített toldalékelemző statisztikai modellt alkalmaznak. Magyar nyelvre azonban ezek a modellek a magas toldalékvariancia miatt nem adnak jó eredményt [8]. Ezért a jelen ismeretlenszó-elemző rendszer egy parciális szimbolikus elemzőn alapul, amely a lehetséges lemma plusz toldalék szekvenciákról a statisztikai modelleknél intelligensebb hipotéziseket képes generálni.

Más megközelítésektől eltérően a szimbolikus ismeretlenszó-elemző által felhasznált adat nem nagyszámú szóalak elemzése feletti általánosítás eredménye [9]. Magyar nyelvben ugyanis a nyílt szóosztályok tagjainak lehetséges toldalékolt alakjai túlságosan nagyszámúak ahhoz, hogy kezelhetőek legyenek ilyen általánosítás megtételéhez. E helyett az ismeretlenszó-elemző adatbázisa a normál morfológiai elemző építésénél használt nyelvtani leírásnak a nyílt szóosztályok minden lehetséges tővégződésére való alkalmazásával készült.

Az ismeretlenszó-elemző a nyílt szóosztályok (főnév, ige, melléknév) minden inflexiós toldaléksorozatát azonosítani tudja, és néhány nagyon produktív derivációs toldalék is elemezhető. Az ismeretlen szóalakok jelentős része lehet idegen szó, melyek nem követik a magyar kiejtés szerinti helyesírást, ezért bizonyos, az eredeti morfológiai elemzőben meglévő megszorítást az ismeretlenszó-elemzőben ki kellett iktatni, illetve gyengíteni kellett (pl. magánhangzó-harmónia). Azon fonológiai és ortográfiai megszorítások, melyeknek ezen rendhagyó helyesírású alakok is engedelmeskednek, részei maradtak az ismeretlenszó-elemző adatbázisának, és elemzéskor ellenőrződnek is.

Az elemző által megengedett igei alakok formája erősen korlátozott. Mivel a magyar igetövek osztálya zárt, minden új tőnek egyértelműen azonosítható végződése van, amely valamilyen produktív igeképzőt tartalmaz. Az elemző csak abban az esetben javasol igei elemzést, ha ilyen végződés kapcsolódik a (hipotetikus) tőhöz. Ez a lépés jelentősen csökkenti a lehetséges elemzések számát, de egyben feltételezi, hogy a morfológiai elemző ismerni a zárt tőosztály összes elemét.

Minthogy alapvetően ugyanazok az inflexiós toldalékok követhetik a főnévi és melléknévi (valamint számnévi) töveket, ezek csupán morfofonológia alapon történő megkülönböztetése gyakorlatilag lehetetlen. Ezért az ismeretlenszó-elemző adatbázisában nem tettünk különbséget főnévi és melléknévi tövek között. A számnevek zárt osztályt alkotnak, így a számnévi szufixumok sem kerültek be az adatbázisba. Azokban a (ritka) esetekben, ahol egyértelműen azonosítható a melléknévi toldalék, az elemző természetesen felismeri a helyes tőkategóriát, egyébként minden főnévi tövet ajánló elemzés egyben melléknévi tövet tartalmazó elemzésként is tekinthető. A főnévi kategória később felülírható, ha a szóalak melléknévnek bizonyul. Az ismeretlenszó-elemzőnek a lemmatizáló formátuma szerinti kimenetét a 3. ábra illusztrálja.

## 4. Az adatok

Az ismeretlenszó-elemző eszközlánc átfogó tesztelésához, illetve a tő- és szufixumeloszlások statisztikai modelljeinek felépítéséhez a Magyar Nemzeti Szövegtár [10] teljes anyaga szolgált nyelvi erőforrásul. A szöveg minimális előfeldolgozá-

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

```
guesser>Torgyán
Torgyán[FN][NOM]
Torgyá[FN][SUP]
Torgya[FN][SUP]
Torgy[FN][PSe3][SUP]
```

**3. ábra.** A szimbolikus ismeretlenszó-elemző kimenete.

son, tokenizáláson esett át, a speciális tokenosztályok külön kezelése nélkül. Ezt az „először nézzük, milyen morfológiai információt hordoz egy token" megközelítést a magyarban az indokolja, hogy a legkülönfélébb típusú tokenek, mint például rövidítések, tulajdonnevek, címek, tisztségek mind toldalékolhatók, ezért a speciálisan ezek kezelésére kifejlesztett nyelvfeldolgozó moduloknak is hozzá kell férniük a morfológiai információhoz.

A korpusz anyaga gyakorisági lista alakjában szolgált a morfológiai elemző (lemmatizáló) (ME) bemenetéül. Az 1. táblázat tartalmazza a morfológiai elemzés főbb adatait. Az ismeretlen alakokat ezután az ismeretlenszó-elemző dolgozta

**1. táblázat.** A morfológiai elemzés összefoglaló adatai.

| Egységek | ME által elemzett | Ismeretlen | Összesen |
|---|---|---|---|
| Szóalak típus | 2.222.280 (69.06%) | 995.396 (30.94%) | 3.217.676 |
| Szóalak token | 125.319.357 (95.50%) | 5.907.372 (4.50%) | 131.226.729 |

fel, amely minden egyes az ME által elemzetlenül hagyott alakhoz hozzárendelte a lehetséges elemzésük listáját. Az ismeretlenszó-elemző összesen 2.360.845 elemzést adott meg a 995.396 szóalakhoz, ami 2,37 elemzés/token átlagnak felel meg. Ez az érték jelentősen magasabb, mint az ME hasonló értéke (ahol 3.065.988 elemzés tartozott 2.222.280 szóalakhoz, 1,38 elemzés/token átlaggal). A különbségnek alapvetően két oka van: egyrészt néhány az ME-ben jelenlévő megszorítás az ismeretlenszó-elemzőből ki lett iktatva az idegen szavak elemzésének elősegítése miatt, másrészt a lemmatizáló gyakran összevon elemzéseket, melyeket az ismeretlenszó-elemző nem. Az utóbbi ugyanis megpróbál minél több derivációs toldalékot és ezen keresztül minél több tövet azonosítani, hogy az elemzések rangsorolását és értékelését végző statisztikai módszerekhez kimerítő alapadatokat szolgáltathasson. Érdemes megjegyezni, hogy amennyiben a lehetséges igei elemzések nem lennének ilyen mértékben korlátozva, illetve a melléknévi elemzés is alapesetben bekerülhetne a lehetséges elemzések közé, a fenti 2,37-es átlag megközelítené az 5-öt.

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

## 5. Az elemző tesztelése és kiértékelése

A szimbolikus ismeretlenszó-elemző által generált hipotézisteret természetesen érdemes szűkíteni a valószínűtlen elemzések kizárásával illetve alacsonyra rangsorolásával. Ezzel kapcsolatban releváns információ nyerhető például a korpuszban található toldalékszekvencák eloszlásából, melynek alapján többféle tesztmodellt is lehet vizsgálni.

### 5.1. Tesztmodellek

Az *1a.*-val jelölt modellben a preferált elemzés kiválasztása az ismeretlenszó-elemző által javasolt tőnek a korpuszban mért előfordulási gyakorisága alapján történt. Tehát az az elemzés számított a helyesnek, ahol az elemzéshez rendelt szótő a legtöbbször fordult elő mint független szóalak a korpuszban. A gyakorisági adatok a szóalakok kisbetűsített formája alapján lettek kiszámítva. A 4. ábrában látható az ismeretlenszó-elemző kimenete, ahol az elemzések a tő gyakorisága szerint vannak súlyozva.

```
19957   Torgyán      Torgyán[FN][NOM] (19957)
                     Torgy[FN][PSe3][SUP] (0)
                     Torgyá[FN][SUP] (0)
                     Torgya[FN][SUP] (0)
1635    mindenképp   minden[FN][_KEPP] (175547)
                     mindenképp[FN][NOM] (1635)
598     Monde        Mond[FN][PSe3][NOM] (6792)
                     Monde[FN][NOM] (598)
```

**4. ábra.** Az ismeretlenszó-elemző kimenete az 1a. modellben.

Az *1b.* modell az előző kissé módosított változata, amennyiben egy szűrő ebben a modellben kizárt bizonyos elemzéseket, mielőtt azok az *1a.*-ban használt mérték (egyszerű tőfrekvencia) szerint rendezve lennének. A szűrő az alábbi módon működik: amennyiben az ME az ajánlott elemzéshez tartozó tövet egyébként tudta elemezni, de ezen elemzések között nincs olyan kategóriájú, amit az ismeretlenszó-elemző tulajdonított a javasolt tőnek (pl. a tőnek az ME szerint ige a kategóriája, viszont a javasolt elemzés főnévi kategóriát adna), akkor a kérdéses elemzést a szűrő kizárja. Az 5. ábra mutatja az ismeretlenszó-elemző szűrt és rangsorolt kimenetét. A *mond* alak igei tő az ME szerint, ezért a főnévi javasolt elemzést a szűrő kizárta.

A *2.* modell szintén figyelembe veszi az ME által szolgáltatott elemzéseket, de az *1b.*-ben alkalmazott szűrőn túl a kompatibilis elemzések tőkategóriája az ME által javasoltra íródott felül. A rangsorolás alapjául ebben a modellben nem az egyszerű tőalak gyakorisága szolgált, hanem a javasolt tő gyakorisága az ME általi elemzésekben. Azoknál a töveknél, amelyeket az ME nem elemzett, az előző

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

```
19957   Torgyán       Torgyán[FN][NOM] (19957)
                      Torgy[FN][PSe3][SUP] (0)
                      Torgyá[FN][SUP] (0)
                      Torgya[FN][SUP] (0)
1635    mindenképp    minden[FN][_KEPP] (175547)
                      mindenképp[FN][NOM] (1635)
598     Monde         Monde[FN][NOM] (598)
```

**5. ábra.** Az ismeretlenszó-elemző kimenete az 1b. modellben.

modellekhez hasonlóan a szóalakgyakoriság maradt a mutató. A feltételezés a 2. modell mögött az, hogy az ME számára ismeretlen szóalakok sokszor nem azért maradnak elemzetleniil, mert tövük hiányzik az ME adatbázisából, hanem vagy paradigmahiba van az ME-ben, vagy pedig az adott alak ortográfiája a tő–szufixum határon nem követi a szokásos eljárást (pl. kötőjel használatos ott, ahol egyébként nem szokás, vagy fordítva.). A 6. ábra illusztrálja a 2. modell kimenetét. A *mond* szótő ki van szűrve, és a *minden* tő gyakorisága megváltozott az előző modellekhez képest.

```
19957   Torgyán       Torgyán[FN][NOM] (19957)
                      Torgy[FN][PSe3][SUP] (0)
                      Torgyá[FN][SUP] (0)
                      Torgya[FN][SUP] (0)
1635    mindenképp    minden[FN|NM][_KEPP] (216310)
                      mindenképp[FN][NOM] (1635)
598     Monde         Monde[FN][NOM] (598)
```

**6. ábra.** Az ismeretlenszó-elemző kimenete a 2. modellben.

A 3. modell olyan hasonlósági mértéket használ fel, amely az ismeretlenszó-elemző által javasolt tövek hasonlóságát próbálja megragadni az adott kategória jellemző töveihez (vagyis a *főnéviség, igeiség stb.* mértékét). Ennek érdekében kiszámoltuk az ME által elemzett összes szóalak tövéhez kapcsolódó toldalékok eloszlását, és ezen eloszlásokat tőkategóriánként tároltuk. Megszámoltuk azon elemzéseket, melyek egy adott tőkategóriával kezdődtek, és ezeket az értékeket elosztottuk az adott kategória összes előfordulásával. Így minden kategóriára ($C'$) kaptunk egy normalizált hisztogramot ($H(C')$). Ugyancz az eljárást ismételtük meg az ismeretlenszó-elemző elemzéseire is, majd a javasolt tövek kategóriájának hisztogramját ($H(S_{C'})$) összehasonlítottuk a kompatibilis tövek teljes hisztogramjával ($H(C')$), és kiszámoltuk a két hisztogram abszolút különbségét ($AD(C', S_{C'})$). Ez a különbség egy 0 és 2 közötti szám ($AD(C', S_{C'}) \in [0, 2]$); 0, ha a két eloszlás azonos és 2, ha egyáltalán nincs közös toldalékksorozat. A hipotetikus tő kategória elemhez rendelt, a $C'$ eredeti tövekhez való hasonlóságot

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

kifejező mérték pedig a következő[2]: $SM(C', S_{C'}) = \frac{2-AD(C', S_{C'})}{2}$. Ezután egy $C'$ kategóriájú tövet tartalmazó elemzéshez rendelt mérőszám ($OM$) a $C$ kategóriájú tő gyakorisági értékének ($F(S_{C'})$) és a hasonlósági mértéknek a szorzata: $OM(C', S_{C'}) = SM(C', S_{C'})F(S_{C'})$.

Mint az 5.2. részben látható, ez a mérőszám nem bizonyult különösebben hatékonynak. Ez egyrészt a lemmatizáló és az ismeretlenszó-elemző működése közötti különbségből adódhat, ugyanis pl. a lemmatizáló általában nem ad vissza nominatívuszi főnév elemzést, ha a kérdéses szóalak derivációs toldalékra végződik, míg az ismeretlenszó-elemző igen, ezért pl. a nominatívuszi főnevek eloszlása jelentékenyen különbözik, ez pedig a melléknévi tőválasztást preferálja, ami sok hibához vezet. Másrészt azonban további vizsgálat szükséges annak érdekében, hogy milyen egyéb okok játszhatnak szerepet, illetőleg milyen más hasonlósági értékkel lenne érdemes számolni. A 3. modell kimenetét a 7. ábra mutatja.

```
19957   Torgyán         Torgyán[FN][NOM] (10100)
                        Torgyá[FN][SUP] (462)
                        Torgya[FN][SUP] (462)
                        Torgy[FN][PSe3][SUP] (218)
1635    mindenképp      mindenképp[FN][NOM] (679)
                        minden[FN][_KEPP] (6) 598
        Monde           Monde[FN][NOM] (338)
                        Mond[FN][PSe3][NOM] (58)
```

**7. ábra**. Az ismeretlenszó-elemző kimenete a 3. modellben.

## 5.2. Kiértékelés

Miután legjobb szándékunk ellenére sem találtunk általánosan elfogadott eljárást ismeretlenszó-elemzők teljesítményének kiértékelésére, az alábbi forgatókönyvet választottuk. Az ismeretlen szóalakok gyakorisági listájában megállapítottunk egy (önkényes) küszöbértéket (10), amely előfordulás alatt nem vettük figyelembe az adott tokent, kizárandó a nagy számú olyan „hulladék" alakot, amihez legfeljebb az *egyéb* elemzés lenne rendelhető — ezek igen nagy méretű korpuszokban elkerülhetetlenek. A maradék listát felosztottuk 10 egyenlő gyakorisági tartományra, és mindegyikből véletlenszerűen választottunk 100 alakot. Az eredményként kapott 1000 szavas listán értékeltük a modelleket pontosság szempontjából.[3]

[2] Lényegében ez az eljárás a két eloszlás különbségét az ún. $L_1$ normával méri.

[3] Ez egy egyszerűsített értékelés, amelyben a modellek minden alakhoz egy elemzést választanak, így külön *fedés* és *pontosság* értékek itt nem számolhatók. Ha a leírt eljárást szófaji egyértelműsítés kontextusában lexikális valószínűség értékek indukciójához használjuk — ez a jelen dolgozat témájának egyik lehetséges továbbfejlesztése —, akkor a különböző értékek már számolhatók.

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

Annak mérésére, hogy a korpuszból nyert statisztikai adatok mennyiben javítják a szimbolikus ismeretlenszó-elemző teljesítményét, két viszonyító alapmodellt is kiértékeltünk. A *0a.* modell az egyenlő valószínűségűnek tekintett javasolt elemzések közül véletlenszerűen választott, míg a *0b.* modell mindig a nominatívuszi főnév elemzést adta. A teljesítményre vonatkozó értékek a 2. táblázatban találhatók.

**2. táblázat**. Az elemző teljesítménye különböző statisztika modellekben.

| | Modell | | Teljesítmény | |
|---|---|---|---|---|
| | | | Típus | Token |
| korpuszadat nélkül | 0a | véletlen választás | 69.76% | 53.39% |
| | 0b | FN alanyeset | 78.09% | 88.72% |
| korpuszstatisztikával | 1a | tőfrekvencia | 84.18% | 91.89% |
| | 1b | szűrt tőfrekvencia | 84.61% | 92.73% |
| | 2 | hibrid frekvencia | 84.61% | 92.69% |
| | 3 | hisztogram összehasonlítás | 84.29% | 91.85% |

Minthogy az alanyesetű főnév nagyon gyakori az ismeretlen szavak között, már a *0b.* alapmodell is meglehetősen jól teljesített. Ugyanez a tendencia kiegészítve azzal, hogy az igék leggyakrabban jelen idő, egyes szám 3. személy kijelentő módban szerepelnek, eredményezi a minimális statisztikával támogatott *1a.* modell jó eredményét. A tövekre vonatkozó szűrés tovább csökkenti a hibák számát az *1b.* modellben. A 3. modell viszonylag gyenge teljesítményét az előző részben már említettük.

Az eredményeket sztenderd metodológia hiányában kissé nehézkes más hasonló próbálkozások eredményével összevetni. Alegria et al. [11] egy szófaji egyértelműsítő rendszer általános teljesítményét adja meg, amely ismeretlenszó-elemzést is használ (93%), míg Chanod és Tapanainen [12] az ittenihez hasonló kiértékelés szerint 85 %-os pontosságot ér el, bár meglehetősen szűk elemzési kódkészlettel (az általunk használt készlet több ezer lehetséges kódot tartalmaz).

## 6. Összefoglalás

Egy olyan ismeretlenszó-elemző rendszer kifejlesztését mutattuk be, amely szimbolikus megszorításokon alapuló részleges elemzőt egészít ki nagy korpuszból nyert olyan statisztikai információval, melynek segítségével az első lépésben előállított hipotézistér a kívánt mértékűre szűkíthető. A szimbolikus elemző és a statisztikai szűrő együttesét alapvetően két feladat ellátására látjuk alkalmasnak. Az egyik feladat a folyó szövegben előforduló ismeretlen szóalakok on-line elemzése és egyértelműsítése, a másik a morfológiai elemző adatbázisának bővítése, illetve javítása (off-line adatgyűjtés).

Novák Attila, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése.
In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, 45–57.

Az első feladat esetében a konkrét szóalakhoz egyetlen olyan elemzést kell kiválasztani, amely a szó tövét és morfoszintaktikai jegyeit (a tő és az inflexiós toldalékok kategóriáját) leírja. A másik feladat megoldásához olyan tövek kell a korpuszból kiválasztani, és a kategóriájukat megfelelően azonosítani, illetve esetleges egyéb megjósolhatatlan morfológiai tulajdonságaikat a korpuszban szereplő toldalékolt alakjaik segítségével megállapítani, amelyeket érdemes lenne a morfológiai elemző adatbázisába felvenni. A rendszer ezen két irányban történő alkalmazása jelenleg folyó kutatás tárgyát képezi.

## Hivatkozások

1. Hajič, J.: Morphological tagging: Data vs. Dictionaries. In: Proceedings of ANLP-NAACL Conference, Seattle, Washington, USA (2000) 94–101
2. Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L., , Palmucci, J.: Coping with ambiguity and unknown words through probabilistic models. Computational Linguistics **19** (1993) 359–382
3. Cucerzan, S., Yarowsky, D.: Language independent minimally supervised induction of lexical probabilities. In: Proceedings of the 38th Annual Meeting of the Associations for Computational Linguistics, Hong Kong (2000) 270–277
4. Karttunen, L.: Applications of finite-state transducers in natural language processing. In: Proceedings of CIAA-2000. Lecture Notes in Computer Science, Springer Verlag (2000)
5. Alegria, I., Aranzabe, M., Ezeiza, A., Ezeiza, N., Urizar, R.: Using finite state technology in natural language processing of Basque. In: Proceedings of the Conference on Implementations and Applications of Automata, Pretoria (2001) 2–12
6. Prószéky, G., Kis, B.: Morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, USA (1999) 261–268
7. Hockett, C.F.: Two models of grammatical description. Word **10** (1954) 210–234
8. Oravecz, Cs., Dienes, P.: Efficient stochastic part of speech tagging for Hungarian. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas (2002) 710–717
9. Daciuk, J.: Finite state tools for natural language processing. In: Proceedings of the COLING 2000 workshop Using Toolsets and Architectures to Build NLP Systems, Luxembourg, Luxembourg (2000) 34–37
10. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas (2002) 385–389
11. Alegria, I., Aranzabe, M., Ezeiza, A., Ezeiza, N., Urizar, R.: Robustness and customisation in an analyser/lemmatiser for Basque. In: Proceedings of the LREC-2002 Workshop on Customizing Knowledge in NLP Applications, Las Palmas (2002)
12. Chanod, J.P., Tapanainen, P.: Creating a tagset, lexicon and guesser for a French tagger. In Tzoukermann, E., Armstrong, S., szerk.: From Texts to Tags: Issues in Multilingual Language Analysis: Proceedings of the ACL SIGDAT Workshop, Geneva (1995) 58–64