A magyar nyelv és a magyar nyelvtechnológia viszonyáról

Prószéky Gábor
(MorphoLogic)

Napjainkban mintegy 300 000 személyi számítógép van a mintegy 3,2 milliónyi magyar otthonban; negyedmillió egyetemista, oktató és tudományos kutató használ számítógépet, félmilliónál is több magyar internethasználót tartanak számon, 100 000-nél több magyar olvassa az interneten a híreket naponta – az esetek jelentős részében magyar nyelven. A számítógépen szövegeket létrehozó és olvasó magyar anyanyelvű beszélők koreloszlásáról viszont azt tudjuk, hogy közel a felük 24 év alatti, és csak alig több mint 20%-uk 30 év fölötti. (Angelusz–Tardos 1999) Mindez azt jelenti, hogy a számítógép hatása a jelen és különösen a közeljövő magyar írásbeliségre igen jelentős. Azt is mondhatnánk, hogy soha nem látott módon lehet a nyelvvel kapcsolatos tudást vagy éppen a nyelvi igénytelenséget "elsajátítani".

A számítógéppel létrehozott szövegek száma természetesen jóval több az internetre felkerülőkénél, hiszen manapság gyakorlatilag minden újságcikk, tudományos írás, előadás, disszertáció, könyv, törvénytervezet, hozzászólás, feljegyzés, fordítás vagy éppen levél számítógépen készül. Óriási tehát azok felelőssége, akik például a magyar felhasználót számítógépes nyelvészeti eszközökkel segítik. A nyelvi kultúra, a nyelvi kulturáltság, a nyelvi műveltség iránt elkötelezett nyelvészeti világ ezekről a jelenségekről viszont – éppen azok újdonság-mivolta miatt – csak részleges információval rendelkezik.

Ez az előadás néhány olyan találkozási pontra szeretne rámutatni, ahol a gépi nyelvkezelés problémái a nyelvészet hagyományos kutatási területein kívül más, eddig nem művelt tevékenységek meghonosítását igénylik.

1. Mi a nyelvtechnológia?

A "nyelvtechnológia" a nyelvleírásnak és a szoftvertechnológiai eszközöknek a találkozása. Jóllehet a magyar nyelvleírás hagyományai nem tették automatikusan lehetővé a számítógép számára készítendő nyelvleírást (Prószéky 1999), a számítógépes feldolgozás számára formalizált szóalaktani, sőt mára már mondattani modellek is elkészültek. Az így készült leírásokat megfelelően működtető szoftvereszközök napjaink legkorszerűbb számítógépes nyelvészeti programjai közé tartoznak – köszönhetően a rendszerváltásnak, s ezáltal a 90-es évek csúcstechnológiájára építő szoftvermegoldások lehetőségének.

Előadásunkban a magyar nyelv gép számára történő leírásának, azaz a magyar nyelvtechnológiának az eredményeit járjuk körül, remélve, hogy a magyar nyelvésztársadalom számára is mondunk újdonságokat, többek között a nyelvtechnológiai eszközöknek a nyelvre gyakorolt hatásáról. Hazánkban ugyanis a magyar nyelvi szoftvereszközöket többszázezren használják naponta, és hatásuk a magyar nyelvhasználókra – ennek következtében a magyar nyelv jövőjére – lényegesen nagyobb, mint azt elsőre gondolnánk.

2. Magyar nyelvtechnológiai kutatások és eredmények

2.1 A magyar nyelv leírása a számítógép számára

A számítógépes nyelvészet súlyát ma már hazánkban is az adja, hogy a számítógép alapvetően és elsősorban a kinyomtatandó vagy felolvasandó – és elektronikus formában egyre inkább felhasználásra kerülő – dokumentumok előállításának eszközévé vált. A nyelv (így esetünkben a magyar nyelv) grammatikájának számítógép számára készített leírása szükségszerűen különbözik a másik ember számára írttól. Annak, aki egész életében az emberek számára készített nyelvészeti munkákat, annak minden bizonnyal nehéz megérteni és – különösképpen – művelni a gépi nyelvészetet. A nyelvtechnológia tehát nem arról szól elsősorban, hogy a

nyelvészt hogyan segíti munkájában a számítógép, hanem sokkal inkább arról, hogy a nyelvészet eredményei hogyan tehetők elérhetővé a számítógép és így a számítógépet használók egyre növekvő tábora számára.

A számítógépes írástámogatás a helyes és választékos írást segítő, illetve a szöveg tördelését és elválasztását végző eszközök által végzett nyelvi tevékenység. A dokumentumok létrehozásában a szerzői eszközök, azaz az igényes szövegek létrehozását támogató nyelvhelyesség-ellenőrző, és elválasztó programok, valamint a számítógépes szinonimaszótárak a legnépszerűbb nyelvi segédeszközök. A szinte minden magyarországi szövegszerkesztő és kiadványkészítő alkalmazásba beépült *Helyesek* nyelvhelyességi rendszer különböző tagjai immár tizedik éve szolgálják azokat, akik magyar szövegeket írnak számítógéppel. A helyesírás-ellenőrzőnek keresztelt első szoftvermodulok – ahogy sokan el is nevezték őket – még csak szóellenőrzők voltak. Ám sokszor azt kellene tudni, hogy egybe- vagy különírandó-e valami, kell-e vessző stb. Ez a feladat nem oldható meg, ha fogalmunk sincs az adott szót megelőző és az azt követő szavakról. Ezzel szemben a mondatszintű helyesírás-ellenőrző több mindent lát, így össze tudja kombinálni a mondat szavainak nyelvi tulajdonságait, ezáltal bonyolultabb jelenségeket, például egybeírást–különírást vagy vesszőhibákat is képes kezelni. Gondoljuk csak el: a mondatellenőrző program működéséhez olyan grammatikát kellett írni, mely nem a tökéletes mondatokat leírni szándékozó nyelvtanokkal valahogy leírható, hanem éppen az eddig formális nyelvi módszerekkel senki által fel nem dolgozott, rosszul formált magyar mondatokat ismeri fel!

A magyar nyelvhelyesség-ellenőrző programcsomag 1993 óta megtalálható az összes magyarországi irodai rendszerben, sőt ugyanez a magyar technológia a román nyelv leírására alkalmazva 1996-tól elérhető az összes romániai irodai termékében is. Azt pedig talán nem kell magyaráznunk, milyen fontos – ha tetszik, (nyelv)politikai – eredmény, hogy teljes magyar nyelvhelyességi csomagunk 2000 óta megtalálható ugyanezen világcég irodai programrendszerének szlovák nyelvű változatában is. Ez azt jelenti, hogy minden szlovákiai számítógép-használó, akinek szlovák nyelvű irodai program van a gépén, az szlovák szövegeit a szlovák, magyar szövegeit a magyar helyesírás-ellenőrző programmal tudja ellenőrizni! Mind a szlovákok, mind a szlovákiai magyar anyanyelvűek könnyen a hivatalos politika eredményének gondolhatják ezt a nyelvtechnológiai–üzleti megállapodást, ám bárminek is gondolják, a pozitív hatást nem kell magyaráznunk. Természetesen abban a helyzetben nem voltunk, hogy az amerikai világcég pusztán a mi befolyásunkra döntött volna úgy, hogy beépíti magyar nyelvtechnológiai rendszerünket a szlovák verzióba, de hogy egy ilyen, nyelvpolitikailag nagy jelentőségű döntéshez hozzá tudtunk járulni, az bizonyos.

A globalizációnak elkerülhetetlen kísérőjelensége, hogy egyre több embernek van szüksége idegen nyelvű szövegek megértésére vagy lefordítására. Mivel sokan devalválták a "gépi fordítás" kifejezést az elmúlt időszakban, fontos elmondani, hogy az általános célú valódi gépi fordítás még mindig várat magára, de a fordítástámogató rendszerek szép sikereket érnek el. A számítógépes fordítástámogatás azt jelenti, hogy a fordítást igazából a fordító végzi, de szükség esetén hatékony segítséget kap az erre a célra kifejlesztett nyelvtechnológiai eszköztől. A legfontosabb ilyen eszköz természetesen az intelligens számítógépes szótár, amelynek alapvető funkciója, hogy a szövegekben előforduló tetszőleges toldalékolt szóalak tövének megfelelő címszót megtalálja, akár több szavas kifejezések részeként is. Az igazán hatékony gépi szótárrendszerek egyszerre korlátlan számú szótár kezelésére képesek. Számítógépes szótárak ma már ezerszám kaphatók a piacon, ám ezek közül csak kevés érdemli meg az "intelligens" jelzőt. Egy ilyen gépi szótárrendszer, a *MoBiDic*, speciális igények szerint továbbfejlesztett terminológiakezelő változatával egy sok tízezer oldalnyi Európai Uniós jogszabály-anyagot lefordító – több mint 100 fordítóból álló – konzorcium támogatása is hatékonyan megoldódott.

Ugyanakkor, ha a felhasználó a neki címzett vagy a világhálón talált szövegek egyszerű elolvasásához, megértéséhez keres segítséget (például böngészés közben), olyan "megértés-támogató" programot igényel, amelynek segítségével azonnal megtekintheti egy-egy idegen szó jelentését, anélkül hogy el kellene hagynia azt a szövegszerkesztő- vagy böngészőprogramot, amellyel éppen dolgozik. E célból készült el hazánkban az Európai Unió Információtechnológiai díját is elnyerő *MoBiMouse* program, melynek egyedülálló érdekessége, hogy három technológiát egyesít: speciális karakterfelismerő technikával – külön program elindítása nélkül – elolvassa a képernyőre írt szavakat, majd nyelvi elemzést végez (előállítja az elolvasott szó szótári alakját, elemzi a szó környezetét és kiválasztja az ott legmegfelelőbb jelentéseket), végül végrehajtja a szótári keresést, és megjeleníti az eredményt.

2.3 A magyar nyelvtechnológia eredményei a korpusznyelvészetben

Mivel a számítógépek kapacitása az elmúlt másfél évtized alatt több nagyságrendet nőtt, a mai gépi nyelvészeti lehetőségek alapvetően különböznek a korábbi idők nyelvészetének kutatási paradigmáitól. E területen is a legjelentősebb a korpusznyelvészet előretörése, mely olyan kutatásokat tesz lehetővé, melyről korábban álmodni sem mertünk. E téma hazai eredményeiről szól majd a jelen előadáshoz kapcsolódó egyik korreferátum.

2.4 A magyar nyelvtechnológia eredményei a beszéd kezelésében

Az írott nyelvvel kapcsolatos nyelvtechnológiai eredmények azért olyan fontosak, mert – az emberrel szemben – a számítógépnek az írott és nem a beszélt nyelv az "elsődleges nyelve". Ugyanakkor az egyre emberközelibb alkalmazások igénylik a beszélő és beszédértő számítógépes rendszerek létrejöttét. E területen is kiváló eredményeket mutathat fel a magyar kutatás-fejlesztés, amiről a jelen eladáshoz kapcsolódó másik korreferátum szól majd.

3. Új típusú nyelvtechnológiai alkalmazások: új nyelvészeti kutatási irányok

3.1 Információ és keresése a világhálón

A globalizáció egyik legjelentősebb mozgatója az internet, és benne az a hatalmas dokumentumháló, amelyet röviden webnek nevezünk. Az "információ robbanása" a legtisztábban a weben levő dokumentumok mennyiségének gyors növekedéséből látszik. Az internet-hozzáféréssel rendelkező ember szinte minden információt először a weben keres, és ügyeit is ott próbálja elintézni. Ezt nem tehetné, ha ott nem volna jelen valóban mindenféle – valós és valótlan, érdekes és érdektelen – információ. Lassan kijelenthetjük, hogy az emberiség eddig felhalmozott teljes tudása megtalálható a hálózatban. A web ugyanakkor strukturálatlan, mentes minden központi ellenőrzéstől, hiszen természetes fejlődés eredménye. Egyre inkább olyan, mint Stanisław Lem másodfajú démona, amely egy hordónyi dohos levegőből minden információt ki tud vonni. Az azonban véletlenszerű, hogy a kiáramló információban hol van az a darab, amelyre a felhasználó éppen kíváncsi: lehet, hogy harminc nem létező ország történelmét végig kell előbb olvasni, de az is lehet, hogy csak néhány reklámot. Ha megvizsgálnánk az internetről a számítógépek képernyőin át kiáramló információ eloszlását, azt találnánk, hogy a rendszer állapota közel jár az entrópiamaximumhoz, vagyis az információk sorrendje és lelőhelye teljesen véletlenszerű; nincsenek olyan támpontok, amelyek alapján következtetni lehetne akár a sorrendre, akár a lelőhelyre. A világhálón éppen ezért kiemelten fontos a keresés szerepe. Vannak a hálózatban olyan gépek, amelyek megpróbálják rendszerbe foglalni a rendszertelenséget. Kísérletet tesznek arra, hogy végigolvassák a hálóban levő valamennyi dokumentumot – ez egyre kevésbé sikerül –, és ezekből olyan adatbázist építeni, amelyben a dokumentumok egységesen megkereshetők. Tehát olyan kivonat – index – készül belőlük, amely csak a keresést irányító számítógépeken található meg (Prószéky–Kis 1999). Már ez is nagy segítség, hiszen maguk a dokumentumok számítógépek millióin vannak szétszórva. Az internetes információkeresés fő problémája viszont továbbra is az, hogy információkeresést ígérünk, de csak egy rövidebb betűsorozatot próbálunk megkeresni egy nagyon hosszúban. A hálózatban levő információ legnagyobb része szöveges dokumentum formájában jelenik meg. A számítógépnek valóban alapfunkciója, hogy rövid betűsorozatokat megkeres hosszú szövegekben. A gépi keresés tehát a legtöbb esetben minden intelligencia nélkül a keresett szövegnek csak azokat az előfordulásait jelzi, amelyek pontosan, betűhíven megegyeznek a keresendő betűsorozattal. A *folyóirat* szót így nem feltétlenül találjuk meg azokban a szövegekben, amelyekben csak a *folyóiratok* vagy az *újság* betűsorozat található meg. Ez pedig nagy baj, hiszen a keresést végző ember szándéka nem a betűsorozat, hanem a betűsorozat által képviselt információ megtalálása. A számítógépnek nincs fogalma arról, hogy a "folyóirat" információtartalma komoly hasonlóságokat mutat az "újság" információtartalmával; a két szó betűkódjai ugyanis nem egyeznek meg, így nem tud köztük kapcsolatot felfedezni. A számítógépnek tehát nincs eszköze a tartalmi, jelentésbeli azonosság vagy hasonlóság megállapítására – sőt, még a felszínes nyelvtani, szintaktikai összefüggések felismerései is nehézséget jelent a számára, sőt a legtöbb esetben hiányzik is. Ez pedig lassan megkérdőjelezi a hálózat mint információforrás használatát, hiszen a számítógép – a

fentiekből láthatóan igencsak korlátozott képességű számítógép – az egyetlen eszköz, amelyen keresztül hozzáférhetünk a weben tárolt dokumentumokhoz. Ekkor jelenhetnek meg a hálózat számítógépein a nyelvtechnológiai eszközök: ezek olyan programok, amelyek a keresés során megpróbálják azonosítani a különböző betűsorozatok közötti nyelvtani vagy éppen tartalmi összefüggéseket, hasonlóságokat. Ezek a keresőgépekbe rejtve, a laikus számára szinte észrevehetetlenül működnek, ám alapvetően megváltoztathatják a hálózat működését, s ezen keresztül jelentőségét és felhasználását is.

3.2 A nemzeti nyelvek szerepe és a globalizáció

A globalizáció nem eredményezte egyfajta "globális nyelv" kialakulását. Annak ellenére, hogy az angol az utóbbi évtizedekben egyeduralkodó világnyelvvé lett, a nemzeti nyelvek és kultúrák szerepe egyértelműen felértékelődött. Ez a tény a – globális és az elektronikus – kereskedelemnek köszönhető: a világszerte megnyíló, különböző helyi piacokon úgy lehet csak számottevő bevételt elérni, ha az adott piacon terjesztett termék megfelel a helyi nyelv, a helyi kultúra és a helyi szokások sajátosságainak. A szövegeket is tartalmazó termékek – számítógép-programok, könyvek és minden olyan produktum, amelyhez használati utasítás tartozik – esetén ez kézenfekvő: a terméknek a helyi nyelven kell "megszólalnia" ahhoz, hogy eladható legyen. (Az idézőjel azt jelzi, hogy a jelzett termékek többnyire írott formában tartalmazzák a szöveget.) A globalizáció ezért a legtöbb esetben lokalizációt jelent. A különböző termékek gyártói – kezdetben, a kilencvenes évek elején főleg a szoftvergyártók – jelentős összeget áldoznak arra, hogy termékeik a legtöbb országban az ottani nyelven, az ottani szokásoknak megfelelően jelenjenek meg. (Esselink 2000) A fenti folyamatot erősíti az is, hogy a weben – amely kezdetben kizárólag angol nyelvű dokumentumokat tartalmazott – a növekedést elsősorban a nem angol nyelvű weboldalak megjelenése jelenti. A jelenlegi mintegy 200 millió weblapból egy-két év múlva egymilliárd lesz, de azok közül már csak 300 millió lesz angol nyelvű – vagyis az angol nyelv még többségi pozícióját is elveszíti. A web azonban globális marad akkor is, ha dokumentumai egy helyett néhány száz nyelv valamelyikén íródnak. A magyar nyelvű dokumentumok tehát – a hálózat természetéből adódóan – elérhetők Amerikában, Kínában, Dél-Afrikában is, mint ahogy mi is el tudjuk érni az orosz, a japán vagy éppen az izlandi nyelvű webhelyeket. Ahhoz azonban, hogy a nyelvek sokfélesége ne váljon bábeli zűrzavarrá, átjárást kell biztosítani köztük. Mit tehet az, aki csak magyarul és angolul tud, ám a létfontosságú információ csak spanyolul áll rendelkezésre a hálózatban? A nyelvtechnológiának tehát nemcsak a keresésben, hanem a szövegek megértésében és megértetésében – vagyis lefordításában – is segíteniük kell.

3.3 Gépi – de főként: géppel támogatott – fordítás

Tehát a globalizáció elkerülhetetlen kísérőjelenségének kell tekintenünk azt, hogy szinte mindenkinek egyre inkább egyre több idegen nyelvű szöveg megértésére vagy lefordítására van szükség. Mérhető tény, hogy ma soha nem látott mennyiségű idegen nyelven írt szöveget kell lefordítani, de legalábbis megérteni – ehhez pedig mostanában egyre többen a számítógéptől próbálnak segítséget kérni. (Prószéky–Kis 1999) Az automatikus gépi fordítás immár fél évszázados múltra tekinthet vissza, és évtizedekkel ezelőtt készültek már működő – de jelentős korlátozásokkal működő – rendszerek. Talán meglepő, de a mai fordítórendszerek javarészt a hetvenes években készült programokra épülnek. Miért jobb képességűek mégis? Az informatika az elmúlt évtizedekben jelentős extenzív – mennyiségi – fejlődésen ment keresztül, vagyis megnőtt az egy gépen tárolható és az egységnyi idő alatt feldolgozható (nyelvi) adatok mennyisége. Egyszerűen szólva: gépeink nagyobbak és gyorsabbak, így a régi programok sokkal gyorsabban és eredményesebben működnek. Sőt, az erőforrások gyarapodása azt is lehetővé tette, hogy most megvalósítsunk korábban gazdaságtalannak tartott és elvetett eljárásokat. Mai személyi számítógépeink memóriakapacitása ugyanis három nagyságrenddel nagyobb a húsz évvel ezelőttiekénél, és a feldolgozási sebesség is körülbelül két nagyságrendet nőtt.

Ha a felhasználó az idegen szöveget gyorsan szeretné megérteni, megértés-támogató programot használ, ám ha magára a lefordított célnyelvi szövegre is szüksége van, akkor nem a gépi, de nem is a puszta emberi, hanem a gépi támogatással végzett emberi fordítás adja a legmegfelelőbb minőséget. Ilyenkor a számítógép nem bonyolult program segítségével állítja elő az egyes mondatok, szövegrészek fordítását, hanem egyszerűen megpróbálja megtalálni a lefordítandó mondatot az erre szolgáló adatbázisban, és ha megtalálja, visszaadja az ott tárolt fordítást. Ennek az adatbázisnak a neve: fordítómemória. Gépeink kapacitása ma már akkora, hogy

bizonyos feltöltési idő után egy szűkebb szakterület szinte minden mondata lefordítható az adatbázisból. A fordítómemória közeli rokona a sakkmemória: a sakkprogramokban sem a lépések szabályait ismerjük jobban, mint régen, hanem a lejátszott játszmákat tudjuk a korábbiaknál lényegesen nagyobb számban tárolni. Nem is olyan régen az IBM Deep Blue számítógépe azzal verte meg Kaszparovot, hogy nagyságrendekkel több múltbeli sakkjátszmára "emlékezett", mint a nagymester.

A gépi fordításnak természetesen csak tudományos, szakmai, esetleg köznapi szövegek (hírek, hirdetések stb.) lefordításában vagy megértésében van szerepe. A kutatások nem tudnak és nem is szándékoznak kiterjedni a szépirodalmi szövegek számítógépes vizsgálatára és a műfordításra. Viszont a mennyiség tényleg átvezet minőségbe: ha ugyanis egyetlen gép kapacitása akkora, amekkorának fentebb érzékeltettük, gondoljuk el, milyen számítási kapacitást képvisel az a hálózat, amely több százmillió (vagy lassan már több milliárd) hasonló gépből áll! Napjainkban megjelentek az olyan projektek, amelyekben az internetfelhasználók felajánlhatják számítógépük kapacitásának egy részét, és a felajánlott kapacitások konglomerátumából egyetlen soha sem látott óriási számítógép alakul ki...

3.4 Szokatlan lexikográfiai feladatok: szótárak "új ruhában"

A lexikográfiai kutatás feladatai nem válnak sem egyszerűbbé, sem bonyolultabbá a számítógépes szótárak megjelentével. A hatás más irányban jelentkezik: a gépi szótárak új problémákat hoznak, és egyben korábban még ki nem dolgozott elvek meghonosodását ígérik. Az új eszközök segítségével a mindenkori szótárhasználó munkája lesz, lehet könnyebb. Ehhez viszont a lexikográfiát művelők figyelmét rá kell irányítanunk az elektronikus szótárak kínálta rengeteg, eddig még nagyrészt kihasználatlan számítógépes nyelvészeti lehetőségre. Az "igazi" elektronikus szótárak nem pusztán a papírszótárakéhoz hasonló funkciót látnak el, hanem élnek a számítógép adta, azaz a nyomtatott szótárak által nem megvalósítható lehetőségekkel. Az első osztályba tartozó szótárprogramok – egy lexikográfiától távoli világ analógiájával – a 20. század elején megjelenő gépjárművek "ló nélküli lovas kocsi" formájára emlékeztetnek, míg a második kategóriába sorolt és általunk "igazi"-nak nevezett elektronikus szótárak az új elektronikus lexikográfia termékei – a fenti analógiával: a légellenállás és egyéb felhasználói szempontok figyelembe vételével tervezett autók világának megfelelői. A tárgyalandó kérdések egy része is szükségszerűen új, hiszen a szótárak célnyelvi lekérdezhetősége, a gépi szóalaktani elemzés aktív jelenléte, vagy az eredeti szövegkörnyezetnek a szótárhasználatra való közvetlen hatása megvalósíthatatlan és elképzelhetetlen a hagyományos szótárak esetében. Legfontosabb állításunk ebből következik: ezek a kérdések – bár jellegüknél fogva első látásra technikaiaknak látszanak – komoly hatással vannak a lexikográfiai munkára, ezért a szótártan művelőinek is mielőbb meg kell ismerkedniük az itt felmerülő problémákkal és ezek első megoldásaival.

A szótárak szerkezete tehát értelemszerűen meg kell változzék, ha az elektronikus szótárnak olyan funkciókat is el kell látnia, melyekre hagyományosan nem volt szükség. Ilyen például a több szavas szerkezetek címszavak alá sorolásának kérdése, az utaló szócikk szerepének kiváltása az internetes területről jól ismert hiperlinkek segítségével, de legfőképp ilyen a szócikk dinamikus megjelenését biztosító újfajta szócikk-strukturálás. Vegyünk csak egyetlen példát: a hagyományos szótárakban az önálló szócikk a lehetőségekhez képest a címszóra vonatkozó valamennyi tudnivalót magában foglalja, ezzel szemben az utaló szócikk nem nyújt érdemi tájékoztatást a címszóról, hanem csak azt közli, hogy a tüzetes felvilágosítást hol, melyik önálló szócikkben kell keresni. A számítógépes programok azonban a kért szó azonosításakor azonnal képesek egyetlen lépésben az utalási helyre ugrani, így a számítógépes nyelvfeldolgozó eszközök szótáraiban nincs szükség utaló szócikkre. A számítógépes szótárak különböző bemenő kérdésre is adhatnak egyféle választ, azaz mutathatják egyazon szócikk tartalmát. Például ha a tejfel szó csak a szótárban önálló szócikként szereplő tejföl-re való utalásként szerepelne a szótárban, akkor mindössze arra van szükség, hogy a tejföl címszó a tejfel bemenet estén is azonnal elérhető legyen.

A számítógépes morfológia szerepének hangsúlyozásánál rendkívül fontos, hogy felhívjuk a figyelmet arra a kevéssé ismert tényre, hogy egy komolyan megszerkesztett szótárban a szótári alapalakban álló címszórészletek száma összemérhető a nem szótári alapalakban állókéval. Ennek oka a szótári szócikkekben előforduló nagy számú kifejezés és a kifejezésekben előforduló toldalékolt formák viszonylag nagy száma. A morfológia működtetése ezért is rendkívül fontos: a kifejezések minden szavának kulcsszónak kell lennie, bármilyen alakban fordulnak is elő. Ugyanakkor, ha csak alapalakban írnánk be őket a keresőablakba, a toldalékolt alakot tartalmazó

kifejezések nem adnának találatot, pl. a *zavar* szó keresésekor a *zavarba hoz* kifejezés morfológiai komponens nélkül nem adna találatot. A többtagú kifejezések egyetlen címszó alá sorolása egyébként nem is egyértelmű, illetve a felhasználó és a lexikográfus nem mindig gondolkozik egyformán, hiszen a felhasználók egy jelentős része a szótár készítőjénél kevesebb nyelvi ismerettel rendelkezik. A papírszótárakban viszont az egyes kifejezések az érthető terjedelmi korlátok miatt csak egyetlen helyen – az ún. kulcsszó mint címszó alatt – találhatók meg. Például a *tiszta vizet önt a pohárba* kifejezés vagy a *víz*, vagy az *önt*, vagy a *pohár* alatt található, de semmiképpen sem mind a három helyen. Ezért az említett három szócikk valamelyike a papírszótárakban abban az értelemben sosem teljes, hogy a vele alkotott kifejezések, idiómák mind fel lennének sorolva. Egy számítógépes rendszer – a megfelelő strukturális szervezettség esetén – képes felsorolni mindazokat a helyeket, ahol keresett szavunk – mi több: keresett szavunk valamely toldalékolt alakja – előfordul, így a nyelvet még csak töredékesen ismerők számára is pontos segítséget képes nyújtani. Ez a probléma különösen jól illusztrálható az önállóan egyébként soha elő nem forduló címszavakkal. A *vérszem,* a *szabadláb,* vagy a *közkéz* alakok lekérdezésekor általában csak olyan címszavakat adhat vissza a szótár, melyek szócikkében szerepel a megfelelő *vérszemet kap, szabadlábra helyez, közkézen forog* kifejezés: azaz igen valószínű, hogy az adott kifejezések a papírszótárakban a *kap*, a *helyez*, illetve a *forog* címszavak alatt találhatóak meg, bár igazi különlegességük épp abban áll, hogy ezeket a különleges alakokat tartalmazzák. Nyugodtan kijelenthetjük, hogy a fenti példák nominális tagjának "közvetlen" megtalálása az idegen ajkúak számára az elektronikus szótárak használhatóságát a szokásos papírszótárak elé helyezi. Hasonlóképpen nem lehet önálló címszó például a *kedvetlenedik* ige, de ha a mondatban ott áll az *el* igekötő is, akkor helyesnek és az *elkedvetlenedik* címszó alatt megtalálhatónak kell tekinteni. Gondoljunk a nyelvünket tanuló szótárhasználóra, aki az igekötő nélküli alak megtalálást kísérli meg az igekötős formát akár tartalmazó szótárban. Ha az elektronikus szótár figyelni tud a *kedvetlenedik* alak pontos, elvált igekötős előfordulási feltételeire is, nagy – és korábban meg nem valósítható – segítséget nyújthat a gépi szótárhasználóknak.

Egy hagyományos papírszótár esetében az a tény, hogy melyik a forrásnyelv és melyik a célnyelv, az egyik legfontosabb ismérv. Ezzel szemben, a szótárt alkotó két nyelv szerepe teljességgel más az "igazi" gépi szótárak esetében. Az ok az újféle szótárszervezésben keresendő: a hagyományos címszó–szócikktest aszimmetrikus párt a címszó–jelentés, címszó–kiejtés, címszó–szófaj, azaz általánosságban a címszó–X típusú párok n-ese váltja fel. Ez által a megoldás által szimmetria vezethető be a papíron aszimmetrikusnak látszó szócikkleírásba. A gyors keresést szolgáló számítógépes indexelés tehát nem pusztán a címszavak, hanem maguknak a címszó–X pároknak a gyors megtalálását szolgálja. Az eredmény megdöbbentő: bármely szó ugyanolyan sebességgel és pontossággal található meg egy szócikkben, ha a célnyelvi oldalon szerepel, mint ha a forrásnyelvi oldalon keresnénk. Ezáltal lehetővé válik például egy magyar szó összes előfordulásának azonnali megtalálása egy angol–magyar szótár "jobb oldalán". Eredményül megkapjuk mindazokat az angol szócikkeket, melyekben valamely szócikkbeli angol szónak vagy kifejezésnek ekvivalenseként megjelenik az adott magyar szó. Például a magyar *ló* szót magyar oldali jelentésként tartalmazó angol nyelvi címszók (*horse, knight, pommel horse*) meg tudják mutatni a magyar nyelvi jelentések közötti esetleges jelentésbeli viszonyokat is. Természetesen ezek egy megszokott magyar–angol szótárból tökéletesen hiányoznak. A hagyományos szótárakban ugyanis a *ló* szomszédságában az alfabetikus környezet segítségével idekerülő *lóbál* vagy *lobbanás* szavak szerepelnek, de a *knight* címszó által tartalmazott és a *lovag* és a *huszár* szemantikus rokonságát is kimutató viszony soha. A különféle találatok tehát az eredeti magyar szó angol nyelvi megfelelőin túl az egyes angol szavak magyar megfelelői egymás szinonimáit is adják, pl. a sakkbéli *ló* a hivatalosabb *huszár* szinonimája, sőt még a *lovag* is a valódi *huszár* valamiféle jelentéstani rokona. Nyugodtan kijelenthetjük, hogy a fentiekhez hasonló, szemantikus, esetleg etimológiai vagy stilisztikai csoportokat a nyelvtechnológiai eszközöket nem használó szótárakban nem találunk.

Egy-egy szó vagy kifejezés másik nyelvi megfelelőjét a leggyakrabban szövegek olvasása közben keressük. Amennyiben ez a szöveg a számítógép képernyőjén található, a keresendő szavak teljes környezetükkel együtt vannak jelen. Az ilyen szavak esetleges egyértelműsítése, illetve egy nagyobb kifejezés részeként való előfordulás felfedezése éppen a szövegkörnyezet alapján történhet meg. Efféle lehetőség a nem-számítógépes szótáraknál nem volt lehetséges – hiszen gondoljuk csak el, honnan vehetnénk a környezetre vonatkozó információt, hacsak nem a szótárhasználó "fejéből". Ez pedig nem teszi lehetővé, hogy a hagyományosan megfogalmazott szócikkek méretét bármi módon is csökkentsük az aktuális környezet igénye szerint – éppen ellenkezőleg: az összes lehetséges környezetre fel kell készíteni az ilyen szótárat. Ugyanakkor az új típusú, dinamikus elektronikus szótáraknak mindig csak annyi információt kell adniuk, amennyi az adott

szövegkörnyezet megértéséhez szükséges. Az egyetlen követelmény, hogy a szótárazandó szó beolvasásakor rendelkezésre álljon annak eredeti környezete, akár dokumentumfájlról, akár egy internetlapról, vagy bármilyen egyéb elektronikus dokumentumról van szó. Ennek a technikája is létrejött az utóbbi időben (Clark 2000), így az utolsó olyan akadály is elgördült az új szótártípus létrehozása elől, mely értelmetlenné tehette volna a pusztán elméletileg létező szótárkonstrukció kidolgozását. A gépi szótárhoz csatlakoztatott morfológiai komponens kiegészül egy szövegkörnyezet-elemző modullal, és a kívánt szóval alkotott összes több szavas vagy igekötős szótárbeli kifejezést ezzel a szövegkörnyezettel veti össze a program. Ha valamely szótári kifejezés minden szavát (illetve ennek tövét) megtalálja a modul a kívánt szó környezetében, ezeket is a dinamikusan összeálló virtuális szócikk részévé teszi. A megjelenítendő szócikk tehát sohasem tartalmaz olyan elemeket, melyek elvileg létrejöhetnek a kérdéses szó közreműködésével, de a jelen szövegkörnyezet ilyet nem tartalmaz.

4. A nyelvtechnológia hatása a nyelvhasználatra

Sok felhasználó büszkén állapítja meg, hogy ő még mindig jobban tudja s helyesírást, mint az erre szolgáló programok. Leszögezhetjük, hogy ez így is van jól. A felhasználó a legtöbb nyelvi programtól vagy többet, vagy kevesebbet vár, mint amennyit ezek a programok teljesíteni képesek. A problémák általában ebből a jelenségből adódnak. A számítógép előtti időkben az írógép billentyűzetének használata több okból sem okozott a számítógépes gépeléshez hasonló nehézségeket: részben azért, mert az írógép nem játszhatta el az "okos gép" szerepét, s így senki nem várhatta el tőle a hibák kijavítását, részben pedig azért, mert hiába volt ugyan szabványos, minden magyar ékezetes betűt tartalmazó billentyűkiosztás, a legtöbb gépen akkor sem lehetett tökéletes helyesírással gépelni, ha valaki szeretett volna.

A nyelvi programrendszer mint minden számítógépes rendszer tartalmazhat hibákat. Sokszor viszont még az is egyfajta hibaforrás, ha a pontos, szabályos megfogalmazás következtében olyan alakokat is helyesnek tekint a program, melyeket az anyanyelvi beszélők nem. A szavak értelmezését a programok formai alapon végzik, ezért komoly hibaforrás lehet például azoknak a szavaknak a csoportja, melyek betű szerint tökéletesen helyesek, de a beszélő szándéka szerint helytelen helyesírásúak volnának. Illusztrációként vegyük a *kör–kőr* szópárt, melyek esetében gyakorisági alapon szinte kizárólag a *karika* értelmű rövid ékezetes alaknak kellene előfordulnia. A hosszú ékezetes forma megengedése a – francia kártya második legerősebb színét takaró – jelentés ismerete nélkül csak azt eredményezné, hogy sokan elhinnék a hosszú ékezetes írásmód helyességét a *karika* értelmű kör esetében. A probléma az ilyen és hasonló szavaknak a program adatbábózisából való kihagyással oldható meg, ám ennek az a következménye, hogy az esetlegesen helyesen használt *kőr* és rokonaik ismeretlenek maradnak a programrendszer számára. A nem szótári, azaz a toldalékolt alakok hasonló okok miatt még több hibaforrást jelenthetnek. A magyar főnevek esetében tökéletesen működő *-i* képzőt mindig követheti *-t* tárgyrag: *kert, kerti, kertit* vagy *fal, fali, falit* stb. Ha a főnév a viszonylag ritkán használt tan szó, a *tan, tani, tanit* sorozatnak az előzőekhez hasonlóan jónak kellene lennie. A legutolsó szó viszont sokkal valószínűbb, hogy a *tanít* ige helytelenül, rövid *i*-vel írt alakja, mintsem a tárgyragos, *i*-képzős alak. Szisztematikus változtatásra nincs mód, hiszen a *ház* szó esetében csak a rövid *i*-s *házit*, míg pl. az *alak* esetében mind a névszói *alakit* (pl. *alaki* foglalkozást), mind az igei *alakít* helyes. A számítógépes szóalaktani program számára tehát a magyar főneveknek egy olyan osztályozása szükséges, melyre korábban nem volt még szükség (Prószéky 2000b).

Az elválasztó programok jelentősége különösen nagy: a helyesírás ugyanis jó korrektorral megfelelő szinten tartható, még ha munkájuk hatékonyságát nagyban növelheti a gépi megoldás, ám elválasztani a mai elektronikus nyomdai rendszerek korában minden körülmények között a gépnek kell. Fontos működési elv, hogy az elválasztó program úgy működjön, hogy akik használják, szinte észre se vegyék a jelenlétét. Tehát az elválasztó rendszernek nem szabad interaktívnak lennie: az automatikus rendszer attól automatikus, hogy nem igényli a felhasználó közbeavatkozását. Ha a tördelő úgy látja, hogy túlzottan szét van húzva egy sor, mondjuk egy hosszú elválasztatlan idegen szó miatt, a kézi elválasztás lehetőségével bármikor élhet. Addig viszont, amíg ilyen gond nincs, elválasztási kérdésekben nem kell hozzányúlni a szöveghez. Ne feledjük a különbséget: a rossz elválasztás helyesírási hiba, az elválasztás hiánya viszont pusztán esztétikai! Így tehát, ha valaki helytelen elválasztást talál egy többre érdemes nyomdatermékben, nem kell azonnal a gépi rendszer számlájára írni a jelenséget. Sokkal inkább a nyelvművelőnek kellene elmagyaráznia, hogy a sokak által nem is ismert feltételes elválasztójel használata mi módon egészítheti ki a gép meglehetősen megbízható elválasztási modulját.

Az igazi helyesírás – szemben a korai helyesírási programok szó szintű tudásával – nem áll meg a szóhatáron. Az első ilyen szoftvermodulok – ahogy sokan el is nevezték őket – még csak szóellenőrzők voltak. Néhány kritikus nyelvi jelenség helyes kezelése arról szól, hogy egybe- vagy különírandó-e valami, kell-e vessző stb. Ezt a feladatot támogatja a mondatszintű nyelvhelyesség-ellenőrző program. A szószintű helyesírás-ellenőrzőnek csőlátása van, hiszen mindig csak azt az egy szót látja, amit odaadott neki a hívó program; fogalma sincs az előző és a következő szavakról. Ezzel szemben, ha valaki mondatszinten ellenőriz, akkor több mindent lát, össze tudja kombinálni a mondat szavainak nyelvi tulajdonságait, és ezáltal bonyolultabb jelenségeket, egybeírást–különírást, vesszőhibákat is képes kezelni. Eddig a szószintű helyesírás-ellenőrzők csak akkor adtak tanácsot egybe- és különírásról, ha helytelenül egybeírtunk valamit; a különírást ugyanis – csőlátó természetükből adódóan – mindig elfogadták, lévén a szavak legnagyobb része (a *gyógy-*, *al-* és a hasonló előtagokat leszámítva) önállóan helyes. Ezért mindig érdemes kipróbálni a kritikus szavak egybeírását, mert az egybeírási hibát lehet szószinten kezelni. Viszont ha külön írták, akkor már csak a nyelvhelyesség-ellenőrző segíthet. Ha valaki tehát nem ismeri a mondatellenőrzőt, vagy olyan alkalmazást használ, melyben nem érhető el ez a szolgáltatás, résen kell lennie: miután a helyesírás-ellenőrzőnek nevezett szóellenőrző nem alkot véleményt, a döntés a felhasználóra marad. Ha azonban ezzel valaki nincs tisztában, elfogadja, hogy ez esetekben a gép nem jelez hibát, azaz esetleges igénytelenségből hibás helyesírási szokások alakulhatnak ki.

A számítógépes nyelvek fordítóprogramjaiból kölcsönzött szakkifejezésekkel úgy jellemezhető a kétféle helyesírás-ellenőrző viszonya, hogy a helyesírás-ellenőrző hibaüzenetet (error message) küld, a nyelvhelyesség-ellenőrző pedig csak figyelmeztetést (warning). A figyelmeztetés egy jelzés arra nézve, hogy itt és itt probléma lehet, de nem biztos, hogy van. További kérésre a mondatellenőrző idézi a *Helyesírási szabályzat* ide vonatkozó passzusát. A döntés talán még a szóellenőrzők javaslatainak elfogadásánál is jobban a felhasználó kezében van. Például, ha a szövegben egymás után szerepel az a két szó, hogy *vendég* és *fogadókat*, még nem biztos, csak igen valószínű, hogy egybe kell őket írni. Lehetséges ugyanis, hogy egy másik szövegkörnyezetben helyes a különírt változat is, pl. az *Ez a vendég fogadókat foszt ki* mondatban. A helyes döntéshez egyrészt egyfajta nyelvi igényesség, másrészt a lehetséges döntéstámogató eszközök, szabályzatok elérésének hatékony biztosítása szükséges. A modern nyelvművelés feladata itt tehát az, hogy az első elengedhetetlen voltára és a második elérhetőségére felhívja a felhasználók figyelmét.

5. A magyar nyelvi technológia felelőssége

Kész a szöveg, s mielőtt még elégedetten hátradőlnénk, végigfuttatjuk rajta a helyesírás-ellenőrző programot. Így tesz ma mindenki, íróembertől diákig, tanártól titkárnőig. A képzettségünkön, szellemi önállóságunkon múlik, vitába merünk-e szállni a gép ítéleteivel, vagy feltétel nélkül megbízunk bennük. Honfitársaink egy része van annyira bizonytalan a saját helyesírásában, hogy örömmel veszi, ha eligazítják az írott szövegben végzett tévelygései közben. A magyar helyesírás-ellenőrző program nem csupán számítógépes program, és nem is csak piaci termék, hanem a magyar helyesírás jövőjének meghatározó szereplője. Ugyanakkor a mai napig nincs tisztázva, hogy egy magánvállalkozáson kívül ki vállalná az ezzel járó felelősséget. A nyelvművelőknek kellene valójában tudatosítaniuk az emberekben, hogy mit kell, mit lehet és mit nem szabad "ráhagyni" a számítógépes nyelvhelyességi rendszerre, a nyelvművelők nagy része viszont nem is ismeri a gépi eszközök logikáját.

Sajnos sokszor még ezen az ismereten túl a nyelvi szoftvereszközök (külföldi) készítőinek logikáját is ismerni kellene. A szövegszerkesztő programokban létezik például egy olyan modul, amelyik automatikusan nagybetűre javítja a pont utáni első szó kezdetét. Ezt az automatizmust – mely nem része a magyar nyelvhelyességi rendszernek, pusztán egy technikai programmodul – ki lehet kapcsolni, ha nincs rá szükségünk. Aki ezt nem tudja, az eleinte nem is veszi észre, aztán ha észreveszi, nem érti, hogy az évszám után miért nagy kezdőbetűvel jelennek meg a hónapok nevei. Mivel a felhasználók többsége nemcsak hogy nem tudja, hogy ezt az automatizmust ki lehet kapcsolni, de arra sem emlékszik, hogy ő maga írta-e nagybetűvel a hónap nevét vagy sem, így ki sem javítja a végleges szövegben, sőt, egy idő után talán még szokatlannak sem találja. A magyar helyesírási gyakorlatba tehát így látszanak bekerülni a nagybetűs hónapnevek, amiről többen azt gondolják, hogy pusztán az angolt követjük ebben is, ám – mint láttuk – a magyarázat egészen más.

A magyar nyelvvel kapcsolatos nyelvtechnológiai kutatások, főként az alapkutatások finanszírozását az erre szakosodott egyetlen cég éppen az e kutatások segítségével létrehozott nyelvtechnológiai fejlesztéseiből fedezi, bár a magyar nyelv számítógépes védelmét talán nem egyetlen kicsiny magáncégnek kellene vállalnia. Ahogy

Glatz (1999) fogalmaz: “a kis nyelvek korszerűsítési programja sohasem történhet üzleti alapon: nem kifizetődő befektetés” Ha tehát a jövőben koordinálni lehetne a fő kutatási irányok nyelvtechnológiára vonatkozó elképzeléseit a már megvalósított, illetve a most is megvalósításra váró kutatásokkal, vagy intézményi formát lehetne találni esetleges szorosabb együttműködésnek létező és megvalósítandó intézmények között, annak valószínűleg a magyar nyelvhasználók látnák legnagyobb hasznát. Úgy gondoljuk, hogy egy nyelvtechnológiai magánvállalkozás több mint tíz év nyelvtechnológiai tapasztalatával és eredményeivel komolyan tudna segíteni annak az állításnak a minél teljesebbé tételében, hogy “a nyelvi technológiák kifejlesztése a magyar nyelv modernizációjának legalapvetőbb tényezője és feltétele” (Kiefer 1999)

Irodalom

Angelusz–Tardos (1999)
: Angelusz Róbert – Tardos Róbert. A számítógépes és az internetkultúra magyarországi eltterjedésének adatai. In: Glatz Ferenc (szerk.) *A magyar nyelv az informatika korában,* 177–186. MTA, Budapest, 1999.

Clark (2000)
: Clark, Bob: MoBiMouse, the World's First "No-Click" Dictionary Program. *International Journal of Language and Documentation*, Issue 3, January 2000, 26–27

Esselink (2000)
: Esselink, Bert: *A Practical Guide to Localization*. John Benjamins Publishing Company, Amsterdam, 2000.

Glatz (1999)
: Glatz Ferenc: Tézisek a magyar nyelvről. In: Glatz Ferenc (szerk.) *A magyar nyelv az informatika korában,* 13–15. MTA, Budapest, 1999.

Hanthy (2002)
: Hanthy Kinga: Gépnyelvművelők. *Magyar Nemzet* (2002. március 30.)

Kiefer (1999)
: Kiefer Ferenc: Néhány gondolat a nyelvi technológiákról. In: Glatz Ferenc (szerk.) *A magyar nyelv az informatika korában*, 128-132. MTA, Budapest, 1999.

Papp (1989)
: Papp Ferenc: *Alkalmazott nyelvtudomány*. (Akadémiai székfoglaló, 1986. május 19.) Budapest: Akadémiai Kiadó, Budapest, 1989.

Prószéky (1993)
: Prószéky Gábor: Nyelvművelés számítógéppel? (A helyesírás-ellenőrzés új útjai). *Magyar Nyelvőr* 117: 509–511, 1993.

Prószéky (1999)
: Prószéky Gábor: Természetes nyelvek. In: Futó Iván (szerk.) *Mesterséges intelligencia*, 756–814, Aula, Budapest, 1999.

Prószéky (2000a)
: Prószéky Gábor: A magyar morfológia számítógépes kezelése. In: Kiefer Ferenc (szerk.) *Alaktan (Strukturális magyar nyelvtan 3),* 1024–1065. Budapest: Akadémiai Kiadó, Budapest, 2000.

Prószéky (2000b)
: Prószéky Gábor: A nyelvtechnológiai alapkutatások hiányáról és szükségességéről. In:T. Molnár István – Klaudy Kinga (szerk.) *Papp Ferenc akadémikus 70. születésnapjára,* 157–165. Kossuth Egyetemi Kiadó, Debrecen, 2000.

Prószéky–Kis (1999)
: Prószéky Gábor – Kis Balázs: *Számítógéppel emberi nyelven. Természetes nyelvi feladatok megoldása számítógéppel.* SZAK Kiadó, Bicske, 1999.