Novák Attila, Wenszky Nóra (2007): Mire jó és hogyan készül egy számítógépes morfológia
In: Alberti Gábor, Fóris Ágota (eds.): A mai magyar formális nyelvtudomány műhelyei. Nemzeti Tankönyvkiadó, Budapest. 157–169.

Mire jó és hogyan készül egy számítógépes morfológia

Novák Attila – Wenszky Nóra
novak@morphologic.hu, nora@nytud.hu

Dolgozatunk célja, hogy bemutassuk a morfológiai elemzőprogramok létrehozását, működését és felhasználását. Egy rövid bevezető után bemutatjuk a nyelvtudomány történetének néhány olyan állomását, melyek a számítógépen is megvalósítható nyelvi modellek létrejöttében szerepet játszottak. Ezután rátérünk a számítógépes alkalmazásokra, és megpróbáljuk bemutatni, hogy mire jó és hogyan készül egy számítógépes morfológia. Konkrét példánk a MorphoLogic Kft. Humor elemzőprogramjának adatbázisa lesz.

Bevezetés

Ha megkérdezünk valakit arról, mit tanult nyelvtanból az iskolában (általános vagy középiskolában), leggyakrabban azt a választ kapjuk, hogy „Semmit". Esetleg a helyesírásról vagy a mondatelemzés feleslegességéről és kínjairól számol be a kérdezett. Sokszor a „semmit" válasz elég közel áll a valósághoz, mert nyelvtanórák helyett irodalomórát tartanak inkább. Így a nyelvészetet legtöbben haszontalan, értelmetlen tudománynak tekintik. Dolgozatunkban először röviden felvázoljuk, milyen fejlődésen ment keresztül a nyelvtudomány az utóbbi két évszázadban, majd részletesen bemutatunk néhány, a közember számára is hasznos számítógépes alkalmazást, melyek kifejlesztését a modern nyelvtudomány tette lehetővé.

Egy kis nyelvtudomány-történet

A nyelvészetet hagyományosan a bölcsészettudományok körébe utalják, és általában bölcsészkarokon oktatják az egyetemeken. A klasszikus görög (később latin) nyelvtanírási hagyományt követő hagyományos nyelvészet meglehetősen eklektikus módszertana valóban elüt a mai természettudományok egzakt módszereitől és modelljeitől (*Kálmán*, 2006). A hagyományos nyelvtanok a formai és jelentéstani szempontokat következetlenül keverik, az atominak tekintett szavakat merev és rosszul definiált kategóriákba skatulyázzák a nyelvtani jelenségek és a bennük részt vevő elemek leírása helyett. Nagyon sok pusztán intuícióra alapozott megkülönböztetést és összehasonlítást tesznek, bízva a nyelvtan olvasóinak nyelvérzékében, illetve nyelvtudásában. Az iskolai nyelvtantanítás kihasználja azt a helyzetet, hogy a tanulók kompetens beszélői a tárgyalt nyelvnek, így a leírások és a módszertan hiányos vagy ellentmondó volta nem feltétlenül tűnik fel, különösen, ha a problémás eseteket gondosan elkerülik.

A nyelvtudomány történetében a hagyományos paradigmától jelentősen eltérő módszerek olyankor jelentek meg, amikor a nyelvtan írója vagy nem hagyatkozhatott a saját, illetve célközönsége nyelvi intuíciójára, vagy más okból tekintette fontosnak, hogy olyan pontos leírást adjon a vizsgált nyelvről, amely minden fontos részletre kiterjed, és világos fogalmakra épül. Az alábbiakban *Robins* (1999) alapján áttekintjük a nyelvtudomány ezen állomásait.

Az első pontos nyelvtanok Indiában íródtak az i. e. I. évezredben a védikus szanszkritról, amelynek pontos leírása azért volt fontos számukra, mert a beszélt nyelvváltozatok a szent szövegek nyelvétől kezdtek erősen eltávolodni, és ez a folyamat egyre bizonytalanabbá tette a szövegek értelmezését. Bár ez a helyzet nem ütött el lényegesen attól, amelyben a görög grammatikusok a klasszikus görögről készült hagyományos nyelvtanjaikat írták, az indiai nyelvtudósok mégis a fonológia, a fonetika, a morfológiai és a szintaxis területén is messze meghaladták az európai nyelvészek eredményeit, és olyan pontos és ma is tökéletesen egyértelmű és részletes leírást adtak a szanszkrit nyelvről, amilyenhez foghatóval egyetlen európai ókori nyelvről sem rendelkezünk.

Az indiai nyelvészek felismerték, hogy a fonetika és fonológia terén a homályos akusztikai benyomások alapján nem lehetséges a beszédhangok pontos és rendszeres osztályozása (így volt ez legalábbis a hangszínkép-elemző eszközök kifejlesztéséig). Ezzel szemben a hangok artikulációs jegyei, a

beszédszervek mozgása a megfigyelés számára könnyen hozzáférhető, így pontos leírás alapjául szolgálhat.

A pusztán impresszionisztikus akusztikai fogalmakkal operáló hagyományos nyelvtan például csak annyit állapít meg a magyar toldalékolásban alapvető szerepet játszó hangrendi illeszkedésről, hogy vannak „mély" és „magas" hangrendű tövek (pl. *kutya*, illetve *egér*), és ezekhez „mély", illetve „magas" hangrendű toldalékok illeszkednek (pl. *kutyá+nak*, de *egér+nek*). A „mély" és „magas" fogalmaknak az artikulációs jegyeket használó terminológiában a magánhangzó képzési helyének vízszintes dimenziója az elöl-, illetve a hátulképzettség felel meg. Hogy a toldalék-magánhangzók illeszkedésének ezen kívül még két másik dimenziója is van, arról az iskolai nyelvtanban nem esik szó. A két másik dimenzió az elöl képzett magánhangzók, illetve tövek esetében az ajakkerekítés megléte vagy hiánya (ebben különbözik a *tejfel* a *tejföl*-től: *tejfel+es*, de *tejföl+ös*), valamint a nyitás, a toldalék-magánhangzó képzési helyének függőleges dimenziója (ebben különbözik a *harcos*, mint főnév a *harcos* melléknévtől: *harcos+ok*, de *harcos+ak*). Az utóbbi két artikulációs dimenziónak megfelelő akusztikai jellegű fogalmak nem is szerepelnek a hagyományos nyelvtan terminológiájában, de az iskolások nem panaszkodnak, mert úgyis tudnak magyarul.

Az indiai nyelvészeti eredmények Európa számára a XVIII.–XIX. század fordulóján váltak ismertté. Az indoeurópai nyelvcsaládon belüli rokoni kapcsolatok felfedezésének hatására a XIX. század nyelvészetét a nyelvtörténet és nyelvváltozás szabályszerűségeinek feltárására irányuló kutatások uralták, melynek módszerei már alapvetően természettudományos jellegűek voltak. A XX. században a nyelvtudomány érdeklődése ismét a szinkron leíró nyelvészet felé fordult.

A század első felében az amerikai strukturalista nyelvészek rengeteg bennszülött amerikai indián nyelvet írtak le. Az ezeket a nyelveket kutató nyelvészek nem beszélték a leírandó nyelveket, amik ráadásul szerkezetükben is igen jelentősen különböztek az angoltól vagy más, korábban részletesen vizsgált nyelvektől. Így a kutatók nem tudtak az anyanyelvi intuíciójukra hagyatkozni, nem állíthattak fel szemantikai kategóriákat – hiszen ilyen információk nem álltak rendelkezésükre. Az amerikai strukturalisták legnagyobb érdeme, hogy kidolgozták annak a módszertanát, hogy egy teljesen ismeretlen nyelvet hogyan lehet leírni, és a terepmunkás milyen módszerekkel bizonyosodhat meg arról, hogy az általa készített leírás helyes és pontos. Módszereik teljesen empirikus alapokon nyugodtak.

A XX. században a nyelvészet fejlődését alapvetően meghatározó másik tényező a számítógép feltalálása, majd elterjedése volt. A számítógépeket a második világháború idején először rejtjelezett szövegek megfejtésére használták, tehát az első alkalmazásuk is bizonyos értelemben nyelvi természetű volt. Azaz már a kezdetek kezdetén úgy gondolták, hogy ez az eszköz alkalmas lesz arra, hogy az emberi nyelvekkel kapcsolatos feladatok megoldására használják, hiszen képes arra, hogy szimbólumokat manipuláljon, hasonlóan a nyelvet használó emberhez. A számítógép azonban végképp olyan „célközönség" a nyelvtaníró számára, amely egyáltalán nem bír semmiféle nyelvi intuícióval. Például míg egy magyarul tudó ember az „*eszt a labdát*" szövegrészt olvasva azonnal tudja, hogy itt az „*ezt a labdát*" szövegről van szó (hiszen ezt így is mondjuk ki), addig egy helyesírás-ellenőrző program számára az *észt* szó éppolyan jó javaslat. Míg egy ember „úgy születik", hogy tudja, hogy a *kktp* nem olyan jó szó, mint a *baba*, a gép számára egyik sem jobb a másiknál, hacsak jól definiált eljárást nem adunk a számára az emberi nyelvekben lehetséges és gyakori szótagszerkezetek és a ritkák, illetve a soha elő nem fordulók megkülönböztetésére. A számítógép teljesen ész nélkül hajtja végre a beprogramozott procedúrát, és ha az hibás, akkor hibás lesz az eredmény is.

A gép nyelvi intuíciójának hiányában semmire sem megyünk a hagyományos nyelvtan – egyébként is téves – olyan definícióival sem, mint például hogy „az igék cselekvést vagy történést jelentő szavak", vagy hogy a „melléknevek tulajdonságokat jelentő szavak". A huszadik század második felének nyelvészetének egyik meghatározó szempontja volt, hogy számítógépen is implementálható nyelvleírásokat készítsenek. Ez nem jelenti azt, hogy minden modern „generatív" nyelvészeti keretben készült nyelvleírás alkalmas lenne arra, hogy a mindennapi gyakorlatban használható számítógépes programok készüljenek belőle. Sőt még csak azt sem, a többségük jó alapot nyújtana egy hatékony számítógépes nyelvi modell számára. De ennek ellenére szép számmal készültek olyan nyelvleírások, amelyek jól működő számítógépes alkalmazások alapjául szolgáltak.

A gépi fordítás – nagy ambíciók

A számítógépes nyelvészet az ötvenes–hatvanas években nagy ambíciókkal, és óriási erőforrások mozgósításával indult. A célkitűzés az automatikus gépi fordítás megvalósítása volt. Idővel világossá vált, hogy ez a célkitűzés nem megvalósítható olyan színvonalon, mint ahogy azt eredetileg elképzelték. Mára – óriási erőfeszítésekkel, gyakran több évtizedes fejlesztés eredményeképpen – sikerült létrehozni számos

Novák Attila, Wenszky Nóra (2007): Mire jó és hogyan készül egy számítógépes morfológia
In: Alberti Gábor, Fóris Ágota (eds.): A mai magyar formális nyelvtudomány műhelyei. Nemzeti Tankönyvkiadó, Budapest. 157–169.

fordítóprogramot, amelyek a közben nagyságrendekkel megnőtt sebességű és memóriakapacitású gépeken a piaci szempontból fontos nyelvek között fordítanak. Ám amit ezek a programok létrehoznak, az csak nyersfordításnak tekinthető, színvonaluk még ma is csak arra elegendő általában, hogy a forrásnyelven nem értő olvasó valami – szerencsés esetben nem nagyon homályos – képet alkosson arról, hogy miről szól a szöveg.

A problémát az okozza, hogy máig sem sikerült igazán jó formális modellt kidolgozni a nyelvi jelentés leírására, illetve mindannak a világismeretnek az ábrázolására, amelyre egy emberi fordító támaszkodik. A fordítóprogramok pusztán formális átalakítást végeznek a szövegen, nincs módjuk arra, hogy a forrásszöveg grammatikailag lehetséges elemzései közül kiválasszák az értelmeset. Nagyon érzékenyek a forrásszövegben levő hibákra, illetve képtelenek a beléjük épített nyelvtan és lexikon hiányainak okos áthidalására.

Egy sikertörténet: a számítógépes morfológia

Bár az emberi nyelvtudás, szövegértés és gondolkodás teljes formális modellezése még mindig távoli délibábnak tűnik, az emberi nyelvek számítógépes feldolgozásának bizonyos részterületein az utóbbi évtizedekben nemcsak jelentős eredmények születtek, de ezek egy része már jó ideje a mindennapi számítógép-felhasználók számára is hozzáférhető számítógépes alkalmazások része. Ilyen részterület a számítógépes morfológiáé.

A számítógépes morfológia feladata az (általában írott) szöveget alkotó szóalakok kezelése. A leggyakoribb feladat, hogy a szóalakokat fel kell ismerni (hogy melyik szó [lexéma] melyik alakjáról lehet szó), vagy – például fordítás esetén a célnyelven – a megfelelő szóalakot generálni kell.

A felismerés a morfológiai elemzőprogramok feladata: egy morfológiai elemző az adott nyelv szóalakjaihoz megadja a szó lehetséges elemzéseit a szövegkörnyezet esetleges egyértelműsítő hatásának figyelembevétele nélkül. Hogy az elemző milyen elemzéseket produkál, az attól függ, hogy milyen morfológiai modellen alapul a működése.

A nyelvészet története folyamán több különböző modellt dolgoztak ki a szóalaktan leírására. A klasszikus görög nyelvtani hagyományt követő hagyományos nyelvészet „őshonos” morfológiai modellje a *szó és paradigma* (Word-and-Paradigm) névre hallgat. Ez a modell a szavakat a nyelv atomi elemeinek tekinti, semmilyen belső szerkesztettséget nem feltételez (eltekintve attól, hogy a szavak beszédhangokból vagy betűkből állnak), egy szó különböző alakjai nem úgy viszonyulnak egymáshoz, mint egy tő különbözőképpen toldalékolt alakjai, hanem csak mint egy szó paradigmájának különböző tagjai.

A morfémát, mint a nyelvi szerkezetek szónál kisebb (de a beszédhangokkal ellentétben jelentéssel vagy legalábbis grammatikai szereppel bíró) építőelemét, az ókori indiai nyelvészek „találták fel”. A XX. század első felének strukturalista nyelvészei is a morfémát tekintették a megnyilatkozásokat felépítő alapvető nyelvi egységnek, és a nyelvész feladatának azt tekintették, hogy feltérképezze, hogy mik alkotják egy nyelv morfémakészletét, milyen alakváltozatai vannak az egyes morfémáknak (ezeket allomorfoknak nevezik), és hogy ezek az elemek milyen környezetekben fordulhatnak elő (mi a disztribúciójuk). A strukturalista morfológiamodellt *elem és elrendezés* (Item-and-Arrangement) morfológiának hívják.

Számítógépes morfológiák többsége alapvetően az utóbbi *elem és elrendezés* típusú modellen alapul. Egy ilyen elemző a szavak lehetséges (grammatikus) morfémasorozatokra bontásait keresi meg. Egy szóalak egy vagy több tövet (az utóbbi eset összetett szavak esetén fordul elő), és különböző grammatikai morfémákat (prefixumokat, képzőket, ragokat) tartalmazhat. Ezek együttesen megadják, hogy az adott szóalak melyik szótő paradigmájának melyik tagja lehet. Az alábbi példa azt mutatja be, hogy a MorphoLogic Kft. Humor elemzőprogramja milyen elemzéseket ad a *Fejetlenséget* szóra:

```
analyser>Fejetlenséget

fej@etlen@ség[S_FN]=Fejetlenség+et[I_ACC]
fej@etlen[S_MN]=Fejetlen+ség[D=FN_PROP]+et[I_ACC]
fej[S_FN]=Fej+etlen[D=MN_FFOSZT]+ség[D=FN_PROP]+et[I_ACC]
fej[S_IGE]=Fej+etlen[D=MN_IFOSZT]+ség[D=FN_PROP]+et[I_ACC]
```

A *fejetlenség* tő önállóan is szerepel az elemző tőtárában, mert annak ellenére, hogy ez egy minden magyar beszélő számára felismerhető módon több morfémából álló tő (a programnak az itt idézett változata a morfémahatárokat `@` jelekkel meg is jelöli), önálló lexikalizálódott jelentéssel bír (’káosz’), amely nem áll elő a szót alkotó morfémák jelentésének szokásos kombinációjaként. Azt is megtudjuk ebből az elemzésből, hogy a *fejetlenség* főnév (az `FN` címke utal erre), és hogy tő (a címke `S_` előtagja utal erre),

valamint hogy az elemzett szóalakban nagy F-fel és `@`-k nélkül szerepelt a tő. Ebben az elemzésben egy toldalék is szerepel még: a tárgyrag. Ennek alakja itt *et*, címkéje `ACC`, a tárgyeset latin és angol nevének (accusativus/accusative) rövidítése; a címke `I_` előtagja utal arra, hogy ez egy inflexiós végződés (rag).

A második elemzés töve, a *fejetlen*, ismét a lexikalizálódott jelentése okán szerepel (nem arról van szó, hogy valakinek nincs feje), ez melléknév (`MN`), és ebben az elemzésben egy képző is szerepel (erre utal a `D_` előtag, a képzés latin, ill. angol neve: derivatio/derivation). A *-ság/-ség* képző *-ség* alakban szerepel itt, a címkéjéből azt is megtudjuk, hogy ez egy főnévképző (`=FN`), a `PROP` címke pedig arra utal, hogy ez a tulajdonságnév-képző *-ság/-ség* (a *rendőrség* szóban (legalábbis annak 'rendőri testület' értelmében) pl. nem ez, hanem a csoportok megnevezésére szolgáló kollektívumképző *-ség* szerepel).

Az utolsó két elemzés, amelyekben a szót alkotó morfémák valóban teljesen produktív módon kombinálódnak, azt mutatja, hogy az előbbiek mellett a fejetlenség annak az állapotnak a megnevezésére is szolgálhat, hogy valakinek nincs feje, illetve – amire talán első hallásra nem is gondoltunk – egy olyan kellemetlen állapot is lehet, amelyben például egy tehén találhatja magát.

Helyesírás-ellenőrzés

A számítógépes morfológia (és egyben talán a számítógépes nyelvészet) legközismertebb alkalmazása a szövegszerkesztő és tördelőprogramokba integrált helyesírás-ellenőrző. Ennek az a feladata, hogy a helyesírási normának nem megfelelő szavakat megjelölje, és javaslatot tegyen arra, hogy az adott szót milyen helyes szavakra lehetne lecserélni.

A helyesírásra azért van szükség, hogy az egy adott nyelven írott szövegek egységes képet mutassanak (*Papp*, 1979), bár a magyar helyesírási norma azon vonása, hogy minden egyes szó esetében (még akkor is, ha a helyesírási alapelvek több írásmódot tennének lehetővé) csupán egyetlen helyes alakot fogad el, erősen vitatható. Emellett az is igaz, hogy napjainkban, amikor a számítógépek elterjedésével a nyomtatott vagy írásos termékek előállítása többé nem a nyomdák monopóliuma, egyre nagyobb szükség van arra, hogy egy jó helyesírás-ellenőrző program, amely a helyesírási normáknak megfelelő szövegek létrehozásában nyújt segítséget, mindenkinek a rendelkezésére álljon. Hiszen bár a könyvek előállítása még mindig nagyrészt a könyvkiadók és nyomdák feladata, a gondolatok publikálásának lehetőségei demokratizálódtak. Bárki készíthet honlapot, ahol akár közérdekű információkat is közzétehet. Ennek egyik jó példája a Meteorológiai Szolgálat időjárás-jelentése mellett párhuzamosan létrejött www.idokep.hu oldal. Ezen megtekinthetők az önkéntes adatgyűjtők által szolgáltatott aktuális időjárási adatok, sokkal részletesebben és sokkal inkább a pillanatnyi időjárási helyzetnek megfelelően, mint a hivatásos szervezet honlapján, amely a költségvetésében tátongó réseket oly módon igyekszik betömni, hogy a pontos és részletes időjárási adatokat csak az előfizetői számára teszi hozzáférhetővé. Ha egy ilyen hasznos, a viharok előrejelzésével akár milliós nagyságrendű károkat megelőzni képes internetes oldal tele lenne helyesírási hibával, valószínűleg jóval kevesebben vennék komolyan. Az írásbeli igényesség tehát fokmérője egy adott írásos dokumentum színvonalának – függetlenül annak információtartalmától.

A helyesírás-ellenőrzők egyik funkciója azon hibák kijavítása, amit egy helyesen nem jól író ember vét (pl. *hüje*). Ezek azok a típusú hibák, amit egy iskolai magyarfüzetben találunk – *j/ly* csere, betűkettőzések (*t/tt*), ékezethibák, egybe- vagy különírás. Ám ennél sokkal lényegesebb – mert gyakoribb – a spontán elütésekből származó hibák javítása, hiszen ezek a számítógép használata mellett óhatatlanul előfordulnak. Ezek a hibák más jellegűek – például betűfelcserélés, betűkihagyás, nagybetű/kisbetű csere.

Bár történtek kísérletek arra, hogy csupán szóalaklistán alapuló leírást használjanak különböző nyelvtechnológiai feladatok (elsősorban a helyesírás-ellenőrzés) megoldására, ezek a kísérletek még napjainkban is kudarcra vannak ítélve például a magyar esetében, annak ellenére, hogy több száz millió szóalakot tartalmazó elektronikus szövegtárak (korpuszok) állnak a kutatók rendelkezésére. Akármilyen hatalmas egy ilyen szövegadatbázis, a nem túl gyakori szavaknak még a gyakori formái is hiányozhatnak belőle. A 150 millió szavas Magyar Nemzeti Szövegtár vizsgálata (*Nagy Viktor* személyes közlése) azt mutatta, hogy az elméletileg lehetséges ragmorféma-kombinációk 60 százaléka egyáltalán nem jelent meg az adatbázisban (a produktív képzőket nem is tekintve). A hiányzó toldalékkombinációk nem rosszak, vagy furcsák, egyszerűen csak ritkák, ezért nem jelennek meg még egy ekkora adatbázisban sem. A Szószablya projektum keretében összegyűjtött 500 millió szóalakot tartalmazó Webkorpuszban is csak 50 százalék ez az arány.

Ezért a magyarhoz hasonlóan bonyolult morfológiájú nyelvek számítógépes feldolgozása elképzelhetetlen hatékony számítógépes szóalaktani modell nélkül. Ez elsősorban az agglutináló nyelvekre igaz, de flektáló nyelvek esetében is sokkal hatékonyabban lehet dolgozni egy ilyen modell segítségével, mint a lehetséges alakok puszta felsorolásából előálló listákkal. Az agglutináló nyelvekre az jellemző, hogy igen gyakoriak a toldaléksorozatok a szóalakokban (pl. *megehetnétek* = *meg+e+het+né+tek*). Mivel igen sok produktív végződés van és igen sokféle módon kombinálódhatnak (egyes Uráli nyelvekben egy

végződés a többi raghoz képest szabadon több lehetséges helyen is megjelenhet, például a Mari többesjel), a nyitott szóosztályok esetében az egy tőből előálló lehetséges szóalakok száma igen magas (akár több ezer).

Emellett például a magyarban az összetett szavak jelentős részét egybeírjuk, így az összetételek rekurzív módon lényegében nem korlátos elemszámú szóalakhalmazt adnak még véges lexikon esetén is. Tény, hogy egy összetett szavakat szabadon létrehozó nyelvtan jelentéstanilag furcsa alakokat is létrehoz (például *kéményfa, tintamadárkönyv*), a tapasztalat azt mutatja, hogy egy ily módon túlgeneráló programmal mégis jobban járunk, mint egy összetételeket nem megengedő változattal. A produktívan létrehozott összetételeket ugyanakkor sok alkalmazásban érdemes kiszűrni, ha más elemzések is vannak (pl. filmgyár+tó, feleség+ének, gyermek+ében, kisgyerek+ként stb.).

A helyesírás-ellenőrzés könnyebb meg nehezebb feladat is, mint az elemzés. Könnyebb, mert egyrészt nem kell elemzést adni, csak annyit, hogy jó-e a szó, így ha akár csak egy jó elemzést is találunk a szóra, az már elfogadható, nem kell azzal foglalkozni, hogy lenne-e még esetleg másik is. Ugyanakkor nehezebb is, mert az ismeretlen szavak helyett javasolni is kell valamit, és a felkínált javaslatok némelyike előtt a felhasználók általában értetlenül állnak. Ez a helyzet a magyar esetében menthetetlenül bekövetkezik a korábban említett nehezen értelmezhető ad hoc összetételek miatt még akkor is, ha a helyesírás-ellenőrző nyelvi adatbázisa egyébként megfelelő körültekintéssel készült, és nincsenek olyan szavak, amelyeket hibásan toldalékol.

A morfológiai elemző és a helyesírás-ellenőrző nem csak az általuk produkált kimenet információgazdagságában különböznek. Míg a helyesírás-ellenőrzővel szemben az az elvárás, hogy ne fogadjon el és ne javasoljon a helyesírási normáknak nem megfelelő szavakat, addig az elemző esetében kifejezetten előny, ha azokat a szóalakokat is meg tudja elemezni, amelyek valamely gyakori szó elterjedt, de a helyesírási szabályzatnak nem megfelelő alakjai (pl. *kefír, szervíz, csevely, főbelő* stb.).

A szavak szintjén túl

A helyesírás-ellenőrzésen kívül egy számítógépes morfológiának számos más – talán kevésbé közismert – alkalmazása is van.

A morfológiai elemző által adott elemzésekre szükség van minden olyan számítógépes alkalmazásban, amely a szónál magasabb szintű nyelvi egységek valamiféle feldolgozására vállalkozik. Egy kevésbé ambiciózus ilyen alkalmazás az egyes szövegszerkesztőkben szintén fellelhető mondatszintű nyelvhelyesség-ellenőrző eszköz, amely csak bizonyos gyakori helyesírási problémák (pl. a központozási hibák, összetételek különírása stb.) esetleges felismerésére vállalkozik, változó sikerrel. Persze mivel a szövegek jelentése ennek az eszköznek a számára is tökéletesen homályban marad, pusztán bizonyos mintákat ismer fel, javaslatai néha tévesek, esetleg még komikusak is, pl. amikor a *„magyar ember evés közben nem beszél”* helyett felismerve, hogy itt két főnév áll egymás mellett, amelyeket normális esetben összetett szóként egybe kellene írni, azt javasolja, hogy inkább írjuk ezt: *„magyar emberevés közben nem beszél”*.

Jóval ambiciózusabb vállalkozás a fordítóprogramok készítése. Ezek, mint említettük, egyelőre csak durva nyersfordításra képesek, és legfeljebb megértéstámogató eszköznek használhatóak. Mindazonáltal, minden gépi fordítórendszer alapvető komponense a számítógépes morfológia: a forrásnyelvi mondatok elemzésének alsó szintjét morfológiai elemző testesíti meg, a célnyelvi szöveg szóalakjainak előállításáról pedig szóalak-generátor gondoskodik. A fordítóprogramokban alkalmazott szóalak-generátor nyelvtana abban különbözik a megfelelő morfológiai elemzőétől, hogy míg azokban az esetekben, ahol egy szó paradigmájának valamelyik tagja alaki ingadozást mutat (pl. *gyere/jöjj/jöjjél, saras/sáros, tanítasz/tanítsz* stb.), az elemzővel szemben az az elvárás, hogy mindegyik változatot felismerje, a generátortól viszont azt várjuk, hogy csak a leggyakoribb (legjelöletlenebb) alakot generálja.

Dokumentumok indexelése és kategorizálása

A szövegfeldolgozással kapcsolatos számítógépes alkalmazások egy részében nincs szükség teljes morfológiai elemzésre, csak arra, hogy a szövegszavak lehetséges töveit megtaláljuk. Erre van szükség például a különböző dokumentum-osztályozó rendszerekben, amelyek bizonyos kulcsszavak alapján előre megadott csoportokba sorolják a szövegeket, például egy hírügynökségben a híreket. Itt is elkél némi körültekintés: egy szó grammatikailag lehetséges tövei közül némelyiket hiba lenne automatikus osztályozás alapjául használni: a *csecsen* szót tartalmazó szövegek általában sem a *nőgyógyászat* sem a *pornó* kategóriába nem tartoznak, és általában a *román* szóalak előfordulása sem utal arra, hogy a szöveg romákról szólna.

Egy másik alkalmazás, ahol a tövesítés hasznos lehet, a dokumentumok abból a célból való indexelése, hogy a bennük szereplő szavak alapján meg lehessen találni őket. Aki barangolt már a világhálón,

az tudja, hogy lényegében lehetetlen bármit is megtalálni valamilyen keresőszolgáltatás igénybevétele nélkül. A legnépszerűbb internetes keresőszolgáltatás, a *Google*, ugyan nem alkalmaz tövesítést a robotjai által bejárt lapok indexelésekor, de biztosak is lehetünk benne, hogy nem is talál meg egyetlen olyan releváns dokumentumot sem, amelyben a keresendő kifejezés csak toldalékolva fordul elő (ami nem ritka eset). Azokban az alkalmazásokban, ahol az ilyen esetekre is fel kell készülni, nem nélkülözhető a tövesítő alkalmazása. Ebben az esetben viszont persze azzal a problémával kell megküzdenünk, hogy a csecsenekről szóló szövegek is a találatok között lesznek, ha csecset keresünk. Itt ez talán kisebb probléma, mint az automatikus osztályozás esetében, mert a felhasználó könnyebben felismeri, hogy miért kapott téves találatot, ugyanakkor külön dolgoznia kell, hogy az ilyen általálatokat kiszűrje.

Szótárprogramok

Ugyancsak nagyon hasznos szolgáltatás a tövesítés az elektronikus szótárprogramokban. Így a szótár nemcsak a szótári alakban beírt szavakat tudja megkeresni, hanem a benne szereplő szavak bármely alakját felismerve meg tudja jeleníteni a megfelelő szócikket. A MorphoLogic MobiMouse szótárprogramja, amellyel úgy lehet szótárazni, hogy az egér mutatóját a kérdéses szóra állítjuk, és egy másodperc múlva egy buborékban megjelenik a szóhoz tartozó szócikk (feltéve, hogy a szó benne van a szótárban), lényegében használhatatlan lenne, ha a ragozott szavak tövét nem tudná megállapítani, hiszen például ebben a mondatban is a szavaknak majdnem a fele nem a szótári alakjában szerepel. A szótárprogram esetében is kevésbé kritikus, hogy olyan szócikkek is megjelennek, amelyek az adott körülmények között valószínűleg nem relevánsak (pl. a *csecs* szócikke a *csecsen* szóra, illetve a *roma* a *román*-ra), mert az emberi felhasználó nyelvi intuíciójára támaszkodva felismerheti ezeket az eseteket.

A számítógépes morfológia, mint a nyelvészeti kutatás segédeszköze

Egy számítógépes morfológia implementálása magának az adott nyelvvel foglalkozó nyelvészeti kutatásnak is rendkívül hatásos segédeszköze lehet. A kézzel írott nyelvtanokban általában számtalan részlet homályban marad. Ezeket a számítógépes nyelvtanban elkerülhetetlenül explicitté kell tenni, így a számítógépen ténylegesen implementált nyelvtanok jóval pontosabbak lehetnek, mint azok, amelyeket gépen soha nem próbáltak ki. A létrehozott elemző- és generálóeszközök megfelelő nyelvi adatok (korpusz vagy kikérdezhető anyanyelvi beszélők) megléte esetén a bennük implementált nyelvtan adekvátságának messzemenő tesztelését teszik lehetővé (hogy valóban helyesen és pontosan modellezi-e a nyelvi adatokat), olyan alapossággal, amely – különösen egy bonyolult fonológiájú és morfológiájú nyelv esetében – kézzel elképzelhetetlen. Mivel a gép „ész nélkül" dolgozik, minden hibát könyörtelenül kimutat, amivel szembetalálkozik.

Ugyanakkor ha rendelkezésre áll olyan korpusz, amelyen az elemzőt futtatni lehet, az így kapott morfológiailag annotált szövegeket a nyelv mondattanát tanulmányozó kutatók is eredményesen használhatják mondattani modelljeik megalkotásához és tesztelésére. Nemrégiben például egy kis Uráli nyelvekkel foglalkozó projektum keretében olyan nyelvekhez is készítettünk számítógépes morfológiákat, amelyek lényegében a kihalás szélén állnak, és még élő beszélőik valószínűleg soha nem fognak számítógépet használni. A számítógépes modell viszont abban is nagyon hatékony segítséget tud nyújtani ezeket az – egyébként sok szempontból elképesztően bonyolult – nyelveket kutató nyelvészeknek, hogy megtervezzék, hogy a terepen végzendő kutatás milyen kérdések tisztázására irányuljon, hogy minél pontosabb leírásunk legyen ezekről a nyelvekről, és ennek segítségével minél többet megőrizhessünk a világ sokszínű kulturális örökségéből az utókor számára.

Hogyan készül egy morfológiai elemző?

A morfológiai elemzőt készítő nyelvész munkája számos, jól körülhatárolható feladat elvégzéséből áll. Itt azt mutatjuk be, hogy a MorphoLogic Kft-ben kifejlesztett morfológiai elemzők[1] (*Novák*, 2003) adatbázisát létrehozó nyelvész milyen feladatokat kell, hogy elvégezzen ahhoz, hogy elkészüljön egy nyelv számítógépes morfológiája.

Egyrészt a nyelv morfémakategória-készletének leírása (szófajok, toldalékkategóriák) mellett fel kell térképezni a tő- és toldalékalternációkat. Ez azt jelenti, hogy meg kell állapítani, hogy mely morfémáknak van több alakja és hogy melyik változat milyen körülmények között (milyen környezetekben) fordul elő. Ha a váltakozásnak hangtani feltétele van, akkor közvetlenül ezekre a tulajdonságokra lehet

[1] Bár az itt bemutatott számítógépes morfológiai formalizmus, már több, mint öt éve elkészült, a nagyközönség által használt Microsoft-termékekben ma is a helyesírás-ellenőrző egy tíz évvel korábban kifejlesztett változata működik, amely nem az itt ismertetett módon készült.

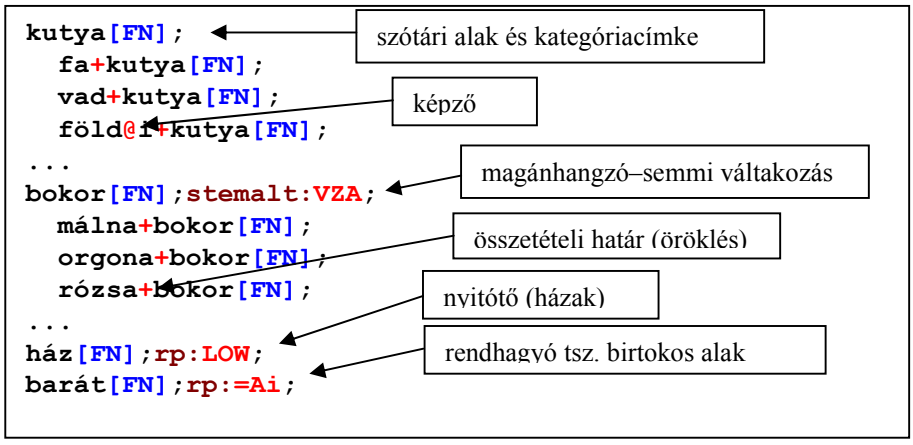
hivatkozni (pl. az, hogy a magyarban a középfok végződése kötőhanggal kapcsolódik-e a tőhöz, attól függ, hogy az mássalhangzóra vagy magánhangzóra végződik). Ha idioszinkratikus lexikai jegyek is szerepet játszanak (mint pl. a morféma által kiváltott pusztán hangtani okokkal nem magyarázható váltakozások esetében), akkor ezeket be kell vezetni. Például meg kell adni, hogy a *tél* típusú szótövekben mely toldalékok előtt van rövidülés *tel-* (pl. *telek, telet*) és mikor marad meg a hosszú magánhangzó (pl. *télen*, *télre*). Fel kell térképezni az összes olyan tulajdonságot, amely a nyelv morfológiájának leírásánál szerepet játszik. Ezek az előbb ismertetetteken kívül vonatkozhatnak a morféma kategóriájára is (például hogy tőről vagy képzőről van szó, illetve egy tő esetén a szófajra, a nyelvtani nemre vagy mondjuk szláv nyelvek esetén az élő–élettelen megkülönböztetésre).

Az így feltárt tulajdonságokra hivatkozva lehet a szomszédos szóelemek közötti szelekciós megszorításokat definiálni. Ez azt jelenti, hogy meg kell adni, milyen tulajdonságokkal rendelkező szóelemek kapcsolódhatnak össze. Például a névszói toldalékok (pl. a tárgyrag) csak névszókhoz kapcsolódhatnak, a hangrendi illeszkedést mutató toldalékok alakjai csak a megfelelő hangrendű tőhöz járulhatnak (a *-ség* toldalékalak pl. nem kapcsolódhat a *piros* tőhöz). A Humor elemző által használt modellben minden egyes szóelem (allomorf) két tulajdonsághalmazzal rendelkezik: az egyiket a vele balról, a másikat a vele jobbról szomszédos szóelemek „látják”. Hasonlóképpen minden szóelem megszorításokat tehet mind a vele balról mind a vele jobbról szomszédos morfémákra nézve. Egy szóelemet csak akkor követhet egy másik, ha mind a bal oldali szóelem jobbról látható tulajdonságegyüttese kielégíti a jobb oldalinak a bal szomszédjával szemben támasztott követelményeit, mind pedig a jobb szóelem balról látható tulajdonságegyüttese kielégíti a bal oldalinak a jobb szomszédjával szemben támasztott követelményeit.

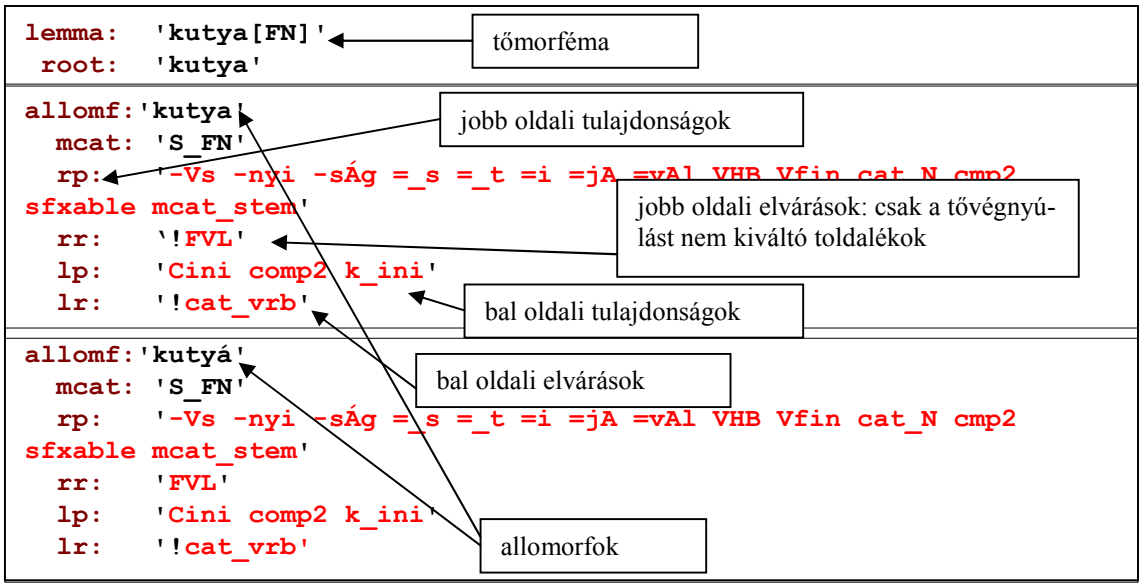
A morfológiai elemző a hatékony működés érdekében nagyon redundáns adatbázist használ: ez allomorfok (morféma-alakváltozatok) leírását tartalmazza mindazoknak a morfológiailag releváns tulajdonságoknak és megszorításoknak az explicit leírásával együtt, amelyek az adott elemre igazak. A nyelvész által létrehozott lexikonok ezzel szemben redundanciamentesek: nem allomorfok, hanem morfémák leírását tartalmazzák. A morfémákat a lexikai alakjuk, a kategóriájuk és a megjósolhatatlan vagy rendhagyó tulajdonságaik és elvárásaik megadásával kell leírni. A rendhagyó toldalékolt alakok és szuppletív allomorfok (ilyen pl. a *jön* ige *gyer-* tőalakja) is megadhatók a lexikonban.

Az összetett szavak konzisztens és gazdaságos leírásának elősegítésére beépítettünk a rendszerbe egy egyszerű öröklési mechanizmust, amelynek segítségével az összetett lexikai egységek alapesetben az utótagjuktól öröklik a tulajdonságaikat. Az öröklési mechanizmus működésének az a feltétele, hogy a szavakat az összetételi határok jelölésével kell a lexikonba felvenni. Tehát a *rózsa+bokor* szó ugyanúgy fog viselkedni, mint a *bokor* szó. Amennyiben egy összetett szó nem örökli az utótag tulajdonságait (pl. *szó–szavak*, de *névszó–névszók*), akkor ezt attól függően lehet kezelni, hogy a szó összetételi utótagként általában másképp viselkedik, mint önállóan (ez az eset a *szó* végű összetételek esetében), vagy csak egyedi kivételről van szó.

A fentiek illusztrálására tekintsük meg a következő részletet a tőtárból. Ez az az adatbázis, amit a nyelvész készít: minden tőhöz megadja a nem megjósolható információkat. Ezek közé tartozik a kategóriacímke (azaz szófaji megjelölés), a morfémákra bontás (képzők, összetételi határok), valamint az olyan jellemzők, mint a hangkivetés a *bokor–bokrok* esetében, a nyitótőség vagy a bármilyen más szempontból rendhagyó alak (pl. *barátai* de *barátja, házak* de *gázok*). Az ábrán ki nem fejtett rövidítések: `FN` – főnév, `stemalt` – tőalternáció, `VZA` – Vowel–Zero Alternation, `rp` – right property (jobb oldali tulajdonság).



Ebből a morfémalexikonból hozza létre egy szabályegyüttes azt az adatbázist, amelyet a morfológiai elemző fog használni. Ezek a szabályok írják le, hogy az allomorfok redundáns tulajdonságai hogyan számíthatók ki a már ismert (a lexikonban megadott vagy korábban már kiszámított) tulajdonságaikból (ide értve az alakjukat is). Az alábbiakban a fenti `kutya[FN]` tétel kifejtett változatát látjuk.



A szavak belső alaktani szerkezetére vonatkozó megszorításokat (ideértve a nem szomszédos szóelemek közötti megszorításokat is) külön szónyelvtan írja le. A szónyelvtan írja le például, hogy a töveket képzők majd ragok követhetik, más sorrendben ugyanezek a típusú elemek nem következhetnek (a *buta+ság+ot* jó szó, az *ot+ság+buta* nem). Az is a szónyelvtan része, hogy milyen összetettszó-szerkezeteket enged meg a helyesírás, például a [számnév]+[főnév]+[s képző] alakú szerkezeteket csak akkor engedi egybeírni a magyar helyesírási norma, ha a számnévi és a főnévi tag sem összetett.

Összefoglalás

A fentiekben azt próbáltuk meg illusztrálni a számítógépes morfológia néhány alkalmazásának bemutatásával, hogy a modern nyelvtudomány nem a világtól félrevonult tudósok önmagáért való időtöltése csupán, hanem olyan diszciplína, amely a számítógépes nyelvtechnológia tudományos hátterét adva a mindennapi gyakorlatban is hasznos eszközök létrehozását tette, és teszi lehetővé.

Irodalom

Kálmán László (2006): *Iskolai nyelvoktatás Antal László szellemében*. Előadás az „Antal László és a mai magyar nyelvtudomány” konferencián. Budapest, 2006. február 17.

Novák Attila (2003): Milyen a jó Humor? In: Alexin Zoltán – Csendes Dóra (szerk.) *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003).* Szegedi Tudományegyetem, 138–145.

Papp Ferenc (1979) *Könyv az orosz nyelvről.* Gondolat, Budapest.

Prószéky Gábor és Kis Balázs. 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 261–268. College Park, Maryland, USA

Robins, Robert Henry (1999): *A nyelvészet rövid története.* Osiris, Budapest.