Prószéky Gábor
Országos Pedagógiai Könyvtár és Múzeum

Nem pusztán az a kérdés, hogy „Helyes-e?", hanem hogy mennyire intelligens...

A mikroszámítógépek sebesség- és kapacitásnövekedésével a korai gépi fordításból kinőtt tudományág, az annál lényegesen többet lefedő számítógépes nyelvészet és a dokumentációkezeléssel foglalkozó diszciplínák újra igen közel kerültek egymáshoz. Ennek kapcsán sok olyan nyelvi szoftvereszköz készül, mely mind a hétköznapi géphasználatot, információkeresést, mind a speciálisabb kutató-fejlesztő munkát támogatja. Sőt, már a magyar nyelvet ismerő első modulok is megjelentek...

A tanulmány egy megvalósított és egy megvalósítás alatt álló rendszer ismertetése kapcsán megpróbál egyben eligazítást is adni a napjainkban egyre több helyen megjelenő nyelvi szoftverek világában. A DISNET programrendszer bemutatása nemcsak természetes nyelvi moduljai miatt érdekes, hanem mert egy olyan világ – az egységes európai információs rendszerek világa – felé vezet, amelynek mi, magyarok is tagjai lettünk. A DISNET fő célja összekötni a felhasználót az általa igényelt információs rendszerrel, függetlenül attól, hogy az hol található meg Európában, vagy attól, hogy a felhasználó tudott-e arról, hogy melyik rendszert kívánja lekérdezni, sőt még attól is, hogy a felhasználó egyáltalán ismeri-e az adott rendszer formális nyelvezetét. Mindezt persze a legkorszerűbb távközlési és szoftvertechnikával kell megoldani, s mivel a végfelhasználó kényelmének kiszolgálása igen fontos, természetes nyelvet kezelni képes moduljai is lesznek egy ilyen rendszernek. Mivel a számítógépes nyelvi eszközök nem túlzottan ismertek idehaza, a tanulmányt ezek leglényegesebb tulajdonságainak ismertetésével kezdjük. Miközben osztályozzuk a természetes nyelvek számítógéppel történő kezelésére alkalmas programokat, a szerző jelen kutatási-fejlesztési munkái kapcsán mindegyikre konkrét gyakorlati példát is mutatunk. A példák a részrendszerek működését hivatottak elsősorban illusztrálni, de gondolatokat is elindíthatnak az adott eszköz önálló alkalmazhatóságával kapcsolatban. A dolgozat egy olyan, most lezáruló konkrét fejlesztő munka lépéseit ismerteti, melyben a szerző is közreműködött, illetve ennek kapcsán egy nem titkolt terv hazai megvalósításának, a MORPHOLOGIC részben elkészült, részben most készülő magyar nyelvi szoftverrendszerének néhány modulját is bemutatja. Ennek a célja a konkrét érdeklődés felkeltése a már megvalósult, illetve a megvalósítás alatt álló magyar nyelvi szoftvereszközök iránt, melyeknek hazai információs és dokumentációs rendszerekhez való kapcsolódását vázolja a tanulmány utolsó része.

A természetes nyelvek és a számítógép lehetséges kapcsolatai

A *számítógépes nyelvészetnek* (a továbbiakban: SzNy-nek) nevezett diszciplína – mint minden alkalmazott tudományág – nagyon sokféle módon osztható fel ágazatokra (ennek részleteiről lásd [1]). Egy lehetséges közelítésmódot, a SzNy-programok bemenő és kimenő adatok szerinti osztályozását az alábbiak szerint tehetjük meg (TNy: természetes nyelv, FNy: formális nyelv):

	Bemenő adatok	Kimenő adatok
(1)	TNy	FNy
(2)	TNy	TNy
(3)	FNy	TNy
(4)	FNy	FNy

Világosan látszik, hogy az első három eset tekinthető igazán számítógépes nyelvészeti rendszernek, a negyedik, ahol a természetes nyelv közvetlenül nincs jelen, csak egy – egyébként számítógépes nyelvészetinek minősített – rendszer valamely részprogramjaként jöhet szóba. A típusok érthetetlenek a hozzájuk tartozó konkrét alkalmazások ismerete nélkül. A továbbiakban bemutatjuk a legtipikusabb számítógépes nyelvészeti programcsoportokat, ismertetjük gyakorlati hasznosságukat, és – a könnyebb érthetőség kedvéért a magyar nyelvből hozott – példákkal, valamint egy meglehetősen leegyszerűsített formalizmussal próbáljuk még érthetőbbé tenni őket.

Az első típusba tartoznak a *szövegmegértő, szövegkivonatoló* (1. példa), illetve az adatbázist, tudás-

bázist természetes nyelven *lekérdező* rendszerek (2. példa). Ezekben a kiinduló adat valamely TNy-en leírt szöveg, vagy TNy-en megfogalmazott kérdés. Szövegmegértő rendszerekre olyan számítógépes környezetben van igény, ahol az információ bevitele formális módon nagyon nehézkes. Ha a felhasználónak nem áll módjában egy bonyolult beviteli formalizmust megtanulni, esetleg ideje sincs rá – ilyenkor segíthetnek a szövegmegértő szoftverek. A szövegkivonatolás a nagyméretű, géppel olvasható formájú szövegek tartalmának későbbi lekérdezésre alkalmas formalizmusba való fordítását jelenti. Ilyenek például az újságcikkeket, jelentéseket, híranyagokat tároló számítógépes rendszerek, melyek esetében sokszor nem a betű szerint visszakereshető információk, hanem a tartalmiak a fontosak.

A második típusba a *gépi fordító* rendszerek (3. példa), a teljes TNy-választ generáló *dialógusrendszerek* (4. példa) és a TNy-bemenetet korrigáló-átalakító, *nyelvhelyességet ellenőrző*, illetve *nyelvtani-stilisztikai* átalakításokat támogató rendszerek tartozhatnak (5. példa). A gépi fordítás jelentőségét talán nem is kell ecsetelni, hiszen ma már hazánkba is annyi idegen (elsősorban angol) nyelvű dokumentum érkezik, hogy nemcsak lefordítani, de elolvasni is kevés rá az idő. Itt egy esetleg nem is irodalmi igényű, de ma már jelentős sebességű gépi fordító rendszernek nagy szerepe lehet. A dialógusrendszer a sokféleképpen lekérdezhető adatbázisokra és tudásbázisokra épülő olyan információszolgáltató program, mely az ember– ember párbeszédet is képes kiváltani. Ilyenek az utazási, vásárlási vagy éppen általános tájékoztatási információs rendszerek, ahol a felhasználónak sem ideje, sem kedve nincs formálisan megfogalmazott válaszok közt böngészni. A nyelvhelyesség bármely szintjét ellenőrző szoftvereszközök ma már beépültek a legtöbb szöveg- és kiadványszerkesztő programba, az optikai karakterfelismerőkbe vagy éppen a szöveges adatbázis-kezelőkbe, ezzel is támogatván a lehetőségek szerinti minél pontosabb munkát.

Harmadik típusú SzNy-rendszer minden *mondat- és szöveggeneráló* program (6. példa). Ezek általában a "számítógép agyában megfogalmazódott gondolatokat" alakítják át emberi nyelvekre. Ilyen például az időjárással kapcsolatos mérési adatokat begyűjtő számítógép automatikus időjárásjelentés-készítő programrendszere, vagy bizonyos gépezetek, szabályozó rendszerek belső állapotáról időnként szöveges jelentést készítő programok. A legtöbb generáló modul a felsorolt példák ellenére azonban elsősorban mint a gépi fordító vagy a dialógusrendszerek alrendszere ismert.

Igény tehát van a nyelvi tudással megtámogatott számítógépes eszközök használatára, a gépek is megfelelően gyorsak, és tárolási kapacitásuk is kielégítő. Így nem kell ahhoz nagy bátorság, hogy megjósoljuk: a 90-es évek hátralevő részében a hazai információs rendszereknek (pl. a világkiállításéinak) egyre több magyar nyelvi tudással rendelkező moduljával fogunk találkozni.

1. példa
Szövegmegértő, szövegkivonatoló

BEMENŐ TNY SZÖVEG ::
A baglyok őrjöngenek a muzsikáért.
KIVONATOLT FNY SZERKEZET ::
BAGOLY ← SZERET → ZENE

2. példa
Lekérdező

TNY KÉRDÉS ::
Mit szeretnek a baglyok?
FNY VÁLASZ ::
ZENE

3. példa
Gépi fordító

TNY SZÖVEG (FORRÁSNYELV) ::
Mit szeretnek a baglyok?
TNY FORDÍTÁS (CÉLNYELV) ::
What do the owls like?

4. példa
Dialógus

TNY KÉRDÉS ::
Mit szeretnek a baglyok?
TNY VÁLASZ ::
Annyi biztos, hogy a zenét igen.

5. példa
Nyelvtani-stilisztikai átalakító

BEMENŐ TNY SZÖVEG ::
Mit komálnak a bagolyok?
MÓDOSÍTOTT TNY SZÖVEG ::
Mit szeretnek a baglyok?

6. példa
Szöveggeneráló

BEMENŐ FNY SZERKEZET ::
BAGOLY ← SZERET → ZENE
GENERÁLT TNY SZÖVEG ::
A baglyok kedvelik a muzsikát.

A természetes nyelvi interfészről, általában

A most ismertetendő *TNy/FNy* modulok legátfogóbb példája maga a teljes *természetes nyelvi lekérdező rendszer*, melynek egy konkrét megvalósítása a következő fejezetben bemutatandó DISNET NLI. Ennek első nagyobb moduljai, a *morfológiai elemző* (jelen esetben ez a MORPHOLOGIC Humor rendszere) és a *szintaktikai elemző* szintén TNy/FNy rendszerek. E rendszerek formális szerkezeteket feleltetnek meg egy természetes nyelvi bemenetnek, a szöveg szavainak, illetve mondatainak (7. példa).

TNy/TNy rendszer a bemutatandó TNy-interfész *normalizáló* modulja, mely udvariassági formulákkal, és az érdeklődést kifejező formális sallangokkal

telitűzdelt TNy-kifejezéseket alakít át köznapi TNy-kifejezésekké (8. példa). A *helyesírás-ellenőrző programoknak* (esetünkben a Helyes-e? programcsaládnak) mind a bemenetén, mind a kimenetén természetes nyelven írt szöveget találunk (9. példa). Ugyanígy viselkednek a *számítógépes szinonimaszótárak* (esetünkben a Helyette), melyek szintén csak szó-, illetve kifejezéscsere segítségével alakítanak át szövegeket (10. példa).

FNy/TNy rendszer bármilyen *nyelvgeneráló program*, melynek kiindulópontja valamely formális nyelven megfogalmazott ismeretreprezentáció. Mivel a DISNET NLI-nek generáló modulja nincs, a fent bemutatott fejlesztési környezetben a Humor rendszer morfológiai generáló modulja az egyetlen ilyen funkciójú szoftver. Ez van beépítve a Helyette rendszerbe, mely ezáltal képes a kiválasztott szinonim alak megfelelő toldalékokkal való szabályos ellátására, az eredeti szóról leválasztott toldalékok formális definíciója alapján (11. példa).

FNy/FNy rendszer a DISNET NLI több rendszere is, mivel a – későbbiekben ismertetendő – fordítási folyamat a kiinduló TNy-ről a célnyelvre több lépésben zajlik, így köztes formális nyelvek jelennek meg, melyek mindegyike felváltva forrás-, illetve célnyelvként kezelendő.

7. példa
Morfológiai elemző, szintaktikai elemző

BEMENŐ TNY SZÖVEG ::
A baglyok őrjöngenek a muzsikáért.

A MORFOLÓGIAI (FNY) SZERKEZET ::
a = a [névelő]
baglyok = *bagoly* [főnév] + *ok* [t.szám]
őrjöngenek = *őrjöng* [ige] +
+ *enek* [t.szám, 3. személy]
a = *a* [névelő]
muzsikáért = *muzsika* [főnév] + *ért* [esetrag]

A SZINTAKTIKAI (FNY) SZERKEZET ::
ALANY =
a [névelő] *bagoly* [főnév] + *ok* [t.szám]
ÁLLÍTMÁNY =
IGE =
őrjöng [ige] + *enek* [t.szám, 3. személy]
HATÁROZÓ =
a [névelő] *muzsika* [főnév] + *ért* [esetrag]

8. példa
Normalizáló

BEMENŐ TNY SZÖVEG ::
Azt szeretném megtudni, kérem szépen, hogy valójában mit szeretnek a baglyok?
NORMALIZÁLT TNY SZÖVEG ::
Mit szeretnek a baglyok?

9. példa
Helyesírás-ellenőrző

BEMENŐ TNY SZÖVEG ::
Mit szeretnek a bagjok?
MÓDOSÍTOTT TNY SZÖVEG ::
Mit szeretnek a baglyok?

10. példa
Szinonimaszótár

TNY KÉRDÉS ::
szeret
TNY ROKON ÉRTELMŰ SZAVAK ::
kedvel, imád, vonzódik

A DISNET rendszer természetes nyelvet használó intelligens interfésze

E fejezetben röviden bemutatjuk a korábban említett **DISNET** *(Domain Independent Intelligent Information and Services Network Interface)* rendszert, melyet a Közös Piac megbízásából fejlesztett ki a holland IDE cég. A rendszer maga egy eszközkészlet, melynek segítségével Európa elektronikus szolgáltatásokat nyújtó intézményei lehetővé tehetik elérésüket más hálózatok vagy szolgáltatások felhasználói számára, akik sem beletanulni nem akarnak az újonnan csatolt rendszerek használatába, sem azok nyelvét nem akarják elsajátítani. Arról van tehát szó, hogy a szoftvercsomag magára vállalja a felhasználó igénye alapján történő szolgáltatáskiválasztást, a szolgáltatást biztosító számítógéppel a kommunikáció felvételét, valamint a keresett információ megtalálását és visszajuttatását a kérdezőhöz, ismét csak a hálózaton át.

A DISNET *alkalmazásfüggetlen*, bár vannak elsődlegesen lefedni kívánt alkalmazási területek. Az elsőként kiválasztott szakterület a mezőgazdaság– mikrobiológia– élelmiszeripar hármas. Az Európai Közösség mezőgazdasági információs rendszereiről összefoglaló ismeretek kerülnek be egy e célra készített és a felhasználói oldalon elhelyezkedő tudásbázisba. A felhasználó mindaddig ezzel a tudásbázissal, illete ennek az adott szakirány (esetünkben a mezőgazdaság) tezauruszából készített szaktudásbázissal folytat dialógust (természetes nyelven, esetleg olykor menü segítségével is), míg a számára szükséges információt, dokumentumot nagy valószínűséggel tartalmazó információs bázis(oka)t – azaz: adatbázisokat, videotex rendszereket és elektronikus postai szolgáltatásokat – a rendszer elérésre fel nem ajánlja. Ekkor a felhasználó a szolgáltatás pontosságának, árának és sebességének hozzávetőleges ismeretében dönthet, hogy mely rendszerekhez kíván hozzákapcsolódni a felkínáltak közül. A kapcsolatot a DISNET automatikusan létrehozza, majd a kikeresett információt eljuttatja a felhasználóhoz. Mindez az egy adott szoftverkörnyezetet megszokott felhasználó számára az általa használt információs rendszer minimális megváltoztatásával történik, ugyanis bármilyen nagyobb formai változás megzavarhatná a jól bevált hétköznapi használatot. A DISNET szoftver karakteres vagy grafikus képernyőjű számítógépekre vagy videotex terminálra egyaránt fel van készítve.

A *felhasználói interfész* teszi lehetővé a felhasználó és a DISNET rendszer közötti kommunikációt. Az adott információ Európában történő megtalálására irányuló kérdéseket a felhasználó természetes nyel-

ven, esetleg menüből, vagy egy konkrét szolgáltatás explicit kiválasztását követően az adott szolgáltatás célnyelvén teszi fel. Egy ún. kommunikációs processzor kezeli a felhasználó és a különféle szolgáltatások közötti kapcsolatokat. Az információszolgáltatók interfészprocesszora fordítja le a hálózatban belső reprezentációban megjelenő üzeneteket az adott szolgáltatás belső nyelvére, pontosabban annak egy töredékére. Tehát nem a már-már szabvánnyá vált bonyolult struktúrájú nyelvekre, mint pl. az SQL lekérdezőnyelvre, vagy az Európai Közösség CCL parancsnyelvére való fordítás, hanem azoknak csak egy nagyon szűk, kulcsszavas lekérdezésre alkalmas résznyelvére való konvertálás a cél.

A tudásbázis, az ezt használó következtető rendszer, valamint a természetes nyelvi alrendszer is a felhasználó gépen helyezkedik el, és amíg nincs szükség pontos adatokat igénylő külső információra, a beszélgetés a felhasználó és a szoftver közt kizárólag a felhasználó gépén folyik. A hálózat használatára csak akkor kerül sor, amikor erre már elengedhetetlenül szükség van.

A DISNET rendszer vázlatos felépítése a fentiek alapján az *1. ábrán* látható módon foglalható össze.

1. ábra

A DISNET intelligens interfész moduljai és továbbfejlesztésük

Mivel a szakértő rendszerekkel történő szokásos, rendszervezérelt dialógus meglehetősen hosszú is lehet, lehetőség van a párbeszéd természetes nyelven történő indítására. Ezzel egy menüvezérelt dialógus első néhány, vagy akár néhány tíz lépésétől szabadulunk meg. A későbbiek során a rendszer természetes nyelvű válaszadásra is képes lesz, de a bemutatandó fejlesztési fázisban erre még nem volt mód. Az itt megfogalmazott feladatokat valósítja meg a DISNET NLI, azaz a DISNET rendszer természetes nyelvi interfésze (NLI: Natural Language Interface).

A természetes nyelvről a tudásbázis lekérdező nyelvére történő fordítás több lépésben zajlik [2]. Először az alkalmazásfüggetlen *elsődleges logikai nyelvre* (LL = Logical Language) fordul le a TNy-bemenet, aztán az alkalmazásfüggő *másodlagos logikai nyelvre* (DL = Domain Language), majd innen a *tudásbázis-lekérdező nyelvre* (KBQL = Knowledge Base Query Language).

A fordítások során használt alaprendszer tehát a szótárakból, az ezeket kezelő modulokból, a beviteli modulokból, a fordító modulokból és természetesen a szakértő rendszer moduljaiból áll, melyek már nem közvetlenül tartoznak a nyelvi alrendszerhez.

▶ **Forrásszótár**

A nyelvész hozza létre a szótárak forrás-, azaz emberek által is olvasható alakját. A szótárak ebben a kontextusban általában tő- és toldaléktárakat jelentenek. Struktúrájuk:

<MORFÉMA> <MORF.KÓDOK> <SZINT./SZEM.KÓDOK>

A morféma maga a tő vagy a toldalék; a morfológiai kódokat a szóalaktani rendszer használja a szóalakok szegmentálásához; a szintaktikai/szemantikai kódok pedig a morfológiai elemzés kimenetén jelennek meg. A DISNET NLI-nek csak a prototípus-verziója készült el, mintegy 1000 szótári alakkal, ezzel szemben a Humor rendszerek szótárai több mint 80 000 tövet tartalmaznak. Meg kell említenünk, hogy az elemzési sebesség nem változik a Humor rendszerben a szótári egységek számának növekedésével.

▶ **Szótárfordítók**

A szótárakat a szótárfordítók hozzák tömör belső alakjukra. Ezek a struktúrák nagy sebességű visszakeresésre vannak kidolgozva. Maga a fordítás gyors művelet: a MORPHOLOGIC szótárfordítóinak sebessége átlagosan 16 000 szó/perc.

▶ **Tárgyszótárak**

A szótárfordítók működésének eredményeképpen létrejönnek a csak gép által olvasható struktúrájú tárgyszótárak. Ezek méretéről képet kaphatunk, ha a 80 000 tövet tartalmazó magyar szótárét megadjuk: ez mintegy 600 KB, minden kóddal együtt.

▶ **Helyesírás-ellenőrző**

A bevitelkor lehetetlen elkerülni az esetleges elütéseket. Ezek felismerésére és kijavítására szolgálnak a helyesírás-ellenőrzők. A DISNET NLI helyesírás-ellenőrzője észreveszi az elütéseket,

de – mivel angol nyelvre íródott – mindössze a szótárban felsorolt alakok átnézésével, nem pedig egy teljes morfológiai analízis segítségül hívásával, amint ez a magyar nyelv esetében természetes. Erről, valamint a javítások automatikus korrigálásáról részletesebben később szólunk, a Helyes-e? rendszer ismertetése kapcsán.

▶ **Normalizáló modul**
Fő funkciója az udvariassági és egyéb pragmatikai szempontból jelentős, de az információkeresés területén irreleváns kifejezések kiejtése a bemenő szövegből. A normalizáló által előállított kérdések, parancsok és állítások már olyan alakban vannak, melyeket a rendszer további elemző fázisai kezelni képesek. A DISNET NLI-ben 16 normalizáló szabályosztály működik, pl. a
Where can I become informed on X?
kérdés normalizált alakja mindössze
about X
lesz (vö. 8. példa).

▶ **Szintaktikai szabályrendszer**
A morfológiai rendszer kimenetén megjelenő kifejezésekre épülő magasabb rendű grammatikai struktúrák leírására szolgál. A nyelvész egy adott formális, szabályleíró metanyelven fogalmazza meg a mondat szerkezetére vonatkozó ismereteit, melyeket a szabályfordító (l. alább) hoz a rendszer által közvetlenül használható alakba. A DISNET NLI ún. extrapozíciós nyelvtannal megadott szabályrendszere 160 szabályból áll.

▶ **Szintaktikaiszabály-fordító**
A szintaktikai szabályok belső reprezentációra való fordítását végzi. A DISNET NLI esetében – lévén az annak alapjául szolgáló extrapozíciós nyelvtan Prolog-alapú rendszer – egy Prolog program áll elő belőlük. Ez lesz a futó programrendszer része, nem pedig a nyelvész által kódolt eredeti szabályrendszer. A MORPHOLOGIC szintaktikai rendszerét egy a Humor morfológiai rendszer fordításához kidolgozott szabályfordítóhoz rendkívül hasonló program hozza a működő elemzőrendszer által használható alakra.

▶ **Az LL/DL fordító modul**
Az LL-szerkezet az eredeti bemenő mondat olyan logikai szerkezetét írja le, mely független a célnyelvtől (a DISNET esetében a tudásbázis-lekérdező nyelvtől, a KBQL-től). A szerkezetet egyébként a szintaktikai elemzés eredményeként kapott gráf csomópontjai, és a köztük fennálló logikai jellegű relációk alkotják. Erről a formális nyelvről kell egy másik formális nyelvre, a DL-re fordítani. A DL már függ az adott világtól és az alkalmazási területtől. A nyelv maga az elsőrendű predikátumlogika egy olyan részhalmaza, mely Prolog nyelven jól kezelhető.

▶ **A DL/KBQL fordító modul**
A formális logikai nyelvek sorát egy, már tudásbázis-kezelésre is alkalmas Prolog program zárja. Az erre való fordítás a formális deriváláshoz hasonló egyszerű technikai eljárás.

▶ **Tudásbázis-alapú szótárgeneráló**
A rendszer lexikai ismereteinek nagy része a nyelvi modulok alapszótárain alapul. Ezek a szótárak azonban a konkrét alkalmazások esetén ki kell hogy egészüljenek az adott alkalmazási terület speciális szókincsével. Ezek egyik legalapvetőbb forrását a tudásbázisban szereplő kifejezések adják. A tudásbázis-alapú szótárgeneráló modul az efféle információs rendszerek automatikus szótárépítéséhez nélkülözhetetlen.

A MORPHOLOGIC magyar nyelvi számítógépes rendszerei és alkalmazásaik

A fejlesztések alapjául szolgáló nyelvészeti formalizmus a 80-as években méltán népszerűvé lett unifikációs leíráson alapul. Az *unifikációs morfológia* lényege, hogy egy szóalakon belül a morfémák találkozási pontjainak – pl. tő/tő, tő/képző, tő/rag, képző/rag stb. – vizsgálata a két szóban forgó elem szóalaktani tulajdonságait leíró jegyek egyfajta speciális összehasonlításán, unifikációján alapul. Az elvnek, mely a **Humor** (*H*igh-speed *U*nification *Mor*phology) nevet kapta, első alkalmazása a magyarra történt meg, de egyazon rendszerben azóta több más nyelv (angol, görög, latin, lengyel, német, olasz, török stb.) szóalaktanának feldolgozása is megkezdődött.

A Humor morfológiai elemző és generáló rendszer

Az *elemző modul* feladata a szótárban szereplő, vagy szabályos szóösszetétellel, illetve szóképzéssel előállítható bármely (relatív) tő tetszőleges szabályosan toldalékolt alakjának pontos felbontása minden lehetséges módon (11. példa).

11. példa
A HUMOR morfológiai elemzője

BEMENET: *mentek*
KIMENET: *ment*[MN]+ *ek*[PL] (mentesek)
ment[IGE]+ *ek*[e1] (én mentek)
megy[IGE]= *men*+ *tek*[t2] (ti mentek)
megy[IGE]= *men*+ *tek*[Mt3] (ők mentek)

12. példa
A HUMOR morfológiai generátora

MINTASZÓ: *gyerekeimnek*
BEMENET: *nagyapa*
KIMENET: *nagyapáimnak*
BEMENET: *ló*
KIMENET: *lovaimnak*
BEMENET: *barack*
KIMENET: *barackjaimnak*
BEMENET: *sör*
KIMENET: *söreimnek*

A *generáló modul* kiinduló adatai a szótő és a hozzáadandó toldaléksorozatok kódjai. A technikai megvalósításnál azonban a bonyolult toldalékkód-leírás helyett a mintaszóval megadott generálás látszott célravezetőnek (12. példa).

Szótárépítő modulok, szakszótárak

A szótár minden TNy-rendszer egyik legfontosabb eleme, hiszen itt jelenik meg azon információk nagy része, melyet az adott nyelvet beszélő emberek a nyelvelsajátítás során egy-egy szóról megtanulnak. Természetesen ez nem azt jelenti, hogy a grammatikai szabályok nem fontosak, hanem azt, hogy a mai nyelvfeldolgozó rendszerek elsősorban a lexikális információra építenek, ugyanis a hagyományos nyelvtani szabályok száma az itt felsorolt modulok által is használt unifikációs formalizmusok használata következtében a minimálisra redukálódott. (Részletesebben l. [1].)

A szavak szótárba való felvétele a szótövek és a hozzájuk tartozó morfológiai viselkedésre vonatkozó információk együttes bevitelét jelenti. Természetesen a felszíni alakok, különösen, ha toldalékosak (szövegből valók), az esetek nagy részében nem, vagy csak részlegesen alkalmasak arra, hogy segítségükkel automatikusan el lehessen dönteni az adott szótő más toldalékok előtti viselkedését. E célból félautomatikus szótárépítő programok kifejlesztése vált szükségessé. Ezek a – természetesen nyelvészeti ismeretekkel rendelkező – felhasználó gyors és hatékony munkáját támogatják.

Nem a lexikai információk bevitele az egyetlen kapcsolat a nyelvészek és a MORPHOLOGIC nyelvi szoftverei között, ugyanis a Humor rendszert az MTA Nyelvtudományi Intézete többféle nyelvészeti kutatás támogatására is használja: ennek segítségével elemzik végig az Akadémiai Nagyszótár elkészítéséhez felvitt szövegeket is.

Helyesírás-ellenőrző: a Helyes-e? programcsalád

A morfológiai elemző lecsonkított, leegyszerűsített változata a hétköznapi szövegszerkesztő-használatban is nagy segítséget jelenthet, ugyanis egy tetszőleges magyar szöveg szóalakjainak helyesírási ellenőrzése a szóalaktani elemzésen alapul. Ha egy szóalaknak létezik az adott szótárban megtalálható valamely tőből és a produktív magyar toldalékok sorozatából szabályosan összeálló változata, akkor az a szóalak – legalábbis ezen a környezet- és jelentésfüggetlen szinten – helyesnek mondható. Minden más esetben hibát, pontosabban ismeretlen szóalakot kell jelezni. A Humor rendszer lekarcsúsított helyesírás-ellenőrző változatai, a **Helyes-e?** programcsalád tagjai tehát felismernek minden olyan alakot, mely a fenti értelemben nem helyes.

Egy másik alrendszer foglalkozik a *javaslattétellel*, mely szintén a magyar toldalékolás ismeretében dönt az ismeretlen szó fonológiai– morfológiai tévesztéseken, vagy az egyszerűen csak elütésen alapuló hibák valószínű javításáról. Cél, hogy minél kevesebb, de annál valószínűbb alternatíva kerüljön a felhasználó elé.

Ha egy szó töve nincs meg a szótárban, fennáll a lehetőség, hogy a rendszer használója szótárkiegészítést végezzen. Az ily módon felvett szavak nem tartalmazzák a korábban említett nyelvészeti kódokat, mert egy átlagos szövegszerkesztő-használótól nem várható el a nyelvészeti jártasság. Így előfordulhat, hogy később egy ily módon felvett tőnek egy másik toldalékkal előállított variánsát ismét fel kell venni a felhasználó saját szótárába.

Intelligens szinonimaszótár: a Helyette ragozó tezaurusz

A szövegszerkesztőben használatos tezauruszok az általános információ-visszakeresésre készült tezauruszok egyszerűsített – mindössze a szinonimarelációra épülő – változatai. A technikai megvalósítást illetően azonban nincs jelentős különbség. A nyelvészeti tudással rendelkező tezaurusz esetében egy folyó szöveg tetszőlegesen toldalékolt szava a kiindulópont, és egy adott relációláncon át kiválasztott tőnek a kiinduló alak toldalékolásának megfelelően képzett, ragozott alakja a kimenő információ. Az intelligens szinonimaszótár, azaz a ragozó tezaurusz első számítógépes implementációja, a most vázolt tulajdonságokkal rendelkező **Helyette** rendszer három fő modulból áll: a Humor elemzőjéből, a tulajdonképpeni szinonimaszótárból (vagy a későbbiekben más tezauruszokból) és a Humor generáló rendszeréből. A 13. példa vázlatosan bemutatja a Helyette működését.

13. példa
A HELYETTE toldalékoló tezaurusz működése

BEMENET: *kupáimra*
ANALÍZIS: *kupá + im + ra*
TŐ: *kupá*
SZÓTÁRI TŐ: *kupa*
SZINONIM TŐ: *kehely*
TŐVÁLTOZAT: *kelyh*
SZINTÉZIS: *kelyh + eim + re*
KIMENET: *kelyheimre*

A Helyesel elválasztó modul

Az elválasztó programok alapalgoritmusa általában nem túl bonyolult: leggyakrabban nyelvfüggő módon a magánhangzók és mássalhangzók speciális viszonyán alapul. Így van ez a magyar esetében is. Azonban a morfológiai elemzés szinte minden nyelvben nélkülözhetetlen az elemi szabályok felülbírálásakor (a magyarban ilyen pl. az összetett szavak elválasztása). A pusztán kivételszótárral operáló rendszerek az egyedileg felsorolt alakok esetében adnak csak garantáltan helyes elválasztást, míg a problémát algoritmikusan kezelő rendszerek (pl. a Helyes-e? program **Helyesel** alrendszere) a szótárban explicite nem szereplő összetett szavak elválasztását, vagy a hathármas egybeírási szabály alkalmazását is helyesen oldják meg. A 14. példa hoz néhány olyan elválasztási nehézséget, melyekre a Helyesel helyes megoldást ad.

14. példa

A HELYESEL elválasztó modul működése

filétek	→	*fi-lé-tek*
csalétek	→	*csal-étek*
karosszék	→	*ka-ros-szék*
karosszéria	→	*ka-rosz-szé-ria*
átall	→	*átall*
átáll	→	*át-áll*

Szótő-előállítás a HelyesLem rendszerrel

A gazdagon toldalékoló nyelvek szövegeiben való keresés pontossága aligha oldható meg a szóalakokban lappangó tövek felismerésének, azaz a *lemmatizálásnak* az elvégzése nélkül. A szótárialak-előállítás az így megtalált tövek alapalakját adja át az indexelést végző rendszernek. A toldalékok csonkolással történő eltávolítása általában sok problémát okozhat. A magyarban például az *-ek* többesjel leválasztása nem elegendő például a *kelyhek* alakról, hiszen a *kehely* tőalakkal ennek közös része mindössze két betű, a maradék pedig nem is toldalék. Ugyanakkor pl. az összetételekben szereplő *alma* betűsor megtalálandó tőnek minősül a *vadalma*, de nem a *hatalma* szóalakban. Ennek a problémának a helyes kezelését végzi a Humor rendszeren alapuló **HelyesLem** modul [3].

15. példa

A HELYESLEM szótő-előállító működése

BEMENET:	*lelő*
TÖVEK:	*lel*
	lő
	lelő
BEMENET:	*telefonhívásokra*
TÖVEK:	*telefonhívás*
	telefon
	hívás
	hív

Tervek: magasabb szintű elemzés és...

A kutatás a MOPRHOLOGIC szintaktikai és magasabb szintű elemzőjének hatékony, a hétköznapi munkában is hasznos alkalmazásai irányában folyik. Ennek alapjait a fent ismertetett DISNET rendszerben megvalósított ötletek és a morfológiai implementációk alapgondolatának egyfajta ötvözete képezi.

A készülő rendszer egy melléktermeke lesz a mondatszintű helyesírás-ellenőrző rendszer, a **Helyesebb**, de a fő cél sokkal inkább az intelligens ember– gép kapcsolat megvalósítása, a magyar nyelvű adatbázis-lekérdezés, és talán a nem is olyan távoli jövőben a géppel támogatott fordítás lesz. Ez utóbbihoz, azaz a fordítói munka számítógépes támogatásához szükséges, a ragozó szótárakon túli nyelvészeti tudással rendelkező szoftvereszközök kifejlesztése már jelenleg is folyik.

Végül néhány reménykeltő terv a dokumentációs szakemberek részére:

- A DISNET NLI-ben megvalósított elképzelés alkalmazhatónak tűnik hazai keretek között is, amennyiben az **IIF** információszolgáltatásban illetékes szakemberei is így gondolják. (Egy konkrét példa: a cikk írásakor már készül a BRS/Search keresőrendszernek a HelyesLem modullal kombinált változata.)
- A British Library **PRECIS** rendszere magyar adaptációjának, az OPKM-ben kifejlesztett magyar PRECIS-nek először csak morfológiai, később magasabb rendű nyelvészeti szoftvermodulokkal (HelyesLem, Humor) való támogatása megkezdődött.
- A VIXEN könyvtári rendszere, a **D'Lib** pedig minden jel szerint az első olyan rendszer lesz, melyben a könyvtáros munkáját a magyar nyelvet némiképp értő modulok (Helyes-e?, HelyesLem) is támogatják.

"Együtt lenni látszanak az építőkövek..." [4]

Irodalom

[1] PRÓSZÉKY G.: Számítógépes nyelvészet. Számalk, 1989.

[2] PRÓSZÉKY G.: Natural Language Interface Prototype of DISNET. DISNET Internal Report, IDE, 1992.

[3] PRÓSZÉKY, G.– TIHANYI, L.: A Fast Morphological Analyzer for Lemmatizing Corpora of Agglutinative Languages. Papers in Computational Lexicography (COMPLEX '92). Linguistics Institute of H.A.S. 1992. p. 265– 278.

[4] CSABAY K.: "I dreamed I met a Galilean..." TMT, 39. köt. 10. 1992. p. 441– 448.

Beérkezett: 1993. II. 1-jén.