PRÓSZÉKY GÁBOR

Nyelvtechnológia 2000

Kezdetben, úgy az ötvenes években volt a *gépi fordítás* (machine translation, MT). Ez a 60-as évek kutatásai (elsősorban a mesterséges intelligencia-kutatás) következtében átalakult egy általánosabb diszciplínává, a *számítógépes nyelvészetté* (computational linguistics, CL). Amikor az elméleti nyelvészet saját módszereitől egyre inkább eltérő, már a gyakorlatban is használható módszerek megjelentek (részletesebben ld. Prószéky 1989), a terület nevét egyre többször olvasni *természetesnyelv-feldolgozásként* (natural language processing, NLP). Ez már a 70-es évek szóhasználata, amit a 80-as években a gyakorlati alkalmazások előtérbe kerülése miatt a *nyelvtechnológia* (language technology, LT), illetve a magyarban kissé szokatlan *nyelvmérnökség* (language engineering, LE) kifejezéssel illettek. A 90-es évekre egyre többen – az Európai Unióban meghonosodott szóhasználattal – *nyelvi iparról* (language industry, LI) beszélnek.

Ez a kis bevezetés nem csupán a szavak szintjén jelentkező változásokat szándékozott bemutatni, sokkal inkább azt, hogy ezek egyfajta beállítottságbeli változást is sugallnak. Éppen ezért tartjuk fontosnak szélesebb közönség előtt is a témával való foglalkozást, hiszen az ipari méretekben megjelenő gépi nyelvészeti termékek gondolata még sokaknak új lehet. Az Európai Unió eddig még nem látott energiával támogatja az ezirányú fejlesztéseket, nem feledkezve meg a Közösségen kívüli országok nyelveiről sem. Ennek természetesen itt nem tárgyalandó egyéb okai is vannak, de a tény az tény: a 90-es évekre a magyar nyelv számítógépes szempontok alapján történő leírása is fontos lett az egységes európai nyelvi infrastruktúra szempontjából (Prószéky 1988, 1994a, 1994b). Ennek oka egyfelől az, hogy Magyarország térben – és remélhetőleg időben is – meglehetősen közel van az EU-hoz, továbbá megvannak a szükséges fejlesztés hardver és szoftver feltételei, másfelől van Magyarországon olyan számítógépes nyelvészetre szakosodott intézmény (a MorphoLogic), amely aktívan közreműködik az Európai Unió nyelvfeldolgozási kutatásaiban. Az érdeklődés tehát részben magára a magyar nyelvre, részben pedig az annak gépi leírásához használt módszerre irányul. Ez utóbbihoz hasonló elképzelések ugyanis ritkán születnek olyan országokban, ahol a beszélt nyelv morfo-szintaktikai szempontból kevésbé bonyolult.

A nyelvfeldolgozási politika tehát a nyelvek sokfélesége ellenére *egységes formális nyelvleírást*, továbbá közösen kidolgozott *szabványos nyelvi alkalmazásokat* támogat.

Language and Technology 2000: az Európai Unió számítógépes nyelvfeldolgozási politikája

„Language and Technology 2000" a neve annak a kisszerűnek aligha nevezhető projektumegyüttesnek, melyet az Európai Unió arra szán, hogy az ezredfordulóra várható irdatlan információtömeget a számítógép a nyelv bizonyos

szintű megértésén keresztül a jelenleginél lényegesen intelligensebben tudja kezelni. A cél az, hogy azt a hátrányt, amelyet a számítógépes nyelvészeknek Európa soknyelvűségéből következően kell leküzdeni, előnnyé változtassuk. Az egynyelvű Egyesült Államok vagy a szintén egynyelvű Japán nyelvtechnológiája épp a többnyelvűséggel kapcsolatos problémákkal nem kíván megküzdeni, így feldolgozási stratégiáik sem tükrözik a nyelvfüggetlen megoldásokra való törekvést. Az okok különbözőek: Japánban még az írásrendszer is egyedi és bonyolult, ami még jobban elszigeteli a nyelvi szempontból is szigetként jellemezhető japán nyelvet beszélőket a világtól. Ugyanakkor az Egyesült Államokban az egyetlen hivatalos nyelv a világon mindenfelé beszélt angol, így mind a piac, mind a fejlesztők fejében élő világkép egyaránt azt sugallja, hogy a problémák nyelvfüggő részét úgy kell megoldani, hogy elsősorban az angol nyelv esetében adjon használható megoldást. Ha más nyelvek is megjelennek az adott rendszerben, akkor majd elkészítik a megfelelő modult azok, akik erre rá lesznek kényszerítve. Mik is ezek a problémák? Például az *n* darab nyelv közötti *n-szer n-1* fordítás; az egyetlen dokumentum egyszerre több helyre több nyelven való elküldése; több különböző nyelven írt dokumentum automatikus összehangolása; idegen nyelven írt dokumentumok anyanyelvünkön való kivonatolása, és még sorolhatnánk.

Ahhoz, hogy a várt nyelvi modellek elkészüljenek, némi időre van szükség. Azonban addig is előrevetíti árnyékát egy olyan jelenség, amelyet a gépi nyelvészet kritizálói minden alkalommal felemlítenek, nevezetesen: a nyelvi szegényedés. A számítógépes rendszerektől persze nem lehet elvárni azt, hogy Arany János-i szinten használják a magyar nyelvet, de az is biztos, hogy az alkalmazási területek szövegei legkevésbé a szépirodalom területéről valók. Megjegyezhető, hogy a hivatalos nyelvben ma is sok a szegényes, silány fordulat, pedig a számítógépes nyelvfeldolgozó rendszerek még nemigen voltak rá hatással. Tehát nyilván nem a gép okozza a nyelvi szegényedést, de az is biztos, hogy avatatlan kezekben könnyebben fogja támogatni ezt a negatív folyamatot. Ugyanakkor viszont az egyre jobban használható nyelvi szoftvereszközök megjelenése gyorsíthatja is az igényesebb írást, pontosabb helyesírást, választékosabb fogalmazást (Prószéky 1994b).

A „Language and Technology 2000" aktualitását több dolog együttes megjelenése adja. Mindenütt elérhetők már a gépi szövegelőállítás eszközei, minden valamirevaló dokumentum az egyesült Európában már gépen készül – gondoljunk csak az újságokra, könyvekre, jogszabályokra, rendeletekre, üzleti és magánlevelekre, vagy bármilyen egyéb kiadványra –, és akkor még az automatikus dokumentumgenerálásról nem is beszéltünk. Ez utóbbi témakör egyébként az unalmas szerződésséмák kitöltésétől az időjárás-jelentések speciális nyelvi fordulatainak automatikus előállításáig rengeteg mindent felölel. Tény, hogy egyre több dokumentum készül gépen, sőt, nemcsak hogy készül, de óriási tömegek számára azonnal elérhetővé is válik a hálózati rendszerek jóvoltából. Az Európai Unió egyik tanulmányának becslései szerint az ezredfordulóra több lesz a géppel előállított olyan dokumentumok száma, amelyeket nekünk címeznek, mint amit – erre fordítható idő híján – egyáltalán el tudunk olvasni (Danzin 1992). Marad tehát a személyi titkárnő, aki megszűri leve-

leinket, esetleg faxainkat, és csak azt teszi az asztalunkra, amelyik valóban a mi személyes közreműködésünket igényli, amelyik valóban nekünk szól. Ma viszont, amikor az elektronikus levelezés sokkal személyesebb, mint bármilyen eddigi levelezés, titkárnőnknek a számítógépben kellene helyet foglalnia, hogy elektronikus postánkat az előbb felsoroltak alapján rendezze-szervezze. Ez a titkárnő tehát nem lehet más, mint az a nyelvi szoftvercsomag, amely átfutva e-postánkat, osztályozza, felénk vagy a hulladékkosár felé továbbítja, ne adj isten, le is fordítja, sőt, egyszerűbb esetben meg is válaszolja leveleinket...

Az efféle funkciókat ellátó szoftverekre korábban nem is volt ekkora igény, mint manapság, de nem is igen lehetett volna megvalósítani őket, elsősorban a korábbi számítástechnikai eszközök hely- és sebességproblémái miatt. Egyébként ma sem azért lehet a legtöbb nyelvészeti problémára megoldást találni, mert mára sokkal okosabbak lettünk, hanem mert korunk számítógépe elég nagy és elég gyors a korábban megfogalmazott, sokszor meglehetősen egyszerű – vagy mondjuk ki: buta –, de ma már gyorsan végrehajtható megoldások megvalósításához.

Vegyük például a gépi fordítást! Ma már sokszor nem is fordításról, hanem fordítástámogatásról beszélünk. Ez utóbbi rendszerek nyelvi készségei gyakran minimálisak, de amit tudnak, azt nagyon gyorsan tudják. Könnyen lehet, hogy a fordítónak valójában nincs is másra szüksége, csak erre a sebességre. Gondoljunk például egy sakkozóra, aki rengeteg játszmát elemzett végig életében, és ha ezek mindegyikére pontosan emlékszik, általában többet tud, mint az, akinek nagyszerű saját elgondolásai vannak, de kevés játszmaismerete. Ennek az az oka, hogy – a legzseniálisabbakat leszámítva – a saját gondolat korábban már megfogalmazódott másokban is, azaz a spontán elképzelés a „betanult" partik valamelyikében nemcsak ötletként, gondolati csíraként, hanem teljes kifejletében megtalálható. A jó sakkprogram természetesen képes szabály alapú kombinációkra is, ami persze nem lebecsülendő, de azt a gyakorlat igazolja, hogy mások játszmáinak ismerete nélkül a sakkprogram biztos vereségre van ítélve. A fordítóprogramok most hasonló fejlődési irányt mutatnak: a számítógép nem elsősorban fordít, hanem inkább hatékonyan keres a korábbi fordítások között egy-egy hasonló szerkezetet, ugyanis keresni nagyon jól tud, és még az így kapott eredmény is gyakran pontosabb, hiszen profi fordító profi fordítását találja meg (feltéve persze, hogy a meglévő mintafordítások valóban jók).

Korábban említettük, hogy a gép hamarosan maga írja a szöveget. Napjaink technológiája már most is lehetővé teszi, hogy egy erre szolgáló program vezérelje fogalmazásunkat. Persze nem elsősorban stilisztikai szempontok alapján, hanem az általa ismert grammatikai szabályok figyelembevételével, de tény, hogy képes ellenőrizni, hogy hogyan fogalmazok. Amint az általa ismert – nyilván a gépelőénél szegényesebb – nyelvtani szerkezetektől eltérőt tapasztal, jelzést küld. Senki ne gondolja, hogy ilyenkor valamilyen „igazi" hibát észlel, mindössze azt jelzi, hogy ha nem tértem volna el az őáltala ismert nyelvi fordulatoktól, ő például garantáltan le tudná fordítani szövegemet az adott környezetben beállított nyelv(ek)re. A döntés az enyém: ha nem változtatok, akkor a gép fordításaiba valószínűleg bele kell majd nyúlnom. Ha betartom a rendszer tanácsait, akkor szövegem az én további interakcióm nélkül lesz le-

fordítva, természetesen a géptől elvárható, meglehetősen „stílusmentes” formában.

Kulcsszavak: újrafelhasználhatóság és korpusz-nyelvészet

Ahhoz, hogy az eddig ismertetett feladatokat meg lehessen oldani, hihetetlen mennyiségű írott szöveg átvizsgálásából, elemzéséből, statisztikai tulajdonságainak a felderítéséből szerzett tapasztalatokra kell támaszkodni. Ebben a munkában a számítógéppel – de nem nyelvészeti céllal – készített szövegek nagy segítségünkre lehetnek. Az újrafelhasználhatóság fogalma tehát, az élet más területeihez hasonlóan, a nyelvészeti alkalmazások területét sem hagyta érintetlenül. Napról napra világosabb, hogy a nyelvi tudás nagy része nem pusztán a szótárak, lexikonok, enciklopédiák, nyelvtanok formájában, hanem magukban a leírt szövegekben van elrejtve. Ilyenkor elég csak a hagyományos szótárak vagy nyelvtankönyvek példamondataira gondolni. Az utóbbi években a számítógéppel történő szövegszedés annyira eluralkodott, hogy az EU-ban már szinte minden, ami újság, könyv vagy egyéb kiadvány formájában jelenik meg, géppel olvasható formában van. Ezeknek a hatalmas szöveganyagoknak, nyelvészeti szakszóval korpuszoknak az elemzésére, feldolgozására egyre több szoftver készül. Ha egy szövegnek megvan valamilyen fordítása is, probléma lehet az eredeti és a fordítás-szöveg szavainak, mondatainak, bekezdéseinek a lehetőségekhez képest legjobb szinkronizálása. A gépi nyelvészeti kutatás komoly statisztikai módszerekkel ötvözve ma már igen rafinált eszközöket készít az egy- és többnyelvű korpuszok nyelvészeti kutatás céljára való feldolgozásához.

Hazánkban, az MTA Nyelvtudományi Intézetében elsősorban csak irodalmi írott szöveget gyűjtenek. Ez a mai köznyelv részletes leírásához nem elegendő. A napilapokat bár géppel szedik, az előállt hatalmas szövegkorpuszt nem tudván hol tárolni, néhány nap után megsemmisítik. A mai magyar nyelvnek igazi szövegkorpusza tehát jelenleg nincs, így nem csoda, hogy az 1985-ben a European Corpus Initiative által kutatási célokra megjelentetett és 27, javarészt európai nyelven mintegy 100 millió szónyi anyagot tartalmazó CD-n semmiféle magyar anyag nem található. Ebben a helyzetben különösképpen üdvözlendő az a tervezet, mely egyik legjelentősebb napilapunk félévenkénti teljes anyagának CD-n való kiadását célozza.

Szerencsére a magyarországi számítógépes nyelvészeti kutatás jelentős korpuszok nélkül is meg tudott újra indulni.

Magyar számítógépes nyelvészeti fejlesztések: itthon és az EU-ban

A 90-es évekre hazánkban is megjelentek az első számítógépes nyelvészeti alkalmazások. A széles nagyközönség ekkor találkozhatott többek közt a MorphoLogic nyelvi programrendszerének a napi gyakorlatban is használható első darabjaival (Prószéky 1993). Először a *Humor* morfológiai programrendszer leszármazottai jelentek meg: a *Helyes-e?* helyesírás-ellenőrző, a *Helyesel* automatikus elválasztó, a *Helyette* toldalékoló szinonimaszótár (Prószéky & Tihanyi 1993) és a pontos szöveges keresésnél nélkülözhetetlen *HelyesLem*

szótő-visszaállító program (Prószéky, Pál & Tihanyi 1994). A közelmúltban bemutatkozott egy újabb modul, mely már elér a mondatelemzés mélységeibe. Ez a *HumorESK* mondatelemző rendszer. A napi gyakorlatban is megjelentek a magasabb szintű nyelvi tudással rendelkező modulok, melyek az intelligens fordítástámogatást (*MoBiDic* toldalékoló kétnyelvű szótárcsalád) és a fogalmazás-támogatást (*Helyesebb* mondatszintű helyesírás-ellenőrző rendszer) tartják szem előtt. A nagy nemzetközi partnerekkel való együttműködésből következett, hogy az említett modulok a magyar mellett több más, elsősorban kelet-európai nyelvre is létrejöttek.

A magyar számítógépes nyelvészeti kutatás egyfajta nemzetközi elismerése, hogy 1995-től az Európai Unió és a közép-kelet-európai államok közös fejlesztéseire létrehozott Copernicus-együttműködés keretében részt vehetünk a *GLOSSER*, a *GRAMLEX*, a *MULTEXT-EAST*, a *TELRI* és az *ELSnet Goes East* projektumokban. Szerepünk elsődlegesen az, hogy javaslatokat tegyünk a kialakulóban levő gépi nyelvfeldolgozási szabványok olyan irányban történő megváltoztatására, melyek segítségével a magyar és más kelet-európai nyelvek számítógépes használatra alkalmas leírása a nyugat-európaiakéival egységes módon történhet. Ez lehetővé fogja tenni az egységes nyelvi szoftvereszközök használatát is, melyek kifejlesztésén a fent említett pályázatokban részt vevő két intézmény, az MTA Nyelvtudományi Intézet és a MorphoLogic dolgozik.

IRODALOM

Danzin, A (1992). *Towards a European Language Infrastructure.* CEC Doc.

Prószéky, G. (1989). *Számítógépes nyelvészet (Természetes nyelvek használata számítógépes rendszerekben).* SZÁMALK, Budapest

Prószéky, G. (1988). Hungarian – A Special Challenge to Machine Translation? In: Maxwell, Schubert & Witkam (ed.) *New Directions in Machine Translation.* Dordrecht, Foris, 219–231.

Prószéky, G.(1993). Nem pusztán az a fontos, hogy helyes-e, hanem hogy mennyire intelligens... In: T*udományos és Műszaki Tájékoztatás,* 27–41.

Prószéky, G. (1994a). Language and Technology in Hungary. In: *Proceedings of the Awareness Days for Central and Eastern Europe,* Luxemburg, 36–42.

Prószéky, G. (1994b). Industrial Applications of Unification Morphology. In: *Proceedings of the 4th Conference on Applied Natural Language Processing.* Stuttgart, 157–159.

Prószéky, G. & Tihanyi, L. (1993).: Helyette: Inflectional Thesaurus for Agglutinative Languages. In: *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics.* Utrecht, 473.

Prószéky, G. – Pál, M. & Tihanyi, L. (1994). Humor-based Applications. In: *Proceedings of the COLING-94.* Kyoto, 1241–1244.