MAGYAR NYELV

108. ÉVF. 2012. TAVASZ 1. SZÁM

A nyelvtechnológia és a magyar nyelvtudomány*

1. Sokan úgy gondolják, hogy a nyelv különféle bonyolult szerkezeteit is „értenie" kellene annak a gépnek, „aki" egy másodpercet sem töltött homo sapiensként ezen a földön. A számítógép pedig önmagában nem más, mint egy buta doboz, amely nagy sebességgel működik, bár a rajta futó programok adott esetben nagyon sok mindenben adhatnak segítséget a nyelvet beszélő embereknek. Sajnos, az újságok sokszor elektronikus agynak nevezik, amitől még napjaink számítógépe is igencsak távol van.

Ha a számítógép találkozik az emberi nyelvvel, az első kérdés, ami felmerül, az az, hogy az írott nyelvvel vagy a beszélt nyelvvel foglalkozik-e ez a terület? Ha jól meggondoljuk, a számítógép világában előbb-utóbb minden „írottá" válik, még a beszélt nyelv is, ugyanis az is betűknek, karaktereknek, szimbólumoknak diszkrét sorozatává válik, és ezt dolgozza fel a továbbiakban a számítógép. Tehát míg az embernél természetes a beszélt nyelv elsődlegessége, a számítógépnél ezeknek a diszkrét szimbólumoknak az egymásutánjából álló, tehát „írott" nyelvnek van elsődlegessége. Ennek a kutatási területnek népszerű elnevezése a *nyelv- és beszédtechnológia.* A közelmúltban ért véget hazákban egy jelentős méretű együttműködés az ezzel a témával foglalkozó intézmények között: ez volt a Nyelv- és Beszédtechnológiai Platform. Sajnálatos módon a nemzetközi használatban is a *language and speech* terjedt el, bár a *text and speech* helyesebb volna, tehát a *szöveg és beszéd.* Nyilvánvalóan a beszéddel való foglalkozás része a nyelvészetnek, viszont a beszédtechnológia elsődlegesen nem humán, hanem műszaki, mérnöki, fizikai, akusztikai tudományág. A jelen áttekintés elsősorban az írott nyelvre koncentrál, bár említést teszünk az olyan fontos és Magyarországon túl is kiemelkedő eredményekkel rendelkező kutatási területekről, mint a gépi beszédfelismerés, a gépi beszédkeltés vagy a beszélőfelismerés. Ezek részleteinek ismertetéséhez elsősorban nem nyelvészeti típusú háttérre volna szükség. A mélyebben érdeklődőknek ajánlható a gépi beszédfeldolgozásról a közelmúltban megjelent és nemzetközi mércével mérve is igen komoly összefoglaló (Németh–Olaszy 2010).

2. Amikor az ember elkezd olvasni tanulni, addigra már rengeteg beszélt információt hallott. Az ember nem is tud elkezdeni olvasni anélkül, hogy az olvasott anyag értelmezését megelőzően ne beszélt szövegeken edződött volna. Érdemes

* A Magyar Nyelvtudományi Társaság 2011. évi december 14-i közgyűlésén az ELTE BTK tanácstermében elhangzott előadás szerkesztett változata.

belegondolni siket embertársaink helyzetébe, akiknek azért lényegesen nehezebb megérteni az írott nyelvet, mert nekik nincs meg az a hangzó beszéd ismeretén alapuló hosszú tapasztalat, ami a hallók számára rendelkezésre áll. Ennél nyilván rosszabb helyzetben van a számítógép, mert annak aztán végképp nem lehet ilyen tapasztalata. Ha az írott nyelvvel foglalkozunk, akkor az egyik lehetséges megoldás, hogy nagy mennyiségű szöveget átadunk a gépnek, aztán segítjük valami módon, hogy megértsen valamit. Kérdés, hogy az elektronikusan elérhető magyar nyelvű szövegek mennyisége mire elegendő. Mennyi magyar nyelvű szöveg lehet elektronikusan egyáltalán? Korábban annyi volt egy adott nyelven a kutatáshoz felhasználható szöveg, amennyit mi magunk elolvastunk. Jó esetben nyolcvan-száz évig élünk, amit ez alatt hallunk, olvasunk, azok számítanak egy ember számára elérhető anyagnak. Napjainkban az interneten körülbelül milliárd szó nagyságrendű magyar nyelvű szöveg található. Persze ez csak becslés, de a gyakoriság, a statisztika rengeteg mindenben segít: az egy nyelvből az interneten fellelhető szavak számát például úgy kaphatjuk meg, hogy a nyelvtechnológus előveszi az adott nyelvre jellemző gyakori szavakat, és megnézi, hogy azok a vizsgálandó szövegekben milyen gyakran fordulnak elő (KILGARIFF–GREFFENSTETTE 2003). Az, hogy egy nyelvre jellemző egy szó, az azt jelenti, hogy olyan – két betűköz között leírt – betűsorozat, ami lehetőleg más nyelvre nem jellemző. Tehát például a *van* szót nem érdemes kulcsszónak venni a magyarban, ugyanis hiába gyakori, a holland nyelvben is gyakori. A németben a *der*, *die*, *das*-t lehet megfelelően gyakorinak mondani, de ezekből a *die* betűsor mégiscsak többször megjelenik az angolban is (igaz, más jelentésben). Ezeknek a szavaknak ismert az előfordulási gyakorisága az átlagos szövegekben, így ha az interneten megszámoljuk, hogy mennyi található meg belőlük, következtethetünk, hogy mekkora az a szöveg, amiben az adott ismert gyakoriság mellett ennyi előfordulás található. A magyar esetében a weben előforduló szavaink milliárdos száma már akkora mennyiség, amennyivel egyetlen ember egész életében nem találkozik: ha belegondolunk, hogy napi ezer-tízezer szót olvasunk, kiszámíthatjuk, hogy egy nyolcvanéves-százéves élettartam alatt a weben levő magyar szövegek mennyiségének a törtrészével sem találkozhatunk. Tehát a mai nyelvtechnológiai programok biztosan több szöveghez jutnak hozzá, mint egyetlen ember egész életében. És mégis, mit tudunk ezzel az őrületes méretű nyelvi anyaggal csinálni?

Érdemes tehát végiggondolni, hogy a nyelvi szerkezetek gyakorisága tud-e segíteni a nyelv géppel történő elsajátításában? Vegyünk egy egyszerű kísérletet: utánanézhetünk, hogy az interneten hogyan fordul elő az a hasonlító kifejezés, hogy *akkorát esett, mint...* Magyarul: feltehetjük a kérdést, hogy ez a kifejezés hogyan folytatódik a webes előfordulásokban, milyen szavakkal? Lehet erre valamilyen tippünk is előre, például a folytatás: *akkorát esett, mint egy ház*. Nyilván lesz, akinek esetleg az jut az eszébe, hogy *akkorát esett, mint egy ólajtó*, és még más variáció is felmerülhet. Végül 39 600 találatot jelzett a webes kereső arra, hogy *„akkorát esett, mint”*. Meglepő módon, az *ólajtó*-val együtt történt előfordulások száma ebből 20 800, tehát ennyiszer volt meg a weben az *akkorát esett, mint az ólajtó*. Úgy tűnik tehát, hogy a lehetséges előfordulásoknak kicsit több mint a fele az „ólajtós”. Intuíciónknak talán egy kicsit ellen mond,

hogy az *ólajtó* ennyivel gyakoribb, mint minden más, ezért megnéztük, milyen is az a teljes kontextus, amiben ez előfordul. Azt találtuk, hogy ebből a 20 800-ból 16 200-ban a teljes mondat így hangzik: *Stohl Buci akkorát esett, mint az ólajtó.* Magyarul: a bulvárnépszerűségnek örvendő Stohl neve a mondat többi részének a gyakoriságát is megsokszorozta, hiszen azt a hírt minden apróbb hazai internetes hírforrás is megismételte: az ismert ember kijött a színpadra, és elvágódott. Ez a hír a bulvárportálokon népszerű volt, ezért nyelvi szempontokhoz nem sok köze van az ólajtós hasonlító szerkezet internetes gyakoriságának, viszont Stohl bulvárismertségéhez annál inkább. Igen ám, de akkor ezek nem is különböző szövegek: a hírt lehozta egy főportál, és a helyi újságírók rögtön másolták is. Innentől kezdve azt mondhatnánk, hogy ez mindössze egyetlen előfordulás, akkor pedig mit számít, hogy tízezerszer van fent a weben? Lehet viszont arról az oldalról is tekinteni a kérdést, hogy mennyivel többet találkozik egy átlagos internetes olvasó ezzel a fordulattal, mint valamely másikkal, és akkor nem számít, hogy mi módon született egy ilyen mondat. A lényeg, hogy ha valami ezerszer annyi helyen fordul elő a weben, mint egy másik hasonló, akkor sokkal több ember fog vele találkozni. Tehát hiába tudjuk, hogy nyelvészetileg annak nincs jelentősége, hogy valami egyazon mondat pontos másolata; annak viszont van, hogy gyakran megjelenik, tehát a hatása is nagyobb. Ilyen módon működik a gyakoriság a weben.

3. Mindannyian sejtjük, hogy a normától eltérő szöveg nem túl szerencsés, de azt nehezebben tudnánk definiálni, hogy mi is maga ez a norma. Természetesen bármi is legyen, sok tantárgy – így a magyar nyelv – esetében is a normától való durva eltérésért az iskolában egyest is lehet kapni, esetleg még mások megbélyegzését vagy rosszalló tekintetét is kiérdemelni. A hagyományos módszerek viszont ma, az informatika korában már nem feltétlen segítenek valakit rávezetni a norma betartására, például egy egyszerű helyesírási szabályéra. Az már nem működik, hogy adott esetben a rádióban – ami korábban szinte az egyetlen szócsöve volt a széles tömegek nyelvi nevelésének –, valaki azt tanácsolja a hallgatónak, hogy valamit nem úgy kell csinálni, ahogy szokták, hanem máshogy. Ezzel egyidejűleg ugyanis milliók használják a számítógépet, aminek az írásra – és a helyesírásra – ma lényegesen nagyobb a hatása, mint a rádióműsoroknak. Húsz-harminc évvel ezelőtt, vagy még korábban, Lőrincze tanár úr korában az emberek figyeltek a rádió tanácsaira, ám nemcsak a tekintélyelv – ami egyébként nem nyelvi kérdés – változott meg az utóbbi évtizedekben, hanem egyszerűen kevésbé hatnak a kimondott szavak. Talán többeknek ismerős torzulás helyesírási gyakorlatunkban, hogy a hónapok kezdőbetűi az írott szövegekben nagybetűsödnek. Mondhatjuk, hogy nem kellene, hogy így legyen, de így van. Ennek az oka egyszerűen az, hogy van egy olyan számítógépes szoftver a legtöbb hazai számítógépen, amelyik úgy dolgozik, hogy pont után minden következő betűt nagybetűssé alakít. A külföldi szoftvergyártó úgy vélte, hogy ami pont után van, az mondatkezdet, tehát automatikusan nagybetűsítendő. Nem arról van tehát szó, hogy a felhasználó ismeri-e az ide vonatkozó szabályt vagy nem, hanem arról, hogy valahol valakik beállították a programot így. A magyar emberek egy részének fel sem tűnik, hogy ők maguk tulajdonképpen talán nem is nagybetűvel írták a hónap nevét, mégis az lett a szö-

vegben. Egy idő után viszont egyre többen látják, hogy mindenki naggyal írja, hogy *Május* meg *Szeptember.* Így aztán ma már sokkal több nagybetűs hónapnév forog közkézen, mint kisbetűs, mert aki ezt a programot használja, nem tudja, hogy ezt a szerencsétlen pont utáni nagybetűre vonatkozó beállítást le lehet tiltani. Nem elég tehát azt mondani az embereknek, hogy a *december* kis *d*-vel van, mert az nem ér semmit, amikor egy program automatikusan átjavítja nagybetűsre. Meg kell mondani a gépelőknek, hogy melyik az a gomb, amelyikkel ki lehet kapcsolni azt az opciót, hogy pont után ne automatikus nagybetűt tegyen be az a program. Ezzel tehát a nyelvművelés átkerült egy eddig nem művelt területre, a számítógépes beállítások ismeretének világába.

A számítógépes nyelvészet célja szinte sohasem az ideális beszélőnek a modellálása, amit a generatív grammatika olyan előszeretettel hangoztat – mert nagyon nehéz volna megfogalmazni, hogy milyen is az az ideális beszélő. Ideálisnak mondható megnyilatkozásokat nem találunk meg a weben. Itt valóban megjelenik az igazi performancia–kompetencia probléma: mindig csak konkrét megnyilvánulások vannak, és csak remélni lehet, hogy ezek a normákhoz hasonlóak. Ha a gépnek fel kell dolgoznia egy mondatot, nem tehet oda a hibásnak gondolt mondat elé egy csillagot, ahogy ezt a generatív grammatika teszi. Azt, hogy egy mondat nem jól formált, azt csak az ideális ismeretében mondhatom; így csak azt tudhatjuk, hogy nincs túl messze attól, amit normának gondolunk. Célunk a mondatok létrehozásával általában az, hogy megértsenek minket (vö. GRICE 1997). Más szavakkal azt is mondhatnánk, hogy van egy nyelvtől független „beállító gombunk", amit – átvitt értelemben – tekergethetünk, és attól függően, hogy kivel beszélgetünk, mindig olyan módon szólalunk meg, ami aktuálisan oda való. Bár egyetlen nyelvi rendszer van a fejünkben, de ha külföldivel, kisgyerekkel, beteggel vagy részeggel találkozunk, akkor „áttekerjük ezt a gombot, mint egy rádiókeresőt", és toleranciamértékünket megváltoztatjuk. Ezért aztán máshogy fogjuk elfogadni az egyébként ettől eltérő kontextusban nem elfogadható szöveget. A normától nagyon eltérőt is norma szerintinek vesszük: ha ezt a célt próbáljuk a számítógéppel segíteni, akkor egy adott szituációban azokat a mondatokat is meg kell tudnunk érteni, amiknek más esetben, úgy tűnik, se füle, se farka. Ha valaki rákeres az interneten, megnézheti, hogy mennyi normától eltérő alakot írnak le az emberek. Ez a szavak szintjén kezdődik: lényegesen több a weben a *csevej* szóból az *ly*-os, mint a *j*-vel írt. Mondhatjuk, hogy továbbra is *j*-vel kell írni, de rengetegen *ly*-nal írják. Nem tartják be a normát, de meg kell értenünk, amit leírtak. Talán ezek az emberek a *csermely* hangsort vélik áthallani, és a *csevej*-ben nem érzik meg a *röhej* meg a *zörej* analógiáját. Ilyenkor nekünk segítenünk kell azokkal a nyelvi eszközökkel, amit a nyelvtechnológia létrehozott, azaz a géppel működtetett szabályok vagy a statisztika módszereivel. Ez utóbbiról – bár a statisztikai megoldások nagyon divatosak a számítógépes nyelvészet utóbbi évtizedeiben – egyre jobban látszik, hogy nem mindenható. Ezért most a statisztikával reprezentálható antitézis után már a hibrid megoldásokra építő szintézis felé megyünk: lesznek olyan feladatok, ahol a statisztika sokat segít, de bizony hasznos az is, ha az intuíció alapján készített nyelvi szabályokat is segítségül hívjuk. Napjaink kérdése, hogy a számítógépes nyelvfeldolgozásban a statisztikát segítsük-e szabályokkal,

vagy a szabályokat egészítsük ki statisztikával. Nézzük kicsit részletesebben: mikor melyiket?

4. A magyar nyelvtudomány a szóalaktan területén sok olyan szabályszerűséget írt le, amelyeket a számítógépes nyelvészet jól föl tud használni. A szóalaktani szintről azt lehet mondani, hogy ami két betűköz között van, az egy gép számára számítógépes szóalaktani probléma forrása lehet. Tudjuk, hogy a szónak sokféle definíciója ismeretes a nyelvtudományban. Például a feltételes múlt idő nem feltétlenül „áll meg" a betűköznél, vagy az egyetlen szóból álló címszavakra építő korábbi értelmező kéziszótár egyik címszava a *száj- és körömfájás* volt. A számítástechnikában ismeretes az ún. kemény betűköz, ami a *száj-* és az *és* között, meg az *és* és a *körömfájás* között kellene, hogy álljon, jelezvén, hogy ezek nem „igazi" betűközök, hiszen a kifejezésben ezeken a pozíciókon semmilyen más nyelvi elem nem jelenhet meg. Tehát azt mondhatjuk, hogy a számítógépes morfológia olyan betűsorozatokat próbál meg nyelvileg elemezni, amelyek két betűköz között jelennek meg (PRÓSZÉKY 2000).

A morfológia mint önálló terület a számítógépes nyelvészetben az ún. kétszintes morfológiával jelent meg (KOSKENNIEMI 1983). A magyar nyelv egy kicsit „ellenállt" ennek a leírásnak, azaz nem sikerült igazán jó magyar kétszintes leírást adni a kutatóknak, de szerencsére volt több, a magyar hagyományokra jobban építkező irány is. Ennek alapján valósult meg például a hosszú ideig csak kéziratban létező ELEKFI-féle ragozási szótár (RagSz.) számítógépesítése (ELEKFI 1994). A hagyományos morfológiai leírásokban megbúvó szemlélet azonban több ponton is különbözik a számítógépes szemlélettől. A számítógép ugyanis olyan meghatározásokkal nem tud mit kezdeni, hogy „szinte kivétel nélkül", „az esetek túlnyomó többségében" vagy „gyakorlatilag mindig". A gépi leírások esetében pontos definíciókra van szükség. Ezért egy másik elv alapján indult el az 1980-as évek végén egy kutatás, melynek eredményeként létrejött a H u m o r számítógépes szóalaktani rendszer (PRÓSZÉKY 1994). A *Hu* előtag a módszer lelkét adó nagy sebességű unifikációs műveletre (h i g h - s p e e d u n i f i c a t i o n) utal, ahol az u n i f i k á c i ó egy matematikai művelet, melynek a segítségével azt írjuk le, hogy mi módon fogadhatja, illetve utasíthatja el egymást két egymást követő nyelvi elem egy adott szóalakban. A szófajok definícióiról hosszas vitákat lehet tartani (vö. KENESEI 2000), ám egy olyan rendszerben, amely – talán egy kicsit nagyképűen fogalmazva – analóg Mengyelejev kémiaielem-táblázatával, mindennek megvan a maga helye. Tehát minden elvileg lehetséges morfológiai viselkedéstípus ott van, legfeljebb nincs olyan magyar morf, amely egyikbe-másikba beleesnék. Az ötlet abban áll, hogy ha nem magukat a komplex morfológiai viselkedési osztályokat hozzuk létre (pl. hangzókieséses magyar igék osztálya), hanem a morfológiai viselkedés alapelemeit (pl. szófaj, hangrend, hangzókiesés) soroljuk fel, akkor ezek metszetéből kijönnek a nyelv lehetséges osztályai. Lesz ezek közt olyan osztály, amelyikbe mindössze néhány vagy akár egyetlen elem kerül, és lesz, ahová rengeteg. Amiben nincsen egy elem se, arra azt mondjuk, hogy aktuálisan a magyar nyelv ezt a morfológiai típust nem használja. Amely osztályba kevesen kerülnek, arra azt szoktuk hagyományosan mondani, hogy ezek a kivételek. Ezek általában több

szempontból is hasonlítanak egy másik osztályra, de van egy-egy olyan tulajdonságuk, amely alapján oda mégsem sorolhatjuk be őket. A nyelvészeti szakirodalom ilyenkor sokszor a lábjegyzethez nyúl, hogy az apró különbségről szóljon, ám a mi rendszerünk szempontjából teljesen mindegy, hogy például hetvenezren vannak-e egy ilyen osztályban vagy csak ketten: egyedül a viselkedési különbség számít. Nem érdekel minket tehát, hogy ebben vagy abban az osztályban kevesen vannak-e, és nem lenne-e mégis jobb, ha egy adott elemet egy másik, gazdagabban kitöltött osztályhoz sorolnánk, mert nem tehetjük: lesz ugyanis legalább egy olyan ismérv, ami alapján a mi elemünk különbözik az összes többitől. Tehát ha megvannak a szempontjaink, például az ajakkerekítés, akkor e szempont alapján egy magyar magánhangzó (illetve az ezt tartalmazó morféma) vagy ajakkerekítéses, vagy nem. Azt is megkérdezhetjük, hogy az illető morf elöl vagy hátul képzett magánhangzót tartalmaz-e, és így tovább. Ezeket a kérdéseket egyenként feltesszük, és ahány ilyen szempontunk van, annyi igen/nem választ kaphatunk. Ha például tíz kérdést teszünk föl igen/nem válasz formájában, akkor 2^{10}=1024 lehetséges válaszkombinációt kaphatunk erre a tíz kérdésre.

Töveknek egyébként a gépi feldolgozásnál azokat a morfokat mondjuk, amelyek kizárólag a jobb oldalról kapnak olyan toldalékokat, melyek valamilyen fonológiai/ortográfiai hatással vannak az előttük álló morfokra. Ha valami után garantáltan már nem jön semmi, azokat terminális toldalékoknak nevezzük: ilyenek például az esetragok. Ezeknek a záró elemeknek csak „baloldali arcuk" van, viszont léteznek olyan köztes elemek is (hagyományosan ezek képzők és jelek), melyeknek két „arcuk" van. Egy egyszerű képző bal felé úgy viselkedik, mint egy toldalék az előtte levő relatív tő után, jobb felé pedig mint egy relatív tő. Egy szó felépítésében minden elem vagy indító típusú, vagy terminális, vagy köztes. Lesznek olyan jegyek, hogy egy adott morf névszói-e, azon belül főnév-e, vagy hogy szótári alak-e, esetleg elöl vagy hátul képzett, ajakkerekítéses vagy sem, lehet-e neki többes száma, és így tovább. Például a *-nak* toldalék csak annyit kell mondjon magáról, hogy terminális elem, és hogy igényeket támaszt a baloldali morf, azaz az előtte álló relatív tő irányában, legyen az igazi tő vagy csak egy képző. Üzenetét úgy lehetne lefordítani, hogy „szeretném, ha az előttem álló morf névszói lenne, valamint szeretném, ha egyben hátul képzett is lenne, és elfogadná azt, hogy én dativus vagyok". Ez van tehát formális jegyek formájában belekódolva a *-nak* morf szótári leírásába. A *-nek* leírása majdnem ugyanezeket a jegyeket tartalmazza, csak mivel elöl képzett, ezért maga előtt is elöl képzett relatív tövet szeretne. A mindenkori relatív töveknek egyébként rengeteg tulajdonságuk van, de aktuálisan ezekből csak azzal kell foglalkozni egy konkrét szóalak elemzésekor, amelyik éppen megszólíttatik jobbról: ilyenkor meg kell nézni, hogy megfelelő jegyeinek értékei összeférnek-e a támasztott igénnyel. Érdemes még egy kicsit bonyolultabb példán, a *képzésnek* szóalakon megmutatni ezt a működést. Itt nyilván a *képez* igei forma a tő, azaz „nem névszó". Az *-ás*/*-és* nemterminális toldalék, és maga elé igei alakot, vagyis „nem névszó"-t kér. Van neki azonban egy másik, jobb oldali arca is, amely főnévi, hiszen az *-ás*/*-és* a további toldalékok irányából nézve már főnévi tövet takar. A morfológiai illeszkedés ellenőrzésére tehát ez az a nagyon egyszerű elv, amit igen hatékonyan lehet számítógépes programban

megfogalmazni. Egy ilyen elemzőprogram a mai gépeken másodpercenként mintegy százezer szót képes végigelemezni. Az elemzési hatékonyság megtartásához azonban azt igen fontos betartani, hogy ha valamilyen morfológiai viselkedésre utaló felszíni jel nagyon egyértelműnek is látszik a betűalakból, akkor sem szabad elemzéskor magával a betűalakkal foglalkoznia a programnak, hanem a betűalakból (vagy a nyelvész tudásából) korábban létrehozott jegyeket kell egy szótárból elővenni, és csak azokkal megvalósítani ezeket a jegy-összeférhetőségre vonatkozó – unifikációnak nevezett – műveleteket. A gépi feldolgozáshoz olykor kiegészítő információra is szükségünk lehet: van például olyan toldalék, amiből hagyományosan csak egyet ismerünk, ám az itt vázolt működéshez meg kell dupláznunk. Ilyen például az *-i* képző, ami a *ház-i-ak,* illetve a *kéz-i-ek* esetében is átviszi a hangrendet a túloldalra. Ilyenkor a *ház* alak nem tud átszólni az *-i* képzőn túlra, hogy én egy *-ak* típusú többes számot kérő tő vagyok: ezt az információt az *i*-nek kellene tudnia átadni, de akkor a *kéz* tő után egy másik *-i*-nek kellene állnia. A nyelvtechnológiai megoldások olykor tehát eltérnek a hagyományos nyelvészet megoldásaitól. Ez tehát nagy vonalakban a Humor elemzőrendszer alapja (l. PRÓSZÉKY 2000). Más közelítés is ismert a magyar számítógépes morfológiában, például a h u n m o r p h statisztikai alapon közelít (TRÓN et al. 2005). Ez utóbbi az interneten mindenki számára ingyenesen elérhető, magyar nyelvre is működő morfológiai elemzőprogram.

A nyelvtechnológiában van a morfológiához szorosan kapcsolódó, ám az elméleti nyelvészetben nem szereplő terület, ami elméleti nyelvészeti körökben magyarázatra szorul: a s z ó f a j i e g y é r t e l m ű s í t é s. Ez a kategória azért nem létezik a nyelvtudomány más területein, mert az ember számára egy többértelmű szó értelmezésekor mindig létezik egy olyan nyelvi szint, ahol csak egyetlen szófaji értelmezése van az illető szónak. A morfológiai többértelműségek kezelésében mindig segít a szintaxis, a szemantika vagy a pragmatika, vagy valami külső körülmény segítségével el tudjuk különíteni az egyik szófajt a másiktól. Az ember mindig felismeri, hogy például a *nyom* aktuálisan ige vagy éppen főnév. A számítógép, ha nem lép magasabb szintre az elemzésben, akkor a morfológiai feldolgozás után valamilyen „orákulum" segítségét kell kérje, hogy megtudja, ebben a pozícióban a *nyom* főnév, mert például előtte névelő, utána pedig egy állítás van, vagy a *nyom* ige, mert például igei toldalékkal folytatódik. A szófaji egyértelműsítés azért fontos a nyelvtechnológiában, mert nem minden számítógépes rendszernek van módja végigmenni ezen a nagyon sok lépcsős nyelvi hierarchián, így a szóalaktan szintjén kellene valakinek megsúgnia, hogy a sok lehetséges felbontásból melyiket fogadja el, hiszen nincs módja az összes lehetséges utat kipróbálni. Az ember számára egyébként semmi nem annyira többértelmű, mint a számítógép számára. A gépnek sokszor azok az esetek is többértelműek, amelyek az ember számára nem: mi nem gondolkozunk azon, hogy ha azt olvassuk, hogy *nemzetét*, akkor ez tulajdonképpen többértelmű: a *nemzete* tárgyesete, vagy a *nemzeté* tárgyesete, mert ott a kontextus, ahol ez a toldalékolt alak egyértelmű. A gép ezzel szemben végigelemzi a szót az összes lehetséges módon, és azután kiválasztja az aktuálisan legvalószínűbbet. A magyar szófaji egyértelműsítő módszerek kutatása több mint tízéves múltra néz vissza: MEGYESI (1999) Svéd-

országban, ORAVECZ és DIENES (2002) a Nyelvtudományi Intézetben, KUBA et al. (2004) Szegeden, HALÁCSY et al. (2006) a BME-n, valamint legutóbb OROSZ (2011) a PPKE-n adtak a magyar nyelvre különböző algoritmikus eljárásokat a szófaji egyértelműsítőknek a magyar morfológia szolgálatába állítására.

5. A magyar szóalaktan fent vázolt megoldásait alkalmazza a helyesírás-ellenőrzés, az automatikus szövegelválasztás, de adott esetben még az ékezetesítés is. A korábbi számítógépek, ma pedig elsősorban a mobiltelefonok világában ugyanis rengeteg *ekezet nelkuli uzenet* megy, és – egy norma szerint létrehozott szöveg esetében – szükség lehet arra, hogy egy program a megfelelő helyre tegye vissza az elhagyott ékezeteket. Persze mindig lesznek *fokabel = főkábel* vagy *fókabél* típusú szavak, amikről szövegkörnyezet nélkül nem tudhatjuk, hogy hova is valók az ékezetek, de azért ezek ritkák. A „hivatalos" magyar ékezetes betűk között nem szereplő zárt *ë* kezelésére is készült egy hasonló logikájú program (BUVÁRI 2001; NOVÁK–ENDRÉDY 2005).

A számítógépes szóalaktan alkalmazásai között talán még fontosabb a szöveges keresés támogatása, ahol nem mindegy, hogy megtaláljuk-e egy keresett szó toldalékolt formáit is, vagy sem. A szinonimaszótárak esetében – amelyek ugyan átvezetnek a következő témánkhoz, a szótárakhoz –, mint minden szótárban, tövek szerepelnek címszavakként, és ha a felhasználó egy szöveg közepén álló szóhoz keres rokon értelmű szót, akkor az a szó nagy eséllyel toldalékolt formában áll ott, tehát elemezni kell ahhoz, hogy megtaláljuk a tövét. Ezután ezt a tövet keressük meg a szinonimaszótárban és ezt cseréljük le egy másik tőre, méghozzá úgy, hogy az újonnan talált szó megfelelő alakját kell visszaírni a szövegbe. Igen, de ennek nagy eséllyel a morfo-fonológiai, morfo-ortográfiai tulajdonságai eltérnek, így itt is komoly szerepe van a számítógépes morfológiának, méghozzá itt a morfológiai generálásnak. A helyesírás-ellenőrző programok esetében is találkozunk hasonló problémával, hiszen ha egy szó morfológiailag nem elemezhető, lehet, hogy csak a normától eltérően van írva, a programtól viszont azt is elvárja a felhasználó, hogy állítsa elő a norma szerint megfelelő alakot. Ha tehát a szó a morfológiai elemző számára bármi okból nem ismert (akár mert hibás, akár mert például nem ismeri a tőszót a program, ami igen gyakori eset a tulajdonnevek világában), a mai elvárás az, hogy adjon javaslatot a hibásnak gondolt alak helyett. Ha például a nem ismert alak a *kérdésse*, akkor a programnak erre csak annyi tippje lesz, hogy ez vagy a *kérdése*, vagy a *kérdéses*, vagy a *kérdéssé*, esetleg a *kérdet* egyik ritka felszólító alakja, a *kérdesse* pontatlan leírásából származhat. Ezek az eredeti, ismeretlen betűsor lehetséges elütéstípusok szerinti technikai variációiból állnak elő úgy, hogy az elemzőprogram a létrejött „betűsalátákból" kiválasztja a formailag lehetségeseket, és egy listában a felhasználó számára választhatóvá teszi. Nyilván – a gyakorlatban zöld hullámos aláhúzásokkal jelentkező – környezetre is némiképp érzékeny program azt is figyelembe veheti, hogy a formailag lehetséges alakok közül melyik jöhet az adott helyen szóba. Például az *Ez esetben elszeretném a másik feleségét* mondatban az *elszeretném* forma lehet megfelelő, míg az *Ez esetben el szeretném adni az autómat* mondatban nem. Ilyenkor a program azt is meg tudja mondani, hogy a jelenlegi helyesírási sza-

bályzat melyik pontjában olvasható a probléma magyarázata, azaz jelen esetben a segédigék és az igekötős alak viszonya.

Az automatikus elválasztás ismét egy olyan terület, amelyhez szükség lehet a nyelvészek által régóta művelt morfológiára. Igaz, a magyar nyelv elválasztási szabályai szótagokra és betűkapcsolatokra hivatkoznak, de az elválasztást megelőzi egy magasabb nyelvi szint, a morfológia. Ugyanis az összetett szavak vagy az igekötős szavak elválasztásakor először ezek szegmentálását kell elvégeznünk, majd alkalmazhatjuk – szinte mechanikusan – az elválasztási szabályokat. Ha a morfológiai elemző minden összetételi határt felismer, az elválasztó algoritmus pedig csak akkor lép működésbe, ha teljesen biztos az elválasztási pozíció, azaz ha nincs egymásnak ellentmondó morfológiai felbontása az aktuális szóalaknak. Ebből következik, hogy az elválasztóprogram elvileg már nem tévedhet. Ha valaki mégis lát elválasztási hibát valahol a gyakorlatban, azt nem ennek a programnak a kontójára kell írni, hanem tudni kell, hogy más elven működő elválasztó programok is létezhetnek. Ha az elválasztandó szó például a *megint*, akkor az ember sem tudhatja, hogy mire gondolt, aki leírta, ha nincs hozzá szövegkörnyezet. A mindenkori felhasználónak kell majd eldöntenie, hogy *me-gint* vagy *meg-int*, de egy igényesen megírt automatikus elválasztóprogram nem segíthet azzal, hogy például gyakoribb az egyik felbontás, mert az elválasztás nem vezethet be egy újabb hibaforrást. Az ugyanis nem helyesírási hiba, ha valahol nincs elválasztás; legfeljebb nem szép. A rossz elválasztás viszont helyesírási hibának számít. Van még egy – tipográfiai jellegű – probléma is a magyar elválasztásban: a kettős hosszú mássalhangzóké. Az a szöveg- vagy kiadványszerkesztő program, amelyik szeretne elválasztást kérni, nem feltétlenül ismeri azt a problémát, hogy a magyar nyelvben, ha odaadjuk például az *asszony* szót az elválasztónak, akkor a visszakapott szó nem feltétlen ugyanannyi karakterből áll, mint a bemenő szó, ugyanis adott esetben beszúrhatunk egy betűt: *asz-szony.* Ha tehát azt kérdezné a hívó program, hogy hányadik betű után kell elválasztani a szót, akkor erre a kettős hosszú mássalhangzók esetében nem lehet válaszolni, ugyanis azok eredeti alakjukban nem lesznek elválasztva. Például az *asszonánc* esetén ott, hogy *assz-*, nem lehet elválasztani. Marad az a lehetőség, hogy *asszo-nánc*, ugyanis ott minden körülmények között lehet. A magyar elválasztóprogram pontosságával tehát igen közel vagyunk a 100 százalékhoz, ám még a mai napig is hallani olyat, hogy a magyar elválasztás egy tökéletesen soha meg nem oldható probléma.

Az előzőekben tehát igyekeztünk a legfontosabb olyan nyelvtechnológiai eszközöket bemutatni, amelyekben a számítógépes szóalaktan a magyar nyelvészet szóalaktanának eredményeire építve működik a gyakorlatban.

6. A hazai nyelvtechnológus kutatók egy ideje foglalkoznak a kisebb uráli nyelvek morfológiájával egy kutatási együttműködési program keretében, az MTA Nyelvtudományi Intézetével és a magyarországi uralisztikai tanszékek jelentős részével közösen munkálkodva. Elkészült a manysi, két – a szinjai és a kazimi – hanti, a komi, az udmurt, a nganaszan és tulajdonképpen egy tundrai nyenyec morfológiai elemzőprogram is. Ez azt jelenti, hogy ezeknek a nyelveknek a számítógépes szóalaktani feldolgozása magyar kutatók segítségével meglehetősen

előrehaladott állapotban van. Sőt, létrejött néhány nyelvre (például komira és udmurtra) egy helyesírás-ellenőrző program is, de nem volt olyan nagy világcég, amelyik azt mondta volna, hogy ez őt érdekli. Ezek a nyelvek ugyanis a nemzetközi gépi osztályozásban eddig még külön kódot sem kaptak, így egy szöveget nem lehet komira vagy udmurtra formázni, ahogy angolra vagy magyarra lehet. Bár az ezeken a nyelveken beszélők száma nagyobb, mint az izlandi vagy a máltai nyelven beszélőké, és létrehozásukkor nyilván nem az üzleti motiváció, hanem a tudományos volt az elsődleges. Ami viszont egy nyelv eddig le nem írt gépi morfológiájánál érdekes lehet, az a kérdésfeltevés jellege. Ugyanis a gépi leíráshoz sokszor olyan kérdéseket is fel kell tenni, amiket a hagyományos nyelvészeti leíráshoz nem kérdezünk meg. Így történt például a nganaszan nyelvhez készített morfológiai elemző esetében is (PRÓSZÉKY–NOVÁK 2005), ahol a szóalaktan és a szótagstruktúra nagyon komplex módon játszik össze. Ennek a nyelvnek az esetében is felmerül, hogy sajnos, nincs elég adat egyes jelenségek absztrakt leírásához. Mivel a nganaszan beszélők száma rohamosan csökken, nagy öröm, hogy a számítógépes rendszer már ismeri a nganaszan morfológiát, úgyhogy ha esetleg egyszer nem lesz anyanyelvi beszélő, akkor van egy program, ami képes ezt a nyelvi komplexitást produkálni. A Nyelvtudományi Intézet nyelvészei és a MorphoLogic kutatócég által létrehozott és ingyenesen használható nganaszan (és más uráli) elemzők mellett példaszövegeket is találni a rendszer és a nyelvek megismeréséhez[1]. Mivel az uráli nyelveket szóba hoztuk, feltétlen említést kell tenni a még készülőben levő számítógépes „Uráli etimológiai szótár" munkálatairól is, amelynek az eredményeként korábban elérhetetlen anyagok válnak széles körben elérhetővé[2].

7. A magyar lexikográfia rengeteg értékes eredménnyel gazdagította a nyelvtudományt. Napjainkban azonban az egynyelvű és a kétnyelvű szótárak világa is változóban van. A nyelvtechnológia, amelynek a hazai eredményeit itt sorra vesszük, valójában egy olyan új világot teremt, ami a Gutenberg-galaxis létrejötte óta kialakult papíralapú világ felváltása valami mással. Gutenberg óta a papír vastagsága vagy a felhasznált ólombetűk száma rengeteg helyen hatással volt a szótárak formájára. A szótári rövidítések azért alakultak ki, hogy minél több információ elférjen a papíron, hogy megfelelő könyvkötési technikákkal még ki tudjuk adni azt a szómennyiséget, amire szükségünk van. Az informatika, illetve az internet korában a helyszűke vagy az ólombetűk száma nem fog minket befolyásolni, tehát újfajta szótárak keletkezhetnek. Az első elektronikus szótárak – egy kicsit távoli analógiával – Benz autójához hasonlíthatók, amely azt a lovas kocsit vette mintának, amiből kifogták a lovat, és amibe a lovak helyére betették a motort. Ez tehát egy olyan „lovas kocsi" volt, amely már motorral ment. Hosszú idő után jutottunk el oda, hogy a szélcsatorna határozza meg az autónak a formáját, és nem a múlt. Az első számítógépes szótárak olyan papírszótárak voltak, amelyek ugyan a számítógép felületén jelentek meg, de még a régi Gutenberg-galaxisból

[1] http://www.morphologic.hu/urali
[2] http://www.uralonet.nytud.hu

örökölt tulajdonságaikkal. Ez a világ most van éppen átalakulóban az „igazi" elektronikus szótárak felé.

Ma, amikor már szinte közhely, hogy egyre több ember éri el az internetet, az a probléma, hogy az emberek nem tudják, hogy hol a határ a szótár meg a fordítás között. Ezért a valójában statikus szótárak és a dinamikusan működő fordítóprogramok közötti eltérések ismeretét ma már illik a nyelvészeti ismeretek alapjaihoz sorolni. A tömegek által használt internetes és mobiltelefonos szótárak használatáról pedig mérhető adataink is lehetnek. A szótárkészítésre való visszahatás soha nem látott módon tudja a szociolingvisztikát is gazdagítani. A felmérések több millió lekérdezés feldolgozása után már komoly vizsgálatokat tesznek lehetővé. Ezekből kiderülhet, hogy az iskolai mobilos szótárhasználat nyolc óra és kettő óra között az átlagos angoldolgozatok szavaira irányul, ebédidőben viszont az ételek nevei kerülnek előtérbe. Ebből az is kikövetkeztethető, hogy melyik az a korosztály, amely ezeket az eszközöket gyakorlatilag folyamatosan használja. Amikor megjelent egy, a fiatalok életét jelentősen befolyásoló könyv vagy film (pl. a Harry Potter-történetek, A Gyűrűk Ura, a Mátrix, vagy akár egy meghatározó zenekar lemeze), megjelennek a szótári keresésekben azok a szavak is, amelyek ezekben megtalálhatók. Mondhatjuk, hogy alakulóban egy új hazai műfaj, a „valós idejű társas nyelvészet", melynek nyelvtechnológiai eszközökkel való kutatása sok érdekességet tartalmaz. Például a vizsgálatok kimutatták, hogy ha az embereket valami foglalkoztatja, akkor olyan szavakat írnak be, amelyek valamilyen szemantikus értelemben egy szűkebb területről valók. Vagyis anélkül, hogy a szótárlekérdezők szándékosan tennének bármit is, az általuk egyetlen alkalommal bevitt szavakat összekapcsoljuk, ez a gráf több millió lekérdezés után egy olyan hálózatot ad, melyben a szemantikusan összetartozó szavak erősen összekapcsolódnak, azaz egyfajta szemantikus hálót (mai népszerű nevével: ontológiát) definiálnak. Kétségtelenül érdekes az is, hogy két nagy felhasználói világ különül el a gépi lekérdezések esetében: azoké, akik valóban a szavak jelentését szeretnék tudni, és azoké, akik elsősorban büszkélkednek az eszközhasználattal (vagy annak egy-egy extrém formájával) a barátaiknál. Igaz, ez az ismeret is a társas nyelvészeti kutatásokat gazdagítja.

8. Gépi segédeszközök és a meglevő szövegek felhasználásával egyre több új típusú szótár jön létre. A hagyományos egy-, két és többnyelvű szótárak világát jól egészíti ki a számítógépes felhasználásra és adott esetben emberi olvasásra is használható fogalmi szótárak, a már korábban is említett ontológiák világa. Ilyen hatalmas fogalmi szótár, a Princeton Egyetemen készített WordNet adatbázis (MILLER et al. 2000), amely több mint százezer nyelvi egység között definiál fogalmi viszonyokat. A magyar nyelv WordNethez kapcsolásával foglalkozó első kísérletek a 2000-es évek elején indultak el; az eljárás mögött az a hipotézis állt, hogy a WordNet-rendszerben kódolt relációk többé-kevésbé nyelvfüggetlenek, ezért tehát, ha a rendszer csomópontjain álló lexikai elemekhez találunk magyar megfelelőt, a köztük lévő fogalmi kapcsolat az angol WordNetből egyszerűen átörökíthető. A kísérleteket egy több hazai kutatóhely által koordinált kutatás követte,

amelyből 2007-re elkészült a Magyar WordNet szemantikus háló, egy 38 ezer magyar szóból álló adatbázis (PRÓSZÉKY–MIHÁLTZ 2008).

Bár most a hazai számítógépes nyelvészeti kutatásnak a magyar nyelvtudomány eredményeihez fűződő viszonyát tárgyaljuk, mégis érdemes egy kis kitérőt tenni a hazai szótárhelyzet rövid jellemzésére. A szótárak egy részét ugyanis üzleti alapon nem lehet felújítani, mert ez nem igazán jövedelmező terület. Sokszor csak lappang az a lehetőség, hogy megújuljon egy szótár, de legnagyobb szótárkiadóink üzleti alapon dolgoznak, vagyis ha nem látják garantálva a hasznot, eltekintenek a kiadástól. Arra is van példa, hogy már létrehozott, konkrét, nagyméretű szótárak azért nem jelennek meg, mert nincs, aki kiadja. Azt pedig végképp nem lehet üzleti alapon megoldani, hogy legyen a régió nyelveit összekötő, de üzletileg valószínűleg nem elég „erős" szlovák–magyar vagy magyar–román szótár. A gépi nyelvészet nagyszerű technológiákat hozott létre, a hazai szótárkészítésben is ott lehetnének ezek az eredmények, ám a globális választ mégsem a nyelvtechnológiának kellene megadnia erre a kérdésre. (Ezt a kérdést részletesebben tárgyalja PRÓSZÉKY 2011.)

A gépi nyelvészet eredményei ott különösen jól látszanak, ahol a statikus szótárak „működni kezdenek". Mit jelent ez? Amikor a szótárhoz fordulunk, leggyakrabban valamilyen szöveget olvasunk, és eközben kérünk segítséget. Eddig ezt a kiinduló szöveget csak az ember maga látta, ma viszont erre a szövegre a számítógépnek is van „rálátása", hiszen ami a gép képernyőjén megjelenik, annak a nyelvi elemzését is el tudja végezni. Így összeállhatnak akár a mondatban szintaktikai okok miatt szétesett kifejezések, akár a távoli igekötős igék, de valójában minden szó a saját szövegkörnyezetével azt is segít meghatározni, hogy melyik értelemben használjuk most őket. A jelentéskiválasztás eddig kizárólag a szöveget értő emberre maradt, most viszont a dinamikus morfo-szintaktikai elemzést már a nyelvi programok vállalhatják magukra. Ha például a *panel* szó nem érthető számomra, viszont a program észreveszi, hogy *panel* a *control* szó után áll a szövegben, akkor annak a *panel*-nak a *control panel* szócikkben lesz meg a megfelelő jelentése. Vagy gondoljunk a magyar és a német igekötőkre: hány embert „tréfál meg" az, hogy nem tudja, hogy az igéhez valahol a mondatban hozzátartozik egy igekötő is. Mivel egy ilyen intelligens szótárprogram[3] magától fel tudja fedezni ezeket az összefüggéseket, innentől kezdve ez már nem szótári kérdés, hanem számítógépes nyelvészeti, ami a nyelvtudomány eddigi más (morfológiai, szintaktikai) eredményeit kombinálja a lexikográfiaiakkal.

Ahogy korábban jeleztük, az elektronikus szótárnak különböznie kell a papírszótártól. Például az elektronikus szótárakban a célnyelv szerint is lehet keresni. Ez tehát nem a mechanikus ábécérend, hanem tartalmi keresés. Mit jelent ez? Ha például megnézzük, hogy egy német–magyar szótárban az *eszik* hol jelenik meg, akkor különböző találatokat kapunk, ahol az *eszik* megjelenik a német szó valamely ekvivalenseként: *essen*, *manfen*, *fressen* és egy kifejezés, a *sich etwas zu Gemüde fühlen* „jobboldalán" áll ott az *eszik.* Csakhogy ilyenkor a magyar szó a saját szinonimái között szerepel, nem pedig ábécérendbeli társai mellett. A cím-

[3] http://www.morphologic.hu/A-MoBiMouse-6-hasznalata.html

szavakat papíron csak ábécé-rendben lehet egyértelműen rendezni, viszont ha a jobb oldali tartalomban tudok keresni, akkor sokkal további részletes információt kapok a szó igazi jelentéséről, mintha csak egyszerűen föl lenne sorolva ábécében a találati lista. A gépi technológiák elterjedése előtt ez a lehetőség eddig ki volt zárva, vagyis itt egy újabb példa arra, hogy miben különbözik a Gutenberg-galaxis és az elektronikus.

9. Eddigi gondolatmenetünk egyenesen vezet az eredeti szövegek közvetlen felhasználásához a gépi nyelvészet területén. Ez a terület a korpusznyelvészet, melynek segítségével a szövegek elemzéséből közvetlenül, azaz nem a nyelvészet összefoglaló segítségével jutunk nyelvi jellegű információkhoz. A szövegkorpuszokból való kiindulás nem a gépi nyelvészet sajátja, csak a számítógép lehetővé tette a szövegek korábban nem látott méretekben történő (százmillió, sőt, milliárd szavas szövegkorpuszok) gyors felhasználását. Az alapgondolatot megtalálhatjuk már a 19. század végén, lényegesen a számítógépek megjelenése előtt: „Simonyi új grammatikai módszert akar behozni, könyve induktíve halad, azaz a példákból kiindulva tanítja a szabályt, nem pedig dogmatice. A grammatikát tehát valami olvasmány alapján akarja előadni úgy, hogy a szabályokat a tanár tanítványai közreműködésével vonhatja le ésszerű következtetések útján. Ilyenképp tehát ezen módszer véget vet a lelketlen magolásnak, és azt észfejlesztő indukcióval pótolja” (Riedl 1882). Simonyi valamit megérzett, és nyilván emberekről beszél még, akik képesek a feladatot megoldani, amit mi most a számítógépre vonatkoztatunk. Az utóbbi évtizedekben kialakultak hazánkban is a korpuszalapú és különösképpen a korpuszvezérelt nyelvészeti kutatások, ahol a nyelvészeti feladat továbbra is a nyelvészé, de a számítógép különféle összefüggéseket tár fel nekünk, és segít abban, hogy ezeket értelmezzük. A hazai korpusznyelvészet eddigi legkidolgozottabb szövegkorpusza az MTA Nyelvtudományi Intézete által mintegy tíz éve létrehozott Magyar Nemzeti Szövegtár. Ebben ma 187,5 millió szónyi szöveg van, a szépirodalomtól a joganyagokig, de hétköznapi szövegek és sajtószövegek is. Ráadásul a szövegek morfológiailag elemezve és egyértelműsítve is vannak. Az anyag az idők folyamán több mint húszmillió határon túli szóval is kiegészült, ami további új és jelentős kutatások alapjául szolgálhat.

A szövegtár anyaga felett működik egy olyan vonzat-, illetve kollokációkereső eszköz (Sass 2009), amelynek segítségével megtudható, hogy mely szavak milyen más szavakkal állnak tipikusan együtt. Ebből számszerűleg is kimutatható, hogy például *kérni* elsősorban nem tárgyakat szoktunk, hanem *bocsánatot, segítséget, elnézést, engedélyt, tájékoztatást, támogatást* és így tovább. Az általános nyelvészeti eredményeken túl itt is szociolingvisztikai ismeretekkel gazdagodhatunk, ha megvizsgáljuk, hogy az *ad* ige tárgyesetben álló bővítményei közül például a Magyar Nemzetben az *otthont ad* vagy a *hírt ad* sokkal gyakoribb, mint az *igazat ad* vagy *tippet ad*, ami meg az interneten jellemzőbb. Ennek a kutatásnak az egyik legfontosabb eredménye az igéket vonzatkereteikkel ábrázoló szótár (Sass et al. 2011). Tehát azt, amit mindannyian intuitíve érzünk, meg lehet nézni konkrét előfordulási gyakoriságaikkal. Hiába érezzük, hogy például az *emelkedik* és a *növekedik* szemantikusan rokon szavak, hamar kiderül, hogy *emelkedni* lehet *jog-*

erőre is, de *növekedni* nem lehet ugyanerre. Tehát szemantikus osztályzás keletkezik anélkül, hogy a szemantika explicit módon itt jelen lenne, pusztán a nyelvtechnológiai megoldások segítségével a program feltárja azokat a kategóriákat, amelyeket úgy intuitíve olykor mi is tudunk, de eddig nem állt mögötte százmillió szavas írott anyag.

Fontos hazai szövegkorpusz még a Szeged Korpusz, amit a Szegedi Tudományegyetem, az MTA Nyelvtudományi Intézete és a MorphoLogic hozott létre. Itt már a mondatszintig elemzett anyagról van szó. Hat különböző terület kétszázezer szavas, összesen egymillió-kétszázezer szavas szövege, amiből egy statisztikai számítógépes program tanulni tud. Léteznek továbbá nem egynyelvű korpuszok is, amelyeket a szakirodalom párhuzamos korpuszoknak mond. Magyarországon a legnagyobb ilyen a Hunglish, amit a BME-n dolgoztak ki (HALÁCSY et al. 2004). Ez egy szinkronizált magyar–angol korpusz, amivel többek között komoly fordítástudományi kutatásokat is lehet végezni, vagy akár csak ellenőrizni azt, hogy egy adott kifejezést a fordítók mindig konzisztensen fordítanak-e le vagy sem. Olyasmit is meg lehet nézni benne, hogy például milyen kontextusban fordítják az *egyetem*-et *college*-nak és mikor *university*-nek, és így tovább. Tehát a párhuzamos korpusz olyan lehetőségeket nyújt, amire korábban csak egy-egy nyelvtanár tudott volna válaszolni az intuíciója alapján, ma viszont ez az anyag még fordítóprogramok betanításához is használható.

Hatalmas hírkorpuszokon működnek például az olyan, ún. szövegosztályozó programok, melyek segítenek például beosztani a hírügynökségekhez az újonnan érkezett híreket. Egy olyan mondat, hogy *Nem lett cselgáncselnök a soproni futballtulajdonos,* a HiTec algoritmus (TIKK et al. 2005) alapján valamilyen biztonsággal a „Sport” kategóriába kerül, míg az *Obama megszakította nyaralását* a „Nagyvilág” kategóriába tartozik, és nem a „Bűnügyek”-hez. Lehet, hogy ha nem Obama neve állt volna ott, hanem egy bűnözőnek a neve, ez a program a mondatot rendőrségi hírként ismerte volna föl. Nagy mennyiségű szövegek elemzése esetén ma már a nyelvi programok ezekből a szövegekből tanulni képesek. Léteznek továbbá szövegekből terminusokat kivonatoló programok is, melyek az ismeretlen szakkifejezéseket hivatottak kigyűjteni.

10. Az eddig emlegetett kutatások a köznyelvi normát valamilyen értelemben jól betartó szövegekkel dolgoznak. Nyilván a nyelvjárástanban is lehet alkalmazni azonban a számítógépet. A hazai számítógépes dialektológiai kutatások alapjainak éppen napjainkban történő megteremtése elsősorban VÉKÁS DOMOKOS és VARGHA FRUZSINA SÁRA nevéhez köthető. A Bihalbocs program hatására a dialektológia nagyon hosszú ideig számítógépesen mostohán kezelt területe kvantitatívvá vált, azaz gépileg is adatolt lett, és összekapcsolódik mindenfajta földrajzi megjelenítő eszközzel. Például a nyelvterület térképén azonnal megmutatható, hogy egy olyan szó, mint a *bihalbocs* hol használatos, és hogy ugyanerre a fogalomra merrefelé járja inkább a *bivalybocs,* a *borjú,* a *kisbivaly* vagy a *bivalyfiú*. Meg lehet mutatni, hogy a *fésű* szó első szótagjának a magánhangzója szerint a Kárpát-medence különböző részein merre *fe*^{i}*sű*, *fé*^{i}*sű, fö*^{i}*sű* vagy épp *fisű* ejtésű ez a szó. A program az adatokat rögtön térképre tudja vetíteni, amivel hatékonyan támogathatja a kutatót (VÉKÁS 2000). Fontos azonban meg-

jegyezni, hogy a számítógépes nyelvészetnek ez az a területe, ahol a gépi eszköz – a korábban ismertetett kutatások eredményeivel szemben – nem a nem-nyelvész végfelhasználót támogatja, hanem magát a nyelvész kutatót.

11. A magyar számítógépes nyelvészet egyik legnagyobb kihívása annak a kérdésnek a megválaszolása, hogy hogyan elemzendő egy magyar mondat. A világban sokféle mondattani elemző program létezik, de a nyelvünkre viszonylag kevés olyan, átfogónak mondható szintaktikai leírás készült, amelynek gépi alkalmazása is van. Az egyik ilyen a MetaMorpho fordítórendszernek a mondatelemzője (PRÓSZÉKY–TIHANYI–UGRAY 2004), amelyik úgy elemzi a mondatokat, hogy az eredményt még a továbbiakban egy másik nyelvre is le kell tudnia fordítani. Az efféle program rátalál olyan megoldásokra is, amelyekre az ember ritkán, mert az ember érti a mondatot, és nemcsak szolgai módon elemzi. Egy angol nyelvű példa, a híres CHOMSKY-mondat jól illusztrálja ezt. Ugyanis a *Time flies like an arrow* esetében van két olyan elemzési lehetőség is, amelyet az ember a gyakorlatban nem vesz észre: nemcsak az a lehetséges fordítás tehát, hogy *Az idő repül, mint egy nyíl*, hanem az is, hogy *Időzíts legyeket, mint egy nyíl (Time flies!)*, sőt az is, hogy *Az időlegyek kedvelnek egy nyilat (időlegyek = time flies)*.

Lehet, hogy a géppel fordított szövegek nagy részét mi, anyanyelvi beszélők nem így mondanánk, de mindannyian tudjuk, hogy körülbelül ez az, amit a mai fordítóprogramoktól elvárhatunk. Végül is, ha valakinek szüksége van a gép fordítására, megérti (VARGA 2012). Másként fogalmazva: nem feltétlenül fordításról kell itt beszélni, hanem megértéstámogatásról (PRÓSZÉKY 2002). Az ilyen fordításokkal nem mennek el fordításhitelesítőkhöz az emberek, viszont nagyon gyors (és olcsó) eredményt kapnak. Nyilvánvaló, hogy a gépi fordítóprogramok működésének következtében senki nem vesztette el fordítói állását. A gépi fordítóprogramoknak más a célja, tehát nem úgy kell megítélni, mint ami az emberi fordítást kiváltja. Természetesen, ha például az angol forrásnyelvet albánra cserélnénk, akkor már többen éreznék a problémát: nem olyan könnyű hirtelen albán fordítót találni, de még olyat sem, aki képes egy adott albán szöveg hozzávetőleges értelmezésében segíteni. Magyarország egyébként még mindig az utolsó az Eurobarometer[4] szerint idegennyelv-ismeretben, ez pedig azt jelenti, hogy sok honfitársunknak nemcsak az albán, hanem az angol esetében is szüksége lehet ilyen eszközre. Az első és az interneten megjelenése óta folyamatosan – és ingyenesen – működő gépi fordító rendszer, a Webfordítás (www.webforditas.hu) rengeteg nyelvészeti kutatási eredmény integrálásával készült el. Mérésekből lehet tudni, hogy évente több tízmilliószor fordulnak felhasználók a gépi fordító rendszerhez (PRÓSZÉKY–TIHANYI 2009).

A hazai számítógépes nyelvészet elkövetkezendő időszakától azt lehet remélni, hogy egyre jobban fognak közelíteni egymáshoz az emberközpontú és a számítógépes elemzések. Ezt célozza egy új kutatóközösség, az MTA – PPKE Magyar Nyelvtechnológiai Kutatócsoport létrejötte 2012 elején. Kutatási céljuk – az emberi elemzéshez hasonló gépi nyelvelemző rendszer lét-

[4] http://ec.europa.eu/languages/languages-of-europe/eurobarometer-survey_hu.htm

rehozása – a számítógépes pszicholingvisztika irányába mutat. Ez nem szekvenciálisan, hanem párhuzamosan fog működni, méghozzá úgy, hogy megpróbál érzékenyen reagálni a nyelven kívüli hatásokra is. A kutatók annak a gépi modellálásával is szándékoznak foglalkozni, hogy hogyan tud ugyanúgy tévedni nyelvi elemzése közben az elemző, ahogy az ember is téved. Mert a nyelvi programok eddigi tévedései, sajnos, nem tipikusan olyanok, mint az ember tévedései: épp ezért érzik oly sokan idegennek még a gépi nyelvészet eredményeit.

12. A nyelvvel foglalkozó írások és előadások kapcsán szokás – a valójában negatív példát modelláló – Bábel képét előhúzni, ám ha a megoldás irányába szeretnénk lépni, okosabb egy másik bibliai kép ideidézése: a pünkösdé. Ott mindenki a saját nyelvén hallotta ugyanis ugyanazt az üzenetet[5]. Remélem, gondolataimmal segítettem azt a képzetet kialakítani az olvasóban, hogy a magyar nyelvtechnológia folyamatosan ebbe az irányba halad.

Kulcsszók: nyelvtechnológia, számítógépes nyelvészet, magyar nyelvtudomány.

A hivatkozott irodalom

Buvári Márta 2001. Kiejtési szótár és útmutató 15 magánhangzóval. Bárczi Géza Értékőrző Alapítvány, Bp.

Elekfi László 1994. Magyar ragozási szótár (Dictionary of Hungarian inflections). MTA Nyelvtudományi Intézet, Bp.

Grice, Paul 1997. A társalgás logikája In: Pléh Csaba – Síklaki István – Terestyéni Tamás szerk., Nyelv – Kommunikáció – Cselekvés. Osiris Kiadó, Bp., 213–28.

Halácsy, Péter et al. 2004. Hunglish: nyílt statisztikai magyar-angol gépi nyersfordító. In: Alexin Zoltán – Csendes Dóra szerk., II. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Informatikai Tanszékcsoport, Szeged, 81–4.

Halácsy, Péter et al. 2006. Using a Morphological Analyzer in High Precision POS Tagging of Hungarian. In: Calzolari, Nicoletta – Choukri, Khalid szerk., Proceedings of 5th Conference on Language Resources and Evaluation. ELRA, Párizs, 2245–8.

Kenesei István 2000. Szavak, szófajok, toldalékok. In: StrNyt. 3: 75–136.

Kilgariff, Adam – Greffenstette, Gregory 2003. Introduction to the Special Issue on the Web as Corpus. Computational Linguistics 29/3: 333–47.

Koskenniemi, Kimmo 1983. Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. Publications, No. 11. University of Helsinki, Helsinki.

Kuba, András – Hócza, András – Csirik, János 2004. POS Tagging of Hungarian with Combined Statistical and Rule-based Methods. In: Sojka, Petr et al. szerk. Proceedings of the Seventh International Conference on Text, Speech and Dialogue (LNAI 3206). Springer Verlag, Berlin, 113–21.

[5] A gondolattal Simonfai László egyik 2002-es előadásában találkoztam először, és annyira megtetszett, hogy azóta „népszerűsítem”.

MEGYESI, BEÁTA 1999. Improving Brill's PoS Tagger for an Agglutinative Language. In: FUNG, PASCALE – ZHOU, JOE szerk., Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland. ACL, New Brunswick, 275–84.

MILLER, GEORGE A. et al. 1990. WordNet: An Online Lexical Database. International Journal of Lexicography 3/4: 235–44.

NÉMETH GÉZA – OLASZY GÁBOR szerk. 2010. *A magyar beszéd.* (Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek.) Akadémiai Kiadó, Bp.

NOVÁK ATTILA – ENDRÉDY ISTVÁN 2005. Automatikus zárt ë-jelölő program. In: ALEXIN ZOLTÁN – CSENDES DÓRA szerk., A 3. Magyar Számítógépes Nyelvészeti Konferencia előadásai. SZTE, Szeged, 453–4.

ORAVECZ, CSABA – DIENES, PÉTER 2002. Efficient Stochastic Part-of-Speech tagging for Hungarian. In: RODRIGUEZ, MANUEL GONZÁLEZ – ARAUJO, CARMEN PAZ SUAREZ szerk., Proceedings of the Third International Conference on Language Resources and Evaluation. ELRA, Párizs, 710–7.

OROSZ, GYÖRGY 2011. Investigating Hungarian POS-tagging Methods. In: ROSKA, TAMÁS szerk., Proceedings of the Multidisciplinary Doctoral School 2010–2011 Academic Year. Pázmány University ePress, Bp., 77–81.

PRÓSZÉKY GÁBOR 1994. Humor: a Morphological System for Corpus Analysis. In: RETTIG, HEIKE et al. szerk., Language Resources for Language Technology: 1st TELRI European Seminar. Institut für deutsche Sprache, Mannheim,149–58.

PRÓSZÉKY GÁBOR 2000. Számítógépes morfológia. In: StrNyt. 3: 1021–64.

PRÓSZÉKY GÁBOR 2002. Comprehension Assistance Meets Machine Translation. In: TOMAŽ ERJAVEC – JERNEJA GROS szerk., Language Technologies. Institut Jožef Stefan, Ljubljana, 1–5.

PRÓSZÉKY GÁBOR 2011. A szótári világ átalakulási tendenciái az internet megjelentével. Modern Nyelvoktatás 17: 3–13.

PRÓSZÉKY GÁBOR – MIHÁLTZ MÁRTON 2008. Magyar WordNet: az első magyar lexikális szemantikai adatbázis. Magyar Terminológia 1: 43–57.

PRÓSZÉKY GÁBOR – NOVÁK ATTILA 2005. Computational Morphologies for Small Uralic Languages. In: ARPPE, ANTTI et al. szerk., Inquiries into Words, Constraints and Contexts. (Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday). CSLI Publications, Stanford, 116–25.

PRÓSZÉKY GÁBOR – TIHANYI LÁSZLÓ 2009. Webforditas.hu: egy internetes nyelvtechnológiai szolgáltatás tanulságai. In: TANÁCS ATTILA et al. szerk., VI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged, 19–23.

PRÓSZÉKY GÁBOR – TIHANYI LÁSZLÓ – UGRAY GÁBOR 2004. Moose: A Robust High-Performance Parser and Generator. In: ROSNER, MIKE szerk., Proceedings of the 9th Workshop of the European Association for Machine Translation. EAMT, La Valletta, Málta, 138–42.

RIEDL FRIGYES 1882. Simonyi kis nyelvtana. Egyetemes Philológiai Közlöny 6: 573–90.

SASS BÁLINT 2009. „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: VÁRADI TAMÁS szerk., Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból. MTA Nyelvtudományi Intézet, Bp., 117–29.

Sass Bálint et al. 2011. Magyar igei szerkezetek (A leggyakoribb vonzatok és szókapcsolatok szótára). Tinta Könyvkiadó, Bp.
Tikk, Domonkos et al. 2005. Categorizing Gigabytes: Experiments on the RCV1 Corpus. In: Tar József szerk., Proc. of the 6th Int. Symp. of Hungarian Researchers on Computational Intelligence. BMF, Bp., 267–76.
Trón Viktor et al. 2005 Hunmorph: Open Source Word Analysis. In: Martin Jansche szerk., Proceedings of the ACL 2005 Workshop on Software. ACL, Ann Arbor, Michigan, 77–85.
Varga Ágnes 2012. A gépi fordítás minősége és javítási lehetőségei. PhD-disszertáció, ELTE BTK, Bp.
Vékás Domokos 2000. Magánhangzó-rendszerek elemzése informatizált nyelvjárási korpuszon. In: Gósy Mária szerk., Beszédkutatás 2000. Beszéd és társadalom. MTA Nyelvtudományi Intézet, Bp., 75–86.

Language technology and Hungarian linguistics

Language technology and *speech technology* are cover terms for two interrelated directions of research that involve an encounter between computer technology and written, respectively spoken, language and enable computers to give responses that are similar to those of human speakers and listeners and are based on knowledge derived from the regularities of natural language. In our case, the particular natural language at hand is Hungarian; this paper gives an overview of current research on that language within the area of language technology and of the results of Hungarian linguistics in general that this emerging discipline has been able to make use of.

Keywords: language technology, computational linguistics, Hungarian linguistics.

Prószéky Gábor

A stílus szociolingvisztikai meghatározásáról*

1. B e v e z e t é s. – A stilisztika és a szociolingvisztika a huszadik század második felében rendkívül közel kerülnek egymáshoz – akadnak olyanok, akik lehetőséget látnak a két diszciplína egyesítésére is. Ez elsősorban azzal magyarázható, hogy már a nyelvészeti strukturalizmus is számos olyan nyelvi jelenséget (többek között a mondat, illetve a mondat feletti nyelvi jelenségek, a nyelvhasználat egész területe, a nyelvi változás kutatása) kizár a tudományos vizsgálat területéről, melyek ugyanakkor a nyelvi kommunikáció leírásában megkerülhetetlenek. (Ladányi – Tolcsvai Nagy 2008: 19–20.) A remélt objektív tudományos leírás érdekében történő tárgyi szűkítést a generatív nyelvészet is alkalmazza, továbbra is háttérbe szorítva a nyelv(használat) változatosságából következő empirikus sokszínűséget. Azok a megfigyelhető nyelvi megoldások, amelyek nem illeszt-

* A tanulmány a K 81315 sz., Kognitív stilisztikai kutatás című OTKA-pályázat keretében készült. Ezúton köszönöm Kiss Jenő és Tolcsvai Nagy Gábor segítő megjegyzéseit.