

12. A NYELV ÉS A SZÁMÍTÓGÉP

1. A számítógépes nyelvészet vagy újabb nevén: a nyelvtechnológia olyan műszaki tudomány, amely a természetes nyelvű szövegek számítógépes feldolgozásával foglalkozik. Nehéz ennél szabatosabb meghatározást adni, mert valóban minden olyan elméleti és gyakorlati (leginkább programfejlesztési) tevékenység ide tartozik, amely kapcsolatban van a természetes nyelvekkel. Ez pedig rendkívül szerteágazó szolgáltatásokat takar.

A fenti meghatározást szűkíteni kell kissé, hiszen a természetes nyelven írott szövegek már vagy harminc éve jelen vannak a számítógépen, teljesen magától értetődő módon. Elviekben tehát a nyelvtechnológia világába illenének azok a kódolási, megjelenítési eljárások is, amelyek lehetővé teszik, hogy a számítógépen egyáltalán szöveget ábrázoljunk. Megállapodhatunk azonban, hogy ezeket az alacsony szintű eljárásokat nem tekintjük a nyelvtechnológia körébe tartozónak.

Azért ilyen nehézkes a nyelvtechnológia mint tudomány vagy mesterség meghatározása, mert *interdiszciplína*, vagyis olyan szakma, amely több terület eredményeire és tudására épül. A nyelvtechnológia az informatikát, a nyelvészetet és a matematika egyes ágait (formális nyelvek, automataelmélet, statisztika, halmazelmélet) köti össze, s néha nem tudni, egy adott probléma tisztán informatikai vagy nyelvtechnológiai-e. Tovább bonyolítja a helyzetet, hogy a formális nyelvek és az automaták elmélete (a matematika e két ága áll a legközelebb a nyelvészethez) része az informatika legfontosabb elméleti alapjainak is.

Tiszta helyzetet tehát úgy teremthetünk, hogy nem tudományelméleti, hanem kizárólag gyakorlati céllal alkotunk definíciót: eszerint pedig a nyelvtechnológiát azok az elméleti és műszaki tevékenységek alkotják, amelyek révén a számítógép képes természetes nyelvű szövegek ilyen vagy olyan feldolgozására. Ha a mindenki számára látható szolgáltatásokat nézzük, ide tartozik természetesen a gépi fordítás, a gép és az ember közötti természetes nyelvű kommunikáció, de ennél egyszerűbb dolgok is, például a helyesírás-ellenőrző programok és rokonaik, illetve a számítógépes szótárak és rokonaik – és az internetes keresőrendszerek egyes szolgáltatásai is. A nyelvtechnológia persze mindemellett tudomány is, több könyvtárnyi szakirodalommal, se-regnyi elmélettel és kutatási módszerrel, amelyek mind a nyelv szerkezetének gépi modellezésére irányulnak.

A következőkben alapvetően két kérdésre próbálunk válaszolni: először arra, miért olyan nagy probléma a számítógép számára az emberi nyelvek megtanulása; másodsor arra, hogy mindennek ellenére milyen nyelvi szolgáltatásokat várhatunk már most, illetve a közeli és a távoli jövőben a számítógéptől, illetve azoktól a berendezésektől, amelyek számítógépi feladatokat képesek végezni, vagyis van bennük számítógép.

2. Említettük, hogy a nyelv területén a számítógép és az ember között hatalmas a szakadék: az embernek veleszületett képessége a nyelvvel való bányi tudás, vagyis kezdettől rendelkezik azzal az apparátussal, amely lehetővé teszi például, hogy megtanuljon beszélni. A számítógépnek ezzel szemben semmiféle „veleszületett” képessége nincs; ha tehát meg akarjuk tanítani beszélni, létre kell hoznunk rajta az ehhez szükséges hátteret. Ennek számos akadályja van: arról akár ne is szóljunk, hogy a pszicholingvisztika és a neurolingvisztika ma még nem rendelkezik olyan részletes ismeretekkel az ember nyelvi műveleteiről, amelyek átfordíthatók lennének a számítógép nyelvére. Legyen elég arról beszélni, hogy a nyelv szoros kapcsolatban áll az ember világismeretével; arról, hogy a szavak és a belőlük felépülő kifejezések, mondatok, szövegek semmit sem érnek azon kapcsolat nélkül, amely a nyelv elemeit összeköti a világ dolgairól alkotott képünkkel. A számítógép ettől kezdve duplán hátrányos helyzetben van: sem veleszületett nyelvi képessége, sem világismerete nincs. A számítógép – s bárhogy árnyaljuk is a képet később, a lényeg ugyanaz marad – a benne tárolt szöveget számkódok sorozatának látja, semmi egyébnek. Ha tehát el akarjuk érni, hogy a mienkhez hasonló nyelvi képességgel rendelkezzen, egyszerre kell nyelvi apparátust és világismeretet adnunk neki, már amennyiben ragaszkodunk ahhoz, hogy ez a kettő szétválasztható s egymástól függetlenül kezelhető. A nyelvi apparátus dolga még akár rendben is lenne: a következőkben azt látjuk majd, hogy a legtöbb szolgáltatás a nyelvi apparátus felszínes modellezésére épül, és nemigen használ fel semmilyen világismereti elemet. A világismeret gépi ábrázolásának kutatása s ennek a gépi kommunikációban való alkalmazása nem nyelvészeti kérdés, de a kutatók tudatában vannak ennek a területnek a fontosságával, és jelentős haladást értek el az utóbbi években.

Ha ilyen nehéz a gépet megtanítani a nyelvünkre, miért követeljük tőle kezdettől fogva, hogy segítsen a nyelvvel kapcsolatos feladatainkban? Kévsz az a válasz, hogy az ember ambiciózus lény, és szereti a nehéz problémákat. Inkább az a helyzet, hogy a mi nyelvvel kapcsolatos munkánkban akadnak idegölő, unalmas és munkaigényes feladatok is – ilyen sokszor a fordítás, a szövegkeresés vagy a szövegek javítása –, és a számítógéptől megszoktuk, hogy a nehéz, de egysíkú szellemi munkában nagyon sokat tud segíteni. Most mondtuk ki a kulcsszót: *segíteni* – és nem helyettünk elvégezni – és ami ezt a segítséget illeti, már ma is nagyon sokat kapunk a géptől a nyelv területén is.

3. Az informatika kezdetben a kiváltságosok szakmája volt: a legegyszerűbb számítógépes művelet elvégzéséhez is jelentős szaktudásra volt szükség. Amikor a nyolcvanas években tömegesen elterjedtek a számítógépek,

égető problémává vált, hogy a szaktudással nem bíró felhasználónak gyakran aránytalanul sokat kell tanulnia, amíg igazából hasznát tudja venni a gépnek.

Ennek oka egyértelműen az ember és a gép közötti kommunikáció módja: napjaink számítógépének még mindig rengeteg – több száz – kezelőszerve van, s ezek rendszere csak részben hasonlít az írógép kezelőszerveihez (s azokról a felhasználókról sem szabad megfeledkezni, akik írógépet sem használtak azelőtt). Bár a programok az utóbbi időben egyre intuitívabb kezelőfelülettel jelennek meg, amelynek viselkedését háttértudás, azaz műszaki ismeretek vagy speciális nyelvtudás nélkül is meg lehet érteni, ezek még mindig jelentősen különböznek az ember „természetes” kommunikációs eszközeitől. Egyfelől tehát alapvető célnak kellene lennie, hogy a számítógépek kommunikációját igazítsuk az emberéhez, és ne fordítva. Másfelől viszont az is igaz, hogy az emberiség, intellektuális fejlődése során, képes volt új kommunikációs formák megtanulására: erre nyilvánvaló példa maga az *írás*.

Egy szó mint száz, a nyelvtechnológia szempontjából különösen érzékeny terület a természetes nyelvű ember-gép kapcsolat kutatása. Itt ugyanis a gépnek nem elég felismernie a nyelv elemeit, valamilyen mértékig meg is kell értenie a felhasználó közlését. „Természetes” kommunikáció esetén – amikor a felhasználó szabadon beszél – ugyanaz a közlés sok különböző formában megjelenhet: a számítógépnek tulajdonképpen valamennyi változatot elemeznie kell, s képesnek kell lennie arra, hogy felismerje bennük ugyanazt a tartalmat. Ennek megvalósítása már jelentős nehézségeket okoz, s ezt csak fokozza, ha a felhasználó nem írja, hanem szóban közli a géppel mondanivalóját.

A számítógép jelenleg képtelen arra, hogy mélységeiben megértse a felhasználó közléseit. Azonban a felhasználót is, a gépet is „be lehet csapni”. Sok olyan, a felhasználóval természetes nyelven kommunikáló program létezik, amelyek nem törekszenek a felhasználói közlés részletes elemzésére; ehelyett a felszínen „utánózzák” a folyamatot, rendkívül egyszerű közelítő eljárásokkal, amelyek révén a felhasználó az esetek többségében mégis úgy érezheti, a gép „megértette” őt. Mindez felveti az emberi nyelv gépi ábrázolásának legalapvetőbb problémáit: a következőkben ezeket próbáljuk meg összefoglalni.

4. A gép a nyelvet szöveggént, sőt betűk sorozataként érzékeli. Kezdjük azzal, hogy a számítógép számára elsősorban az írott szöveg érhető el. A beszélt nyelv gépi ábrázolása egyelőre munkaigényesebb és kevésbé pontos; nehezebben írható át egyértelmű, jól feldolgozható formába. A gép számára az írott szöveg számkódok sorozata, ahol az egyes számkódok betűket és írásjeleket képviselnek. Ha a szöveget nemcsak tárolni és megjeleníteni kell, hanem fel kell ismerni a benne levő nyelvi szerkezeteket is, belépnek a nyelvtechnológia eszközei. Amikor az a kérdés, milyen módon ismerjük fel a szöveg nyelvi szerkezetét, a nyelvtechnológia apaldilemmájához érkeztünk. Ez voltaképpen két irányzatot jelent: a szabály alapút és a statisztikait. Ezeknek az irányzatoknak az elemei a mai kutatásokban egyre jobban keverednek, s a hibrid megoldások jobb eredményeket is szolgáltatnak.

Az alapkérdés az, hogy adunk-e előzetes nyelvi tudást a számítógépnek a szöveg elemzéséhez, és ha igen, milyen mélységben. Az előzetes nyelvi tu-

dás átadása azt jelenti, hogy a számítógép programjába olyan szabályokat írunk, amelyek az ember nyelvi vagy nyelvészeti tudását tükrözik, leképezve a számítógép programozási nyelvének lehetőségeire. Ekkor a számítógépes nyelvész a saját nyelvérzéke vagy nyelvészeti tudása – megfelelő forrásmunkák – alapján fogalmazza meg a szabályokat. A szabályok gépi megfogalmazása általában többé-kevésbé megfelel valamelyik matematikai nyelvmodellnek. Nem szoktuk ide számítani, de voltaképpen emberi háttérismeret az is, hogy az írott szöveg szavakból, mondatokból (bekezdésekből stb.) épül fel. A szavak megkülönböztetése az első körben igen egyszerű: ha ismerjük a szóköz számkódját, elég egy programot végigfuttatnunk a szövegen, amely kiemeli a két szóköz közötti elemeket. Ez a *segmentálási* lépés olyan egyszerű, hogy tulajdonképpen nem is tekintik nyelvtechnológiai műveletnek – pedig egyáltalán nem magától értetődő, magasabb szinten, például a mondatok szétválasztása esetén különösen nem. Amikor a számítógépes nyelvész *nem* ad előzetes nyelvi tudást a számítógépnek (legfőlegbb segít szavakra bontani a szöveget), akkor a gépen olyan eljárásoknak kell futniuk, amelyek fel tudják ismerni a szövegben önmagukban megjelenő szabályosságokat, ismétlődő mintákat. Ezt általában statisztikai számításokkal érik el, de az eredményt gyakran formalizált nyelvészeti információvá alakítják.

A két megközelítés abban különbözik, hogy az első esetben az ember – nyelvérzéke és nyelvészeti tudása alapján – előzetes hipotézist állít fel arról, hogy a szövegekben milyen nyelvi szerkezetek *lehetnek*: ez tulajdonképpen a nyelvi kompetencia felhasználása. A szöveg elemzésekor a számítógép e szabályok jelenlétét vizsgálja a szövegben, s a nyelvész ennek eredményével igazolja vagy veti el a hipotézisét. A második esetben azt vizsgáljuk, milyen jelenségek *vannak* a szövegben, s a felismert mintákból, szabályszerűségekből fogalmazzuk meg nyelvészeti szabatossggal az egyes nyelvi jelenségek leírását. Ez pedig tulajdonképpen a *performancia*, vagyis a létező – nem pedig a lehetséges – szövegek felhasználása. Az utóbbi esetben ahhoz, hogy megfelelő következtetésekhez juthassunk, előbb rendkívül sok szöveg elemzését kell elvégeznünk. A nyelvészeti kutatás céljából összegyűjtött nagy tömegű szöveget *korpusznak* nevezzük, azon – jórészt statisztikai – módszerek együttese pedig, amelyekkel szabályszerűségeket keresünk, a *korpusznyelvészet*. Ebből a tisztán statisztikai eljárások pedig a *nyelvstatisztika* területét alkotják.

A két megközelítés együtt alkalmazható a legjobban. Az előzetesen – a nyelvi kompetencia alapján – átadott nyelvi tudás hátránya, hogy spekulatív jellege miatt nem teljes, s hiányosságai nem jósolhatók meg. Ez a nyelvi tudás mindazonáltal rendkívül értékes, így ha valaki kizárólag a második, korpuszos, statisztikai megközelítést alkalmazza, eldobja a nyelvi kompetenciát, s tulajdonképpen szándékosan mond le olyan tudásról, amelynek révén sokkal jelentősebb eredményekhez jutna.

5. Mindenképpen igaz, hogy a természetes nyelvekben szabályszerűségek vannak, s a számítógép a szövegekből úgy tud – a betűk és írásjelek kódján túl – információt kivonni, ha ezeket a szabályszerűségeket megtalálja benne. Amikor a szövegben felismer egy-egy olyan részt, amely megfelel egy vagy

több korábbról ismert szabálynak vagy mintának, megjelöli valamiféle absztrakt jelöléssel (ennek a nyelvtechnológiában szabványai vannak). Ezt persze csak akkor tudja megtenni, ha a számítógépes nyelvész vagy nyelvtechnológus leírja e szabályok gépi modelljét, vagyis a programot arra, hogyan lehet azonosítani a kérdéses szövegrészeket. Az eredmény olyan eljárások együttese, amellyel a gép fel tud ismerni meghatározott nyelvi szerkezeteket.

A nyelv a gép számára is lehet réteges szerkezetű: külön írják le a szavak, és külön a mondatok viselkedését. A szó és a mondat között további szintek is lehetnek, amelyek a mondatok meghatározott szerkezeti elemeinek felelnek meg. A nyelvtechnológia két legalapvetőbb leírási szintje a számítógépes *morfológia* és a számítógépes *szintaxis*. A gépi morfológia célja, hogy az írott szöveg szavait elemi alkotórészeikre (*morfémáikra*) bontsa, s megjelenítse az egyes alkotóelemek bizonyos nyelvtani tulajdonságait. A gépi morfológia mindig egy meghatározott nyelv szóalakjaival foglalkozik. A legegyszerűbb gépi morfológia olyan szótár, amelyben az adott nyelv szóalakjai vannak felsorolva, mellettük a lehetséges felbontások és a nyelvtani információk. Van azonban olyan nyelvek, ahol a toldalékolás és a szóösszetételek révén olyan sok szóalak jöhet létre, hogy a mai számítógépeken egyszerűen nem férne el a lista. Ilyen nyelv Európában a német is – a bonyolult szóösszetételek miatt –, de a problémák nagy része az olyan agglutináló nyelvekből származik, mint a finn, az észt, a magyar és a török. E nyelvekben elvileg több milliárd különböző szóalak létezik, de más, toldalékolást alkalmazó nyelvekben (például a lengyelben) is tízmilliókra rúghat. Egyszóval, a gépi morfológia bonyolultsága mindig nyelvfüggő.

A morfológiai felbontásból elviekben megtudhatjuk, hogy egy adott szóalaknak mi a szófaja, milyen toldalékokat tartalmazhat. Nagy probléma viszont, hogy ez a részegység (szakszóval: modul) minden szót külön, a környezetéből kiszakítva vizsgál. Emiatt a gépi morfológia nem tudja megmondani, hogy a szóalak ott éppen milyen szerepet tölt be, csak azt, hogy a környezettől függően milyen szerepeket tölthet be. Példa:

termet:

termet [FŐNÉV] = „termet” + [ALANYESET] = „Ø”

terem [FŐNÉV] = „term” + [TÁRGYESET] = „et”

Itt tehát nem tudjuk, hogy a *termet* szó alanyesetéről vagy a *terem* tárgyesetéről van-e szó – mindaddig, amíg meg nem vizsgáljuk a szó környezetét, hogy az adott helyen alanyi vagy tárgyi szerepben van-e szükség erre a főnévre (amelyről tehát még azt sem tudjuk, a kettő közül melyik). Ezt az ember sem tudja, amíg nem látta, mondjuk, a *Kiürítettem a termet!* mondatot. Ez a *többértelműség*: a rendelkezésre álló információ alapján a szónak több elemzése is lehet.

Amikor a számítógép eldönti, az adott környezetben melyik szerepben van szükség a szóalakra, az *egyértelműsítés* műveletét végzi el. Azt az összetett tevékenységet pedig, amelynek során egy nagyobb szövegben először el-

végzi a szavak morfológiai elemzését, majd egyértelműen meghatározza a szó szófaját (kiválasztja a megfelelő felbontást), *szófaji címkézésnek* nevezzük. A szóelemzésben egyébként szerencse, hogy az elemzendő egység – az írott szó – viszonylag kicsi és jól körülhatárolható; mondhatni, „természetes” nyelvi egység a számítógép számára.

A következő szint a mondatelemzés vagy inkább *szintaxis*. A mondattal mint nyelvi egységgel azért kell vigyázni, mert nagyon bonyolult szerkezete lehet. Olyan sokféle mondat létezik, hogy gépi szintaxissal jórészt lehetetlen egy adott nyelven leírt *összes* mondatot teljes egészében elemezni. A gépi szintaxis tehát leggyakrabban a mondatnál kisebb szerkezeteket ismer fel; gyakorlat, hogy elkészítik néhány jól meghatározható, jól elemezhető mondatelem modelljét – nyelvtanát –, majd ezeket az elemeket megjelölik a szövegben, ezzel mintegy kiemelve belőle a mondatok kulcselemeit. Ha abban a mondatban, hogy *A francia király fogadta a török követet*, ismerjük a szavak szófaját: A[NÉVELŐ] *francia*[MELLÉKNÉV] *király*[FŐNÉV] *fogadta*[IGE] *a*[NÉVELŐ] *török*[MELLÉKNÉV] *követet*[FŐNÉV]. Levonhatjuk például azt a következtetést, hogy az alanyi, illetve tárgyi szerepben levő szócsoport egyformán a NÉVELŐ + MELLÉKNÉV + FŐNÉV szerkezetet mutatja. Az ilyen szócsoport a *főnévi csoport*; ha egy szövegből kiemeljük a főnévi csoportokat, tulajdonképpen összegyűjtöttük a szövegben szereplő *dolgokat* (tárgyakat, személyeket, fogalmakat, eseményeket stb.). A főnévi csoport így a nyelvtechnológiában az egyik legfontosabb nyelvi alakulat.

A számítógépes nyelvész a fenti minta láttán tehát úgy dönt, hogy a későbbiekben minden szócsoportot, amely szófaji szerkezetében a NÉVELŐ + MELLÉKNÉV + FŐNÉV mintát mutatja, főnévi csoportként kezel. Másképpen: leír egy szabályt, amely szerint a NÉVELŐ + MELLÉKNÉV + FŐNÉV sorozat egy főnévi csoportot alkot. Ugyanez a gyakorlat természetesen más típusú mondatelemekre is alkalmazható.

A *nyelvtan* tehát a fenti módon leírt szabályok, azaz azon minták együttese, amelyek így vagy úgy meghatározott nyelvi alakulatot (például főnévi csoportot) alkotnak. A szabályok lehetnek rendkívül bonyolult szerkezetűek, s sok információt tartalmazhatnak a nyelvi alakulatokról és alkotóelemeikről is; a statisztikai megközelítésben viszont csak azt vizsgáljuk, hogy adott szavak vagy adott szófajú szavak milyen gyakorisággal fordulnak elő együtt, s a gép ekkor a számok alapján próbálja megmutatni a kérdéses mondatelemeket. Az utóbbi esetben nem beszélünk szabályokról. A nyelv szabály alapú modellezésének egyik szabatos matematikai modellje a *Mondatok* című fejezetben bemutatotthoz hasonló nyelvtani rendszer. Természetesen számos további, ezt kiegészítő vagy ezzel versenyző matematikai leírás létezik. Amikor egy adott nyelvi jelenséget modellezünk (például nyelvtant írunk a főnévi csoportok felismerésére), beszélnünk kell a modell *minőségéről* is. A gépi modell mindig csak felszínes közelítése a nyelvben ténylegesen előforduló szerkezetek halmazának. A modellt egyfelől az jellemzi, hogy a szövegekben előforduló szerkezetek mekkora hányadát ismeri fel: ez a modell *fedése*. A másik fontos jellemző a *pontosság*: ez azt méri, hogy a felismerni vélt

szerkezetek közül hány százalék helyes. Egyetlen nyelvmodell sem képes valamennyi kérdéses szerkezetet azonosítani, s minden nyelvmodell felismerni vél olyan szerkezeteket is, amelyek nem tartoznak a megcélzott típushoz. Ha nyelvtanunk kizárólag a NÉVELŐ + MELLÉKNÉV + FŐNÉV szerkezetet tekinteti főnévi csoportnak, akkor az *öreg francia király* főnévi csoportot már kihagyja, hiszen annak szófaji szerkezete NÉVELŐ + MELLÉKNÉV + MELLÉKNÉV + FŐNÉV; ugyanakkor *A vörös Péter kedvenc színe* mondat elején NÉVELŐ + MELLÉKNÉV + FŐNÉV áll, de nem formál egyetlen főnévi szerkezetet.

Összegzésül: a nyelv szerkezetét a számítógépen úgy tudjuk ábrázolni, hogy megpróbáljuk valahogyan leírni, létrehozni a lehetséges szóalakok, mondatok vagy más nyelvi szerkezetek halmazát. Ha egy halmazt nem tudunk felsorolni – vagy azért, mert túl sok eleme van, vagy azért, mert munkaiigényes –, matematikai szabályrendszert s erre épülő számítógépes eljárást alkotunk, amely képes előállítani és felismerni a kérdéses nyelvi elemeket. Ez az eljárás a legtöbbször nem pontosan a kívánt halmazt állítja elő, annak csak közelítése lesz. A fentebb említett *fedés* és *pontosság* e közelítés minőségét határozza meg.

6. A számítógépnek azt a legnehezebb megtanítani, hogy összekapcsolja a nyelv elemeit – szavait, mondatait – és a világ dolgait. A szavak és mondatok valódi tartalmának felismeréséhez és megfelelő kezeléséhez világismeretre van szükség. Ez idáig túl bonyolultnak bizonyult ahhoz, hogy egyszerű, jól kezelhető számítógépes/matematikai modellt készítsenek hozzá. Léteznek azonban olyan modellek, amelyekkel a számítógép egy keveset „megérthet” a természetes nyelvű szövegekből. A problémát nemcsak az okozza, hogy a betűsorozatok vagy a beszélt nyelvet alkotó hangsorozatok önmagukban gyakran többértelműek. Ha csak annyira van szükség, hogy felismerjük a *keres* szó különböző jelentéseit *A kormány szakembereket keres* és a *Péter pénzt keres* mondatokban, ezt a számítógép egyszerűen megteheti: elég csak megnézni, hogy a *keres* ige mellett a *pénzt* tárgy áll-e. Különben ezt a műveletet nevezi a nyelvtechnológia *jelentés-egyértelműsítésnek*. Az előző részben volt szó a *szófaji egyértelműsítésről*: a jelentés-egyértelműsítés némiképp túlmutat ezen, de nem követeli meg, hogy a számítógép valóban mélységeiben kezelje a szavak jelentését.

A feladat akkor válik bonyolulttá, amikor a gépnek az információkeresésben vagy a fordításban kell segítenie. Szándékosan írtunk információkeresést szövegkeresés helyett: ekkor ugyanis a számítógépnek nem a szövegezés, hanem az információtartalom alapján kell megtalálnia dokumentumokat vagy rögzített szerkezetű adatokat. A szövegezés felhasználásával semmire sem jut, mert ugyanazt az információt teljesen más szövegezéssel – például egy másik nyelven – is le lehet írni. Itt tehát a különböző szövegezéssel megfogalmazott dolgok közötti egyezést vagy hasonlóságot kell felismerni. A fordítás esetén pedig azzal kell szembenéznünk, hogy a különböző nyelvek szavai – jelentésüket tekintve – nem feleltethetők meg pontosan egymásnak. Pusztán szótárral *jól* fordítani tehát szinte lehetetlen: a fordítással úgynevezett kommunikációs ekvivalenst, vagyis olyan szöveget kell létrehozni,

amely ugyanazt vagy közelítőleg ugyanazt mondja a célnyelven értő olvasónak, mint a forrásnyelv beszélőjének. Nyilvánvaló, hogy ez túlmutat a szavak vagy nyelvtani szerkezetek egymásnak való megfeleltetésén – a gépi fordítás kezdeti szakaszai ezért is nem jártak sikerrel.

Amikor a nyelvtechnológia nehézségeiről beszélünk, általában azt mondjuk, hogy a számítógép nem tud mit kezdeni a szavak, mondatok jelentésével. Ez ebben a formában nem is igaz: a számítógépnek valójában „csak” világismerete nincs. A szavak, mondatok jelentésének közelítő ábrázolását a kutatók általában azzal könnyítik meg, hogy korlátozzák a feldolgozható szövegek témáját. Ha minden szöveg, amellyel foglalkoznunk kell, jól meghatározható témába (informatika, jog, pénzügy stb.) tartozik, sokkal kevesebb többértelműséget találunk benne; sőt még a szövegek szerkezete is sajátos (lehet), s nem annyira szerteágazó, mint az általános esetben (amikor tulajdonképpen minden elképzelhető szövegre fel kell készülnünk). A téma szűkítésével elérjük, hogy a számítógép látszólag már akkor is „érti” a szöveget, amikor csak az adott terület terminológiáját és a kifejezések közötti kapcsolatot adjuk meg neki. Azonban így sem mondhatjuk azt, hogy a feladat nagyon könnyűvé válik. Meghatározott témára is nehéz olyan információkereső vagy fordítórendszert létrehozni, amely a szavak jelentését is felhasználja valahogy. Amikor tehát a kutatók összetett nyelvtechnológiai eszközt készítenek, a feldolgozott szövegek körét legtöbbször valamelyik szakma nyelvére – vagy újságcikkekre – korlátozzák. Szépirodalmi szövegek számítógépes feldolgozása nem jön szóba: nemcsak azért, mert nehéz feldolgozni ezeket a szövegeket, hanem azért is, mert esztétikai megfontolások is ellene szólnak. Ha a számítógépnek még tárgyilagos világismerete sincs, hogy tudna bármit hozzátenni a szépirodalmi szöveghez, amelynek főleg az a célja, hogy az olvasóban szubjektív élményt keltsen?

A szavak jelentésének felszíni ábrázolásában divattá vált – s teljes joggal – az a módszer, amely a szavak egymással való kapcsolatát írja le nagyon egyszerű eszközökkel, nem mindig a szavak mögötti dolgokat próbálja meghatározni bonyolult, elvont modellek segítségével. Olyan egyszerű kapcsolatokra gondoljunk, mint a hasonlóság vagy azonosság, vagy éppen az adott kategóriához tartozás, akár úgy, hogy adott szó magát a kategóriát, akár úgy, hogy egy, a kategóriához tartozó kisebb dolgot ír le. Az *asztal* és a *bútor* közötti kapcsolat például egyrészt azt mondja, hogy az *asztal* a *bútor* kategóriába tartozó szűkebb dolog, míg a *bútor* olyan tágabb fogalom, amelybe az *asztal* is beletartozik. Egymáshoz képest tehát mindkét szó meghatározható. Ezek a kapcsolatok láncba is fűzhetők: ha az *asztal*-*bútor* pároshoz harmadiknak hozzávesszük a *berendezési tárgy* kifejezést, akkor olyan kategóriát kapunk, amelynek tagja a *bútor*, s *bútorsága* révén az *asztal* is. A szavak között e kapcsolatokkal bonyolult háló is létrehozható; az ilyen hálót *szóháló*nak nevezhetjük. Ezek a hálók pedig egyszerűen leírhatók számítógépes adatbázisokban. A szóháló alkalmazásával a számítógép tudomást szerez például az *asztal* *bútorságáról*, anélkül hogy bármelyikről olyan mentális képe lenne, mint az embernek. A *bútorságon* keresztül pedig meg tudja határozni, mi a

közös az *asztal*-ban és a *szekrény*-ben, s ha ezt természetes nyelvű szövegek elemzésekor felhasználjuk, az már némiképp arra emlékeztet, mintha a gép elvont módon is „megértené” a szöveget.

7. Említettük, hogy a szöveg a számítógép számára írott szöveget jelent. Ugyanakkor nem lehet kézlegyintéssel elintézni azt az igényt, hogy a számítógép a beszéddel is kezdjen valamit: egyrészt beszéljen az emberhez (aki ezért vagy azért nem tudja elolvasni az írott szöveget), illetve értse meg az ember beszédét (amikor annak nincs lehetősége a szöveg leírására). A számítógép – a beszédet illetően – még gyerekkorát éli. A beszéd gépi feldolgozásának valamennyi kísérlete az írott szövegre próbálja visszavezetni a problémát: szöveget beszéddé alakít – ez a *beszédszintézis* –, illetve beszédet szöveggé – ez a *beszédfelismerés*. A két feladat nem egyformán nehéz: a beszédszintézissel eddig több eredményt értek el, mint a beszédfelismeréssel.

Az egyes hangok fizikai tulajdonságai is leírhatók számokkal, s e számok alapján a számítógép nagyjából létre tudja hozni a megfelelő hangot. Csak az a baj, hogy a hangokat nem oly egyszerű számokká kódolni, mint a betűket; ha nagyon egyszerűen fogalmazunk, azt kell mondanunk, hogy – szemben a betűkkel – a hangok végtelen sokan vannak. Szerencsére a betűk fonémákat „kódolnak”, amikhez ügyesen hozzárendelhetjük a környezettől függő hangot. A géphang persze sokszor idegenül szól, sőt néha meg sem értjük. Sokat számít a beszédben az – írásban nem jelölt – hangsúlyozás is: mondat elején, végén járunk-e, kérdezőnk, kiabálunk vagy előadunk, azaz a *hanglejtés*. Nem véletlen, hogy a számítógépes beszéd hangtanával és a tényleges beszédhangok szabatos – betűszerű – leírásával külön tudományág foglalkozik: az előbbi a számítógépes *fonológia*, az utóbbi a számítógépes *fonetika*.

Ha az a célunk, hogy a számítógép érthetően olvasson fel szövegeket – arról tehát még ne beszéljünk, hogy kellemes női hangja lesz-e –, a gépet meg kell tanítanunk a hanglejtés utánzására is. A hanglejtés helyes kezeléséhez például tudni kell, hol járunk a szón, a mondaton belül. Ennek meghatározása nemigen lehetséges másképp, mint a szöveg nyelvtani elemzésével. Más a helyzet, ha a gépnek nem tetszőleges szöveget, hanem egyszerűbb dallamszerkezetű dolgokat – például telefonszámokat – kell elmondania: bár a felolvasott telefonszám lehet kissé idegenszerű, de tökéletesen érthető lesz.

Egy érdekesség: az egyik hazai mobiltelefon-szolgáltató üzemeltet egy olyan szolgáltatást, amely beolvassa a telefonba az ember elektronikus leveleit. Szokásos felolvasórendszerrel lenne szó, ha nem kellett volna felkészülni arra, hogy a magyar nyelven írt elektronikus levelek jó részében nincsenek ékezetes betűk. Ahhoz, hogy a felolvasott szöveg érthető legyen, előbb ékezetesíteni kell: készült tehát egy olyan program is, amely az *ékezetes betűk* szavakat átírja *ékezetes betűk*-ké.

A gépi beszéd-előállítás legfontosabb problémáját megoldották: ma már nem okoz igazi gondot olyan alkalmazás készítése, amely érthetően felolvas tetszőleges szövegeket. A géphang „természetességén”, hanglejtésén van még mit javítani, de a problémák korántsem akkorák, mint a gépi beszédértés esetén.

8. A gép számára egyelőre nagyon nehéz a beszédet írott szöveggé alakítani. A beszéd felismerését a hangok sokfélesége és óhatatlanul pontatlan ábrázolása mellett számos „zavaró tényező” is gátolja. Minden ember hangjának más tónusa van: ha tehát két ember próbálja mondani ugyanazt, a gép számára ez sohasem lesz egyforma. Ezt az akadályt még le lehet győzni, mert a beszédhangoknak vannak bizonyos egyértelműen azonosítható elemeik – a teljesen pontos ábrázolásra tehát nincs feltétlenül szükség ahhoz, hogy a gép azonosítani tudja őket. Következően ott a háttérzaj: ettől el kell választani a beszédhangot. Digitális jelfeldolgozási eljárásokkal még ezen is segíthetünk úgy-ahogy. De az ember nem ott tart közben szüneteket, ahol az írott szövegben a szóközök vannak: ha fel is ismertük a beszédhangokat, nem tudhatjuk, hol van vége a szónak, a frázisnak vagy a mondatnak. Ez a *szegmentálás* problémája. Ha az ember, akit a gépnek meg kellene értenie, még beszédhibával is küszködik, végképp nehéz lesz a hangok azonosítása, mert egyes beszédhangok alapvető fizikai tulajdonságai lesznek mások.

A gépi beszédfelismerő rendszerek készítői három ponton próbálják egyszerűsíteni a problémát. Itt is azt teszik, mint a szöveg jelentésének ábrázolásában: ha nehéz az általános problémát megoldani, szűkítik a szóba jöhető beszédminták körét. Három dolog okozhat problémát: a beszéd folytonossága, a beszélők sokfélesége és a nagy szókincs. Korlátozható tehát a folytonos beszéd: vannak olyan beszédértő rendszerek, amelyek egyszerre egy szóval tudnak kezdeni valamit: ezeket nevezzük *izolált szavas* rendszereknek. Ha tudjuk, hogy a beszélő legfeljebb egy szót mondhatott, sokkal kevesebb lehetőség közül kell kiválasztani a megfelelőt. Ha az a feladat, hogy egy gépet egyszerű parancsszavakkal irányíthassunk, ez a megoldás tökéletesen megfelel. Ugyancsak csökken a lehetőségek száma, ha nem beszélhet akárki a rendszerhez. A beszélőfüggetlen rendszer egy beszélő szavait „érti” csak meg: megtanulja e beszélő tónusát és hanglejtését; ettől kezdve egy betűnek lényegesen kevesebbféle hang felel meg, illetve az egyes hangok egyszerűbben is ábrázolhatók számokkal. Az olyan beszédértő rendszer, amelyhez bárki beszélhet, a *beszélőfüggetlen* rendszer. A szókincs korlátozása pedig nemigen szorul magyarázatra.

A mai „beszédértő” rendszerekben a fenti három probléma közül a szóba jöhető beszédmintákat legalább az egyik szempontból korlátozzák. Ennek ellenére a folytonos beszéd felismerése terén viszonylag szerények az eddigi eredmények: a probléma ráadásul nyelvfüggő. A meglehetősen drága fejlesztés kisebb beszélőszámú nyelvek esetében, sajnos, gazdaságilag nehezen indokolható.

Folynak olyan kísérletek is, amelyek során a nyelvi elemzés eszközeit nemcsak a felismert beszéd utólagos feldolgozására használják, hanem a beszédfelismerés pontosságának javítására. Ha a gép ezt hallja: *dobdideavaszgójót*, egy speciális szóelemző program segítségével fel tudja bontani, és átírni így: *dobd ide a vasgolyót*. A kutatók várakozása szerint ez a fajta visszacsatolás úgy növeli meg a beszédfelismerés pontosságát, hogy közben enyhíti a „klasszikus” beszédfelismerő egységgel szembeni elvárásokat: annak nem

kell többé teljesen egyértelmű választ adnia, hiszen a nyelvi rendszer segít kiválasztani a megfelelőt.

9. A nyelvtechnológiának hazánkban is jelentős eredményei vannak, annak ellenére, hogy a hazai kutatóhelyek száma és nagysága is kisebb, mint sok más országban. A számítógépes feldolgozás szempontjából a magyar nyelv egyes szempontok – például a szavak felépítése – tekintetében bonyolultabb, mint az indoeurópai nyelvek nagy része. Európában a nyelvtechnológiai kutatások túlnyomó része valamelyik indoeurópai nyelv területén folyik: nem is lehet ez másképp, hiszen itt kevés nyelv tartozik más nyelvcsaládokba. Ugyanakkor a nem indoeurópai nyelvek közül a jelentősebb európai agglutináló nyelvek (finn, magyar, török) területein jelentős eredmények születtek, különösképpen a szavak szerkezetének – a gépi morfológiának – a kutatásában. Az a szóelemző program tehát, amelyet hazánkban elsősorban a magyar nyelv céljaira fejlesztettek ki, szükségképpen fejlettebb, mint mondjuk egy angol nyelvi elemzőprogram, s ekképp – legalábbis a magyar nyelvhez képest – szinte gyerekjáték benne leírni az angol, a német vagy épp a cseh és a lengyel szavak szerkezetét.

A magyar mondatban gépi leírása azonban már nem áll ilyen jól. A magyar leíró nyelvészet hagyományai kevésbé alkalmazkodnak a számítógép igényeihez. Ez a helyzet változóban van, ugyanis intenzív kutatások folynak, amelyek segítségével a számítógép néha még mélységeiben is fel tudja deríteni a magyar mondatok szerkezetét.

10. Az előzőekben mindvégig arról írtunk, milyen nehéz feladat a számítógép számára az emberi nyelvvel bánni. Holott vannak olyan nyelvi szolgáltatások, amelyeket a nagyközönség is használ, hovatovább évtizedek óta. Ideje tehát, hogy arról is szóljunk, mit ad a felhasználóknak a nyelvtechnológia – a gépek korlátozott képességei ellenére.

Ma már százazrek, talán milliók használnak szövegszerkesztő programokat Magyarországon. E programok mindegyike tartalmaz olyan nyelvi modulokat (részegységeket), amelyek a „jól formált” – jó helyesírással írt, a sorok végén helyesen elválasztott stb. – szöveg írásában segítik a felhasználót. Az írást segítő szolgáltatások közül a *helyesírás-ellenőrző* programok állnak az első helyen. Két fajtájuk van: a *szóellenőrző* és a – mondjuk így – *szóhatáron túli nyelvhelyességet ellenőrző*. Szóellenőrzőt szinte mindenki használ, a nyelvhelyesség-ellenőrző program azonban nem áll mindenhol rendelkezésre.

A *szóellenőrző* program látszólag azt vizsgálja, helyesen írtunk-e egy-egy szót a szövegben; időnként jelzi, hogy nem, és javítást is ajánl – és időnként téved is. Miért? Mert valójában nem a szavak helyesírását ismeri, csak valami módon tud azokról a szavakról, szóalakokról, amelyek léteznek, létezhetnek a nyelvben. Ez a „legtöbbször” általában egy olyan szótárt jelent, amelyben a program készítői által helyesnek tartott szóalakok fel vannak sorolva. Vannak azonban nyelvek – ilyen például a magyar is –, amelyekben olyan sok szóalak van, hogy a gépen el sem férne az a szótár, amelyben felsorolnánk őket. Ilyen esetben a szóellenőrző program *morfológiai elemző* modult hív segítségül, amely a szóalakokat különböző részek (a morfémák) kombi-

nációjaként ismeri fel. Ennek a modulnak az adatbázisa – ha mondjuk ötmilliárdféle lehetséges toldalékolt magyar szóalakot ismer – már nem nagyobb, mint például az angol nyelv összes (kb. félmillió) szóalakjának szótára.

Amikor a szóellenőrző program úgymond „hibásnak” talál egy szót, igazából csak azt jelzi, hogy az nincs benne a szótárában (illetve a beépített morfológiai elemző modul nem ismeri fel). Nem a *helyesírást* ellenőrzi tehát, arról pedig végképp nincs szó, hogy helyettünk tudná a (magyar) nyelv helyesírási szabályait. A szöveget sem javítja automatikusan: a javítási javaslatokat a felhasználónak jóvá kell hagynia. Nagyon helyesen, hiszen a javaslatok között sok oda nem illő dolog lehet. A szóellenőrző program tehát nem „javítja meg” a felhasználók helyesírását, viszont kiválóan alkalmas a gépelési hibák jelzésére és javítására. Általában azt mondjuk, hogy a szóellenőrző program a hibásan írt szavak 95%-át jelzi, azt viszont már a felhasználónak kell tudnia, hogyan kell ezeket kijavítani. Tehát minél tudatosabb, jobb valakinek a helyesírása, annál több hasznát veszi a helyesírás-ellenőrző programoknak.

A szóellenőrző program egyszerre egy szót lát. A szó ebben az esetben az a szövegrész, amely szóköztől szóközig vagy központozási jelig (írásjelig) terjed. Ez azt jelenti, hogy nem tudja vizsgálni a szó környezetét avégett, hogy kiszűrje az oda nem illő javítási javaslatokat, vagy észrevegye a hibásan különírt szavakat. Erre a szóhatáron túli nyelvhelyesség-ellenőrzés alkalmas.

A *szóhatáron túli nyelvhelyesség-ellenőrzést* a legtöbb alkalmazás *nyelvtani ellenőrzésnek* mondja, és többnyire teljes mondatokat próbál feldolgozni. Nagyon eltérő jellegű és minőségű a különböző nyelvekhez készített nyelvhelyesség-ellenőrző programok működése: például az egyetlen magyar nyelvhelyesség-ellenőrző program nem a mondatok teljes nyelvtani elemzésével dolgozik. A program inkább meghatározott helyesírási, nyelvhelyességi, stílusbeli hibákat, hiányosságokat keres, a mondatok felszíni átvizsgálásával. Ilyen hiba, hiányosság a szavak hibás különírása, a vessző hibás alkalmazása vagy kihagyása, vagy éppen a nyomdafestéket nem tűrő szavak használata. A magyar nyelvhelyesség-ellenőrző program fejlesztői az *Akadémiai helyesírási szabályzatból* kiválogatták azokat, amelyekre gépi szabályt is lehetett írni, így a program által jelzett hibák és javaslatok összevethetők a szabályzattal. A nyelvhelyesség-ellenőrző programról tudni kell, hogy működése sokkal bizonytalanabb, mint a szóellenőrző programoké. Ez nem minőségi probléma, egyszerűen a hibák jellege nem teszi lehetővé a teljesen biztos gépi ítéletet. Emiatt a program még jobban a felhasználó tudására bízza a javítást, csak azt jelzi, hogy az adott helyen „talán” hiányzik egy vessző, vagy hibásan különírtunk két szót, és így tovább. Még inkább igaz tehát, hogy a gép nem „tudja” helyettünk a helyesírást.

Nem annyira az írást, inkább a szövegek nyomdai előkészítését támogatják az *elválasztó* programok. Az elválasztás nem kötelező, ám jelentősen javítja a nyilvánosságnak szánt dokumentumok külső megjelenését. Az elválasztás hiánya nem helyesírási hiba, ha viszont úgy dönt valaki, hogy a sorok végén elválaszt szavakat, azt nem teheti meg a szavakon belül akárhol: a hibás elválasztás már hibás helyesírás! A magyar nyelvben elsősorban a szóta-

golás szabályai határozzák meg az elválasztást, de ez alól vannak kivételek, például hogy az összetett szavakat az összetételi határon kell elválasztani. Mivel a sorok elejére csúszó szavakat nagyon munkaigényes dolog egyenként, kézzel elválasztani, s figyelni, mennyi fér be belőlük az előző sor végére, szükségünk van automatikus elválasztó programra. A programnak viszont ismernie és alkalmaznia kell az elválasztási szabályokat, hogy hasznát lehessen venni. Így például, ha szükséges, ellenőriznie kell, hogy összetett szóval találkozott-e, s ha igen, hol vannak benne az összetételi határok. E feladathoz is morfológiai elemző programot célszerű használni (s a magyar elválasztó programok használnak is ilyen modult), amely a szóalakokat alkotóelemeikre (morfémákra) tudja bontani.

A választékos fogalmazást segíti a számítógépes *szinonimaszótár*, amely a szöveg egyes szavaihoz rokon értelmű megfelelőket ajánl. Az ilyen programok inkább *tezauruszok*, amelyek a számukra ismert szavakat fogalomkörökbe rendezik. Minden szó egy vagy több fogalomkörhöz tartozik: a program a rokon értelmű megfelelők keresésekor előbb meghatározza a lehetséges fogalomköröket, majd megkeresi az adatbázisban a hozzájuk tartozó összes szót. A szinonimaszótárak ezzel szemben szócikkbe vannak rendezve, ahol minden szó mellett megtaláljuk a rokon értelmű megfelelőket. Ha valamelyik ismert szövegszerkesztőben elindítjuk a magyar szinonimaszótárt, láthatjuk, hogy a program a fogalomköröket is meghatározza, tehát inkább tezauruszként működik.

A szinonimaszótár-tezaurusz adatbázisában a szavak *szótári alakjukban* szerepelnek, míg a szövegben többnyire nem. Amikor egy adott szóalakhoz szinonimákat kérünk, a szövegbeli alak gyakran *nincs* benne azonos formában a tezaurusz adatbázisában. A programnak a szóalakokból előbb elő kell állítania a szótári alakot. A legegyszerűbb most is részekre – morfémákra – bontani a szóalakot, vagyis morfológiai elemzést végezni rajta. A szótári alakot pedig akkor kaphatja meg a rendszer, ha arról is információt kap, hogy a szavak egyes alkotóelemei közül melyek tartoznak a szótári alakhoz, és melyek nem. Azt is mondhatjuk, hogy a szinonimaszótár-tezaurusz ekkor a szóalakok tövét határozza meg. Ez a művelet a *szótővesítés* vagy *lemmatizálás*, azaz a járulékos elemek, például a toldalékok eltávolítása, a szótári alak visszaállítása.

Említettük, hogy a szinonimaszótár-tezaurusz adatbázisában a szavak szótári alakjukban szerepelnek. Arra a kérdésre már találtunk választ, hogyan találjuk meg a szótári alakot, ha a szövegben nem úgy szerepel a szó. Amikor viszont kiválasztottuk a megfelelő szinonimát, azt vissza is kell írni a szövegbe. A szövegszerkesztők ilyenkor általában csak a szótári alak visszairását teszik lehetővé; ha az eredeti szó toldalékokkal együtt volt a szövegben, a szinonima beírásával ezek elvesznek. Létezik olyan szinonimaszótár-program is, amely képes a kiválasztott szinonimát toldalékolts formában visszairni a szövegbe, a megfelelő morfológiai generáló modul segítségével.

11. Az információkeresés alproblémájaként mostanában mindig az interneten felgyűlt, elképzeltetlenül nagy mennyiségű szöveget emlegetjük. Való

igaz, szinte nincs olyan téma, amelyhez az internet, amelyen több millió ember több milliárd dokumentumot helyezett el, ne tudna hozzászólni. A megfelelő *tartalmú* dokumentumokat azonban nagyon nehéz megtalálni, mert aki keres, tudnia kell, milyen szavak, kifejezések fordulhattak elő a kérdéses dokumentum(ok)ban, milyen nyelven lehetett stb. A korszerű internetes keresőszolgáltatások sok felszínes művelettel segítik a keresés szűkítését (néha ugyanis az is probléma, hogy a begépelte kulcsszavakra válaszul több millió dokumentumot kapunk). Az igazi megoldás azonban az volna, ha a megfelelő *tartalmú* dokumentumokat találhatnánk meg az általunk *begépelte szavakat szó szerint tartalmazó* dokumentumok helyett. Mindez persze nemcsak az internetes keresés során probléma, bár ott jelenik meg a nagyközönség számára. Szakmai körökben, ahol tematikus információkeresésre van/volna szükség, ugyanúgy fejlesztésre szorul az információ tartalom szerinti megkeresése.

Nézzük, hogyan segíthet a nyelvtechnológia az információkeresésen! Számos fejlesztés folyik: elképzelhető, hogy egy-két éven belül e szolgáltatások a keresők szerves részeivé válnak. Először is, ha szavakat keresünk, találjuk meg a toldalékolt alakokat! Amikor begépeljük a keresendő szavakat, a legritkábban írjuk be az összes lehetséges alakot; s a rendszer nem feltétlenül találja meg az eltérő szóalakokat. Ha például az almatermesztéssel foglalkozó szövegeket keressük, és az *alma* szót tartalmazó összes dokumentumot szeretnénk megtalálni, nem találjuk meg az *almával* szóalakot, ha csak az *alma* szót adtuk meg a keresőablakban. Ha viszont az *alma* és az *almá* tövet is megadjuk, a rendszer az *almárium* szót is megtalálja, pedig arra nem is vagyunk kíváncsiak. A megoldás: olyan keresőrendszer, amely a szóalakokat intelligens módon egymáshoz rendeli. Ez azt jelenti, hogy mind a dokumentum szövegében levő, mind pedig a felhasználó által begépelte szavakhoz megkeresi a szótári alakot, s ezeket hasonlítja össze. Ebben segít a szótövesítő program. Ha azonban például a kutyatartásról keresünk szövegeket, és a *kutya* szó különböző alakjait keressük, nem kapjuk meg azokat a helyeket, ahol az *ebadó*-ról van szó. Ez akkor lenne csak lehetséges, ha a keresőprogram nemcsak a begépelte szavakat, hanem azok szinonimáit is keresné.

Bonyolítsuk tovább a problémát! Kedvencünket Németországba szeretnénk vinni kutyakiállításra. Ha keresőrendszerünk meg is találja a kutyákról és az ebekről szóló oldalakat, még mindig nem kapjuk meg a kutyakiállítások németországi szabályait, amihez legalább a *Hund* szó volna keresendő. Jó lenne tehát, ha a különbözőképpen toldalékolt szóalakok és a szinonimák mellett a kereső megtalálná a szavak idegen nyelvű megfelelőit! A megoldás: két-, illetve többnyelvű szótárak alkalmazása, amelyek megmondják a *kutya* szó különböző idegen nyelvű megfelelőit. Ha azonban megkaptuk az idegen nyelvű oldalakat – mondjuk a németországi kutyakiállításokról szólókat –, bajban vagyunk, ha nem tudunk németül: nem tudjuk elolvasni őket. Ekkor segítené egy automatikus fordítórendszer, amely érthető magyar fordítást adna az eredeti német szövegről.

Nyilvánvaló, hogy a keresőrendszerek nem fogják egyszerre és egyik napról a másikra megvalósítani az összes fenti szolgáltatást. Különösen igaz ez a

gépi fordításra, amelynek fejlesztése még nem tart ott, hogy bármely két nyelv között jól olvasható, érthető fordítást tudna adni. Fokozatosan azonban minden bizonnyal megjelennek majd. E szolgáltatásokat azért nem soroltuk a nyelvtechnológia jövőjéhez, mert az alkalmazott nyelvtechnológiai eszközök nagy része – a fordítógép kivételével – bevált, hagyományos eljárásoknak mondhatók.

Amikor a tartalom szerinti szövegkeresés problémájáról beszélünk, ismét csak a fedésről és a pontosságról van szó. Szeretnénk, ha a kereső minden olyan dokumentumot visszaadna, amely az általunk kívánt témáról szól, és egyet sem, amely nem tartozik a tárgyhoz. Ennek megítélése azonban nagyon nehezen mérhető. Míg a nyelvtani modellek esetén meglevő szövegekben mérhetjük a gépi nyelvtan „okosságát”, az internetes keresés során a felhasználó dönti el, egy adott dokumentum megfelel-e az igényeinek. Semmilyen objektív adatunk nincs arról, hogy egy adott dokumentum mennyire vág a kívánt témába – látni azonban látjuk a hiányosságokat, amikor magunk keresgélünk a hálón.

A tartalom szerinti keresés ideális megoldása az lenne, ha a felhasználó teljesen szokásos kérdést tehetne fel a rendszernek, s a gép a kérdés *jelentése*, nem pedig a benne levő *szavak* alapján keresné a dokumentumokat. Ilyen kutatások is folynak; a közeljövő alkalmazásai között ejtünk róluk néhány szót.

12. A gép még nem tud az ember *helyett* fordítani. Azonban az elmúlt években – többek között az Európai Unió bővítése és az Unión belüli szorosabb együttműködés miatt – olyan sok fordítanivaló keletkezett, hogy az embernek – a fordító embernek – minden elképzelhető gépi segítségre szüksége van, hogy munkáját határidőre és megfelelő minőségben el tudja végezni. Az automatikus gépi fordítás alkalmazása itt nem jön szóba, mert a jelenlegi rendszerek közel sem képesek kiadható fordításokat készíteni; sőt, az általuk lefordított szövegek nyersfordításnak sem jók, kijavításuk több munkát igényel, mint az eredeti szöveg újrafordítása. Nem számíthatunk arra, hogy ez a helyzet lényegesen megváltozik a következő egy-két évben.

A kilencvenes évek elején viszont megjelentek s napjainkra elterjedtek a számítógépes fordítástámogató eszközök. Ezek közül a legegyszerűbbek a számítógépes szótárak. Ezek legtöbbször korábban nyomtatásban megjelent szótárak gépi adatbázisba írt megfelelői. A hozzájuk mellékelt szótárprogram sem alkalmas többre, mint egy-egy szó megkeresésére, amely a nyomtatott szótár fellapozásának felel meg. A gép kicsit gyorsabban megtalálja a keresett szót – mint az ember a papírszótárban –, de a hagyományos gépi szótártól nem kap több segítséget. Vannak azonban olyan szótárprogramok is, amelyek – bár továbbra is nyomtatott szótárak elektronikus változatait tartalmazzák – számos nyelvi többletszolgáltatást nyújtanak. A legegyszerűbb ezek közül a szótővesítés alkalmazása: a szótárakban a címszavak szótári alakban szerepelnek, de a szövegekben nem; a szótárprogram a szótővisszaállítás révén képes lehet arra, hogy a toldalékolt szóalakokat is megtalálja a szótárban, visszavezetve őket a szótári alakra. Nagy segítség az is, ha a szótárprogram egyszerre több szótárt kezel: ha egy szót egyyszer kell

begépelni, de a program sok szótárban keresi meg egyszerre, akkor már úgy dolgozik, ahogyan az ember sohasem tudna: mintha egyszerre tíz-húsz szótárt lapoznánk és olvasnánk.

Ha a szótárt nem fordításra, hanem idegen nyelvű szöveg megértésére kell használni, akkor nem a „hagyományos” szótárprogramra, hanem az úgynevezett gyorsfordítóra van szükség. Ez a program „rejtőzködik” a számítógépen: munka közben igazából nem látjuk működni. Amikor azonban rámutatunk egy szóra az idegen nyelvű szövegben – a konkrét programtól függ, hogy kell-e kattintani –, működésbe lép a szótári keresés, és a képernyőn kis buborékban megjelenik a szó fordítása. Sőt: az ilyen programok általában azt is láthatják, milyen szavak vannak a kiválasztott szó mellett, így nemcsak a szóra, hanem az őt tartalmazó kifejezésre is tudnak fordítást adni.

A fordító munkáját legjobban az összetett fordítástámogató programok segíthetik. Ezek a programok szótár helyett úgynevezett *terminológiakezelő rendszer* használnak. A terminológiakezelő rendszer adott szakmák, szakterületek szókincsét tartalmazza, lehetőleg úgy, hogy egy idegen nyelvű szónak, kifejezésnek, csak egy fordítása legyen. A fordítástámogató rendszerek a terminológiai adatbázis segítségével ki tudják szótározni a lefordítandó szöveget.

Az összetett fordítástámogató rendszerek legnagyobb előnye azonban az, hogy úgynevezett *fordítómemóriát* is tartalmaznak. A fordítómemória olyan program, amely adatbázisban tárolja az eddigi fordításokat (az eredeti szöveggel együtt), és ha a fordítás során olyan mondathoz érünk, amely már benne van a fordítómemóriában – vagyis korábban már lefordítottuk – a program automatikusan felajánlja a korábban megjegyzett fordítást. Fontos, hogy nemcsak azokat a mondatokat lehet így megtalálni, amelyek *pontosan* megegyeznek az épp lefordítandóval – ilyen gyakorlatilag nincs –, hanem a nagyon hasonlókat is. A hasonlóság megkeresésében viszont a legtöbb gyártó olyan matematikai módszereket alkalmaz, amelyekhez nincs szükség nyelvi elemzésre, s amelyek kizárólag a szöveg betűkódjait veszik alapul. Ez sok pontatlanságot eredményez, a nyelvi-nyelvtani hasonlóság megállapításához viszont megfelelő elemzőprogramot kell a rendszerbe építeni. Az utóbbi minden érintett nyelvhez jelentős nyelvtechnológiai eszközöket igényel, míg az előbbi ugyanazt a matematikai eljárást alkalmazhatja minden lehetséges nyelvre.

A fordítómemóriák üresen érkeznek a felhasználóhoz. Ha kizárólag fordítás közben töltjük fel őket, egy-két évnek is el kell telnie ahhoz, hogy a rendszer elég gyakran adjon fordítást az adatbázisából, s valóban lerövidítse a fordítás idejét. Azonban a legtöbb rendszerhez tartozik olyan program, amely lehetővé teszi a meglévő fordítások bevitelét a fordítómemóriába – ha az eredeti szöveg és a fordítás is megvan számítógépen. A program párba állítja – egymáshoz rendeli – az eredeti szöveg és a fordítás mondatait, így nemcsak az egész dokumentumról tudjuk, hogy mi a fordítása, de az egyes mondatokról is. A mondatokat párosító program a szövegek *szinkronizálását* végzi el. A szövegszinkronizáló programok segítségével a fordítók rövid idő alatt a fordítómemóriába vihetik, s ily módon újra felhasználhatják korábbi – „nyelvtechnológia előtti” – fordításaikat.

A fordítástámogató eszközökről is fontos persze tudni, hogy igazából csak szakfordításhoz, szakmai szövegek feldolgozására jók. Irodalmi vagy a szépirodalomról szóló szövegek fordításában a gép segíteni sem tud az embernek – s tulajdonképpen ez így is van jól.

13. Az előzőekben a számítógépek nyelvi szolgáltatásainak sok olyan hiányosságát említettük, amelyek – legalábbis mai tudásunk szerint – könnyen kiküszöbölhetők. Alább megpróbáljuk megjósolni, hogy rövid időn – egy-két éven – belül milyen nyelvtechnológiai kutatások-fejlesztések eredményeit várhatjuk a hétköznapi gyakorlatban.

Számos kutatás van folyamatban, amely a „valódi” információkeresést tűzte ki célul. Ez azt jelenti, hogy a gép a természetes nyelven feltett kérdésre olyan dokumentumokat, adatalemeket keres, amelyek *tartalma* megfelel a kérdés *jelentésének*. Ez persze az ideális eset, de számíthatunk arra, hogy a számítógépes szolgáltatások egyre jobban megközelítik. Ehhez a nyelvtechnológia elméleti alapjainak két létfontosságú ponton kell fejlődniük.

Ha a nyelv szerkezetét a mainál fejlettebb modellekkel tudjuk leírni, a számítógép a mondatokról, a szövegekről részletesebb és mélyebb nyelvtani információkhoz juthat. Ez az utóbbi időben kutatási „divattá” vált, s talán ezáltal gyorsabban fejlődik a szavak, kifejezések jelentésének gépi ábrázolása. Egyre nagyobbak lesznek, s egyre több területen megjelennek azok az adatbázisok, amelyek fogalmakat és azok egymással való kapcsolatát írják le: ezek az ún. *ontológiák*. Filozófiai értelemben vett ontológia csak egy van; azonban itt a szót új jelentésében használjuk, ahol az ontológia technikai eszköz, amely egy adott terület fogalmait rendezi logikai rendszerbe. (Íme, az Olvasó most láthatta, hogyan fejlődik a nyelv, hogyan kapnak régi szavak új jelentést.)

A közeljövőben a jelentés ábrázolása és a tartalomra épülő információkeresés nem válik általánossá, továbbra is egy-egy szakmához, szakterülethez kötődik. Azonban ezek a szakmai keresőrendszerek egyre pontosabbakká válnak. A legnagyobb kihívás a nyelvtani elemzés és a jelentés ábrázolása közötti kapcsolat meghatározása; arra kell választ találni, hogyan írja le a gép az elemzett dokumentum vagy szövegrész összetett jelentését a nyelvtani elemzés eredménye alapján. Egyelőre nincs szó hosszú dokumentumokról; feladat lehet például, hogy a gazdasági elemzők egyes rövidhírek tartalma alapján lássák egyes cégek működését (például a tulajdonosváltásokat).

Bár nem lankad az automatikus gépi fordítás kutatása és fejlesztése, az emberi beavatkozás nélküli gépi fordítás kisebb hangsúlyt kap. Ez a terület az utóbbi időben két, jól elkülöníthető irányt mutat. Az egyik cél az idegen nyelvű szövegek megértésének segítése. A felhasználónak ekkor nem kell lefordítania a szöveget, tartalmának azonban hasznát veszi. A megértéstámogatás alapvető segédeszközei a gyorsfordító programok, amelyek mai formájukban speciális szótárprogramként működnek. Továbbfejlesztésük célja az lehet, hogy ne csak egyes szavakat és a szótárban levő, „konyhakész” kifejezéseket mutassák meg, hanem ténylegesen lefordítsanak mondatokat vagy kifejezéseket. Ehhez alapos nyelvtani elemzésre, jó nyelvi modellre és tulajdonképpen a gépi fordítás szinte teljes fegyvertárára szükség van. Az ilyen

eszköz kifejlesztése rendkívül munkaigényes, mivel minden nyelvhez, minden nyelvpárhoz külön nyelvtant és fordítási szabályokat kell létrehozni. (A nyelvpár a fordításban az eredeti szöveg nyelve és a fordítás nyelve együtt.) A közeljövőben arra számíthatunk, hogy néhány nyelvpárhoz megjelennek jó minőségű mondatszintű gyorsfordítók, amelyek mindig akkora egység fordítását jelenítik meg, amekkorát a nyelvtani elemzőjük még felismert. Azonban nem várható, hogy ilyen programok tetszőleges nyelvpárhoz rendelkezésre álljanak.

A fordítók támogatását olyan összetett fordítástámogató rendszerek javíthatják, amelyekben a *fordítómemória* nem egyszerű matematikai eljárással keresi meg az éppen fordítandó mondathoz hasonló részeket. A keresés ehelyett a nyelvi hasonlóságra épül: a másképpen szövegezett, de nyelvtani szerkezetüket tekintve hasonló mondatok felismerése egy lépés abba az irányba, hogy a gép itt is a szöveg *tartalma* alapján végezze a keresést. Sajnos, a nyelvi szempontból intelligens fordítómemória létrehozása is sok munkával jár, mivel az eredeti szövegek nyelvéhez viszonylag nagy mélységű mondatelemző programra van szükség. Így a közeljövőben – a gyorsfordítóhoz hasonlóan – e programok is csak néhány nyelvpárral lesznek használhatók, később azonban – ahogy a megfelelő nyelvi modellek egyre több nyelvhez megjelennek – univerzálisan hozzáférhetővé válhatnak.

Egy másik probléma, hogy a számítógép számára mindig komoly nehézséget okozott a beszéd, a papírra nyomtatott szöveg és a kézírás felismerése. A gépen rendelkezésre álló felismerési eljárások rengeteg hibalehetőséget rejtenek. Mindhárom kommunikációs forma esetén hasonló hibákra számíthatunk: nem lesz pontos a szavak és mondatok elejének és végének meghatározása, és nem lesz pontos a jelek felismerése sem. A gép sok esetben nem képes egyértelműen meghatározni, melyik jellel találkozott: nyomtatott szöveg olvasásakor keveredhet például a *h* és a *k*, ekkor a gép csak annyi információt ad, hogy a *h* és a *k* betű közül az egyik érkezett. Keveredhet ugyanakkor zaj és információ: halvány betű vagy halk beszéd, illetve foltos papír és háttérzaj zavarhatja a felismerést. A nyelvtechnológia ekkor azzal tud segíteni, hogy szó- és mondatelemző programok révén meghatározza, hogy a beszédet, a nyomtatott szöveget vagy a kézírást felismerő program által jelzett alternatívák közül melyik lehet valódi szó, kifejezés vagy mondat. Mind a nyelvtechnológiai eszközök, mind a felismerési eljárások fejlődnek. Ennek eredményeképp mára olyan pontos szótáralapú kézírás-felismerő rendszerek jöttek létre, amelyek lehetővé tették az *elektronikus palatábla* megjelenését. Ez olyan számítógép, amelynek csak képernyője van, billentyűzete nincs; amikor szöveget és adatokat rögzítünk rajta, a képernyőre írunk. A számítógép leginkább palatáblához hasonlít, miközben képességei megegyeznek bármelyik asztali számítógépével. Ezek a gépeken a kézírás felismerését a nyelvtechnológia támogatja. Ez egyben azt is jelenti, hogy ilyen kézírás-felismerés még nem áll rendelkezésre minden nyelvhez, így a magyarhoz sem. A közeljövőben azonban számíthatunk ennek megjelenésére is.

14. A számítógép rengeteg, nyelvvel kapcsolatos dologban segíti az embert: a jó minőségű szövegek létrehozásában, az információ keresésében, az idegen nyelvű szövegek megértésében és a fordításban, hogy csak a legfontosabbakat említsük. Emögött azonban nincs valódi nyelvtudás: a számítógép nem „érti” és „beszéli” a nyelvünket úgy, ahogy mi; de sok esetben jól utánozza az embert. A nyelv alapvető kapcsolatban van az ember világismeretével: ez a világismeret azonban nem áll a számítógép rendelkezésére. A nyelv *szerkezete* azonban modellezhető matematikai eszközökkel, s ez átültethető a számítógépre; így a számítógép képes lehet szavak és mondatok elemzésére.

Az emberi nyelv számítógépi kezelése még fejlődése kezdetén jár. A jövőben sokkal több szolgáltatást várhatunk tőle, mint amennyit ma nyújt. Ma viszont azt kell mondanunk, hogy a számítógép gyakorlatilag semmit sem tesz emberi beavatkozás nélkül, ami a nyelvvel kapcsolatos. Minden nyelvi szolgáltatás egy vagy több ponton az ember megerősítését igényli. Azzal tehát, hogy az ember nyelvét néhány ponton megtanulta utánozni, a gép még nem vált intelligenssé. A nyelvtechnológia leginkább arra készíti fel a számítógépet, hogy a nyelvvel kapcsolatos unalmas, monoton rutinmunkában segítse az embert. Azonban a nyelvtechnológiai szolgáltatások igen nagy hányada még nem hagyta el a kutatólaboratóriumokat; ezek felhasználásával pedig mind többet és többet tudhatunk meg – nem a számítógépről, nem a nyelv gépi utánzásáról – magáról a nyelvről.